“Big Data: imkanları, multidissiplinar problemləri və perspektivləri”

I respublika elmi-praktiki konfransın

Ə S Ə R L Ə R İ

25 fevral 2016-cı il

“Big Data: imkanları, multidissiplinar problemləri və perspektivləri”

I respublika elmi-praktiki konfransın

Ə S Ə R L Ə R İ

Təşkilatçılar: AMEA Rəyasət Heyəti

AMEA İnformasiya Texnologiyaları İnstitutu

Bakı, 25 fevral 2016-cı il

PROQRAM KOMİTƏSİ

Həmsədrlər

akademik Akif Əlizadə

akademik Rasim Əliquliyev

PROQRAM KOMİTƏSİNİN ÜZVLƏRİ

Adil Bağırov (Avstraliya)
Anar Səmidov
Bikəs Ağayev
Elvin Balacanov (B. Britaniya)
Əfəndi Nəsibov (Türkiyə)
Əlövsət Əliyev
Fərhad Yusifov
Fərqanə Abdullayeva
Firudin Ağayev
Gülnarə Nəbibəyova
Xalid Hüseynov (Koreya)
Julien Ugon (Avstraliya)
Lüdmila Suxostat
Mədinə Səidova
Məkrufə Hacırəhimova
Məsumə Məmmədova
Nadir Ağayev
Nigar İsmayılova
Ramiz Alıquliyev
Ramiz Şıxəliyev
Rasim Mahmudov
Rəna Qasımova
Rəsmiyyə Mahmudova
Rəşid Ələkbərov
Səyyar Abdullayev
Şahin Əmrahov (Türkiyə)
Şakir Mehdiyev
Şəfəqət Mahmudova
Təhmasib Fətəliyev
Tofiq Kazımov
Vüqar Musayev
Yadigar İmamverdiyev
Zərifə Cəbrayılova

Elmi redaktorlar

Ramiz Alıquliyev

Yadigar İmamverdiyev

Texniki redaktorlar

Səkinə Nuralıyeva

Əminə Ağazadə

S E K S İ Y A L A R

- Big Data Analitika
- Big Data və İnformasiya Təhlükəsizliyi
- Big Data Tətbiqləri

M Ü N D Ə R İ C A T

SEKSİYA # 1. Big Data Analitika 9

Big data konsepsiyası və aktual elmi-nəzəri problemlər 10
Rasim Əliquliyev, Məkrufə Hacırəhimova, Aybəniz Əliyeva

An efficient incremental clustering algorithm for very large data sets 14
Adil Bagirov, Sona Taheri, Julien Ugon

Toplu ulaşımda akıllı kart verilerinin analizi ile yolcu yoğunluğunun tahmin edilmesi 19
Ahmet Can Diker, Efendi Nasiboglu, Elvin Nasibov

Big data and fuzzy based techniques in analysis of social networks 22
Ali Abbasov, Shahnaz Shahbazova

Introduction to data science with Apache Zeppelin 25
Khalid Huseynov

Об одной модели анализа данных большого сетевого трафика 28
Рамиз Шыхалиев

Böyük verilənlər analizinin problemləri 31
Rəna Qasımova

Viki-mühitdə big data problemləri və onların həlli yolları 36
İradə Ələkbərova

Böyük verilənlərdən verilənlər haqqında elmə: fənlərarası perspektiv 40
Məkrufə Hacırəhimova, Hicran Gözəlova

Bulud test mühitlərinin böyük verilənlərin emalında reallaşma imkanlarının tədqiqi və analizi..... 43
Fərqanə Abdullayeva

Böyük həcmli mətnlərin yaxınlığının müəyyənləşdirilməsində big data texnologiyalarından istifadə 46
Nadir Ağayev, Əminə Ağazadə

Böyük videoinformasiya resurslarının emalı problemləri haqqında 49
Bikəs Ağayev

Böyük verilənlərdən faydalı informasiyanın aşkarlanmasında mövcud yanaşmalar 54
Rəna Qasımova

Bibliometric analysis of "big data" research 58
Ramiz Aliguliyev, Nigar Ismayilova

Plan məlumatları əsasında virtual hava məkanında proqnozlaşdırılan münaqişə vəziyyətinin imitasiya modeli 61
Nadir Ağayev, Ağası Məlikov, Hüseyn Babayev, Nəsrəddin İsgəndərov

Наиболее распространенные задачи data mining 64
Сайяр Абдуллаев, Судабе Абасова, Сабина Фоменко

О методах сбора и хранения сетевого трафика компьютерных сетей 67
Рамиз Шыхалиев
Veb istifadəçilərin profilinin qurulması metodu 70
Babək Nəbiyev
Проблемы визуализации многомерных данных 74
Айтадж Агабейли
E-dövlətin big data mənbələri 78
Ramiz Alıquliyev, Günay Niftəliyeva
E-dövlət sisteminin analizində data mining texnologiyalarının tətbiq imkanları 81
Ramiz Alıquliyev, Günay Niftəliyeva
SEKSİYA # 2. Big Data və İnformasiya Təhlükəsizliyi 85
E-dövlətin informasiya təhlükəsizliyi üçün big data texnologiyaları 86
Yadigar İmamverdiyev
Big data tətbiqinin kibercinayətlə mübarizədə potensialı və məhdudiyyətlər 91
Elvin Balacanov
İnformasiya təhlükəsizliyi üzrə tədqiqatlar üçün böyük və kiçik verilənlər 95
Yadigar İmamverdiyev
Big data və fərdi məlumatların qorunması problemləri 99
Rasim Mahmudov
Big data və təhlükəsizlik 103
Toğrul Qafarbəyli, Şəbnəm Ələkbərova
Biometrik texnologiyalarda big data problemləri 106
Şəfəqət Mahmudova
Big data və fərdi məlumatların təhlükəsizliyi 109
Yadigar İmamverdiyev
Vacib infrastukturların idarəetmə sistemlərinin bulud texnologiyaları əsasında təşkili problemləri 114
Zəfər Cəfərov, Rauf Həsənov, Rüfət Talıbov
AzScienceNet şəbəkəsində İnternet-resurslardan istifadənin monitorinqi problemləri 117
Yadigar İmamverdiyev
Uşaqların internet təhlükəsizliyində big data texnologiyalarindan istifadə məsələləri 122
Sabirə Allahverdiyeva
Crypto USBToken və SmartSD-də e-imzanın tətbiqi 124
Həbib Abbasov
Вопросы применения методов машинного обучения для решения проблем информационной безопасности 127
Ядигар Имамвердиев, Людмила Сухостат

Botnetlər aşkarlanması üsullarının icmalı 131
Yadigar İmamverdiyev, Gülnarə Qarayeva
"Big data" təhlükəsizliyi və həlli yolları 137
Tural Yunusov
Роль big data в обеспечении безопасности детей в интернете 140
Назакет Меликова
SEKSİYA # 3. Big Data Tətbiqləri 144
Böyük verilənlərin insan resurslarının idarə olunması məsələlərinin həllində tətbiqi, imkanları, problemləri 145
Məsumə Məmmədova, Zərifə Cəbrayılova
Superkompüter texnologiyaları: mövcud vəziyyəti və inkişaf perspektivləri 150
Rəşid Ələkbərov, Tural Mustafayev, Məmmədrəsul Yaqubov
Elektron elmin elmi verilənlər problemləri haqqında 154
Təhmasib Fətəliyev
Texnoparkların innovasiya fəaliyyətinin qiymətləndirilməsində big data texnologiyalarının tətbiqi 157
Ələvsət Əliyev, Roza Şahverdiyeva, Vüsalə Abbasova
Big data neft-qaz sənayesində qərar qəbul etmənin əsas komponenti kimi 162
Məkrufə Hacırəhimova
Ali təhsildə big data texnologiyası tətbiqinin üstünlükləri 166
Havar Məmmədov, İqbal Babayev, Zəfər Cəfərov
E-dövlət sistemində "big data": imkanlar və perspektivlər 169
Fərhad Yusifov
Hava məkanında münaqişə vəziyyətinin müəyyənləşdirilməsi sistemlərində big data texnologiyalarından istifadə 172
Nadir Ağayev, Nəsrəddin İsgəndərov
Big data – to address the business Issues 176
Yashar Hajiyev, Ali Şaintaş
Biznes sahəsində big data-nın tətbiqi perspektivləri 179
Rasim Mahmudov
Böyük verilənlərin saxlanılma problemləri və mövcud həllər 183
Məkrufə Hacırəhimova, Mərziyyə İsmayılova
Применение технологий big data в медицине 187
Гюляра Мурадова
Big data в электронном образовании: большие возможности и большие риски 191
Фирудин Агаев, Гюлара Мамедова

İnkişaf etmiş ölkələrin “big data” təşəbbüsləri 194
Məkrufə Hacırəhimova, Aybəniz Əliyeva

Təbii ehtiyatlardan səmərəli istifadə edilməsi proseslərinin böyük verilənlər əsasında təhlili məsələləri 198
Əlövsət Əliyev, Leyla Əkbərova, Məleykə İbişova

Verilmiş predmet sahəsində terminoloji informasiya sisteminin işlənməsi 202
Əfruz Qurbanova

Proqram mühəndisliyində big data problemləri 205
Tamilla Bayramova

Böyük verilənlər bazasının internet-reklam problemlərinin həllində rolu nədədir? 208
Kəmalə Haşımova

Böyük verilənlərin emalı texnologiyaları innovativ informasiya iqtisadiyyatının inkişafında yeni mərhələ kimi 210
Əlövsət Əliyev, Roza Şahverdiyeva, Elza Musayeva

Virtual məkanda onomastik vahidlərin transliterasiyası 215
Yedgar Cəfərov

Ovuc izləri əsasında insanın identifikasiya olunmasında big datanın rolu 219
Günel Aslanova

Seksiya # 1
Big Data Analitika

Big Data Konsepsiyası və Aktual Elmi-Nəzəri Problemlər

Rasim Əliquliyev[1], Məkrufə Hacırəhimova[2], Aybəniz Əliyeva[3]
[1,2,3]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*rasim@science.az*, [2]*makrufa@science.az*, [3]*aybeniz63z@rambler.ru*

***Xülasə* — Məqalədə “Big Data” konsepsiyası nəzərdən keçirilmişdir. Böyük verilənlərin tətbiqi sahəsində meydana çıxan aktual elmi-nəzəri problemlər təhlil olunmuşdur.**

Açar sözlər— big data, böyük verilənlərin analitikası, vizuallaşdırma, audio analiz, video analiz

I. GİRİŞ

XXI əsrin əvvəllərindən başlayaraq texnika və texnologiyaların köməyi ilə yaradılan rəqəmsal verilənlər hər il həndəsi silsilə ilə artmaqdadır [1]. Belə ki, bəşəriyyətin mövcudluğundan 2003-cü ilə qədərki dövrdə dünyada cəmi 5 ekzabayt məlumat generasiya olunmuşdursa, 2015-ci ildə 8,1 zetabayt, növbəti hər il dünyada informasiyanın həcminin 40% artaraq 2020-ci ildə 44 zetabayta çatacağı proqnozlaşdırılır [2]. Nəticə etibarı ilə verilənlərin emalı, saxlanılması və istifadəsində yeni eranı əks etdirən “böyük verilənlər” (ing. big data) fenomeni meydana çıxmışdır [3]. Bu yeni termin həcm və mürəkkəblik baxımından mövcud idarəetmə metodları və intellektual analiz vasitələri ilə emal oluna bilməyən verilənləri təyin etmək üçündür.

Böyük verilənlər (BV) cəmiyyətin iqtisadi inkişafına təkan verməklə yanaşı elmi ictimaiyyəti bir çox problemlərlə üz-üzə qoymuş, yeni tədqiqat paradiqması yaranmışdır [3]. BV-də bir çox elmi sahələrdə daha böyük nailiyyətlər və yeni elmi kəşflər, müəssisələrdə yüksək mənfəət əldə etmək üçün böyük potensial vardır. Bu potensialı reallaşdırmaq üçün əvvəlcə bir çox texniki, texnoloji problemlər həll olunmalıdır. Bir sıra problemlər də mövcuddur ki, onların həllində yeni elmi baxışlar, yanaşmalar, modelləşmə, riyazi metodlar, optimallaşdırma üsulları və s. tələb olunur. Böyük verilənlərin xarakterik xüsusiyyətlərindən (*böyük həcm, yüksək sürət, müxtəliflik*) irəli gələn problemlərin əsas istiqamətlərini isə miqyaslanma, saxlanma, ilkin emal, idarə etmə, analitika, analizin nəticələrinin interpretasiyası – vizuallaşdırma, təhlükəsizlik və s. təşkil edir. Bu problemlər tədqiqatçılardan real vaxtda böyük verilənlərdən lazım olan informasiyanı, biliyi, məlumatı əldə etmək üçün daha effektiv statistik və analiz üsulları, miqyaslılığı təmin edən arxitektur yanaşmalar, vahid infrastruktur və alətlər tələb edir. Bu mənada böyük verilənlərin elmi-nəzəri problemlərinin tədqiqi olduqca vacibdir və aktualdır.

II. BİG DATA KONSEPSİYASI

“Big data” termini ilk dəfə 1998-ci ildə Silicon Graphics şirkətinin kompüter elmləri üzrə mütəxəssisi Con Meşi tərəfindən istifadə edilmişdir. Bu terminə bir qədər sonra, 2000-ci ildə, Pennsilvania universitetinin professoru, “big data” terminin əsas araşdırıcılarından olan Fransis Dieboldun akademik mühitdə dərc olunan tədqiqatında rast gəlinir [4]. Onun sonrakı tədqiqatlarında bu terminin bir qədər də möhkəmləndiyini, artıq onun bir hadisə, bir fenomen deyil, ciddi tədqiqat istiqamətinə çevrildiyindən bəhs olunur. Lakin, termin professor Klifford Linçin (*Clifford Lynch*) 2008-ci ildə “Nature” jurnalının xüsusi nömrəsində nəşr olunmuş “Big Data: How do your data grow?” adlı məqaləsindən sonra akademik mühitdə populyarlıq qazanmışdır [1]. Vahid tərifin olmaması “big data” fenomeninin anlanmasında qeyri-müəyyənlik yaradır. Elm, sənaye və biznes kimi maraqlı tərəflərin müxtəlif yanaşmaları mövcuddur [5,6.]. McKinsey Global İnstitutunun hesabatında bu termin informasiyanın toplanması, saxlanılması, idarə edilməsi, axtarışı və analizinin tipik verilənlər bazasının imkanları xaricində olan verilənlər dəsti kimi xarakterizə olunur [1,7].

Tətbiq olunduğu sahələrdən asılı olmayaraq böyük verilənləri təsvir etmək üçün ümumi xarakteristikalar mövcuddur. Bu xarakteristikalar böyük verilənlərin əsas problemlərini özündə əks etdirməklə üç əsas qrupa bölünür: həcm (*volume*), sürət (*velocity*) və müxtəliflik (*variety*). İngilis dilli mənbələrdə bunu «3V» də adlandırırlar. Bu parametrlərin konvergensiyası böyük verilənləri təyin etməyə və digər verilənlərdən fərqləndirməyə kömək edir. Bu model ilk dəfə 2001-ci ildə D.Laney tərəfindən verilmişdir [1]. O, “big data” terminini istifadə etməsə də, elektron kommersiyada bir tendensiyanı: verilənlərin idarə edilməsinin daha vacib, daha çətin olacağını əvvəlcədən xəbər vermiş və daha sonra verilənlərin idarə edilməsində verilənlərin ölçüsünü, ötürülmə sürətini və müxtəlifliyini əsas problem kimi təyin etmişdir. Bu xarakteristikalar isə ümumilikdə “big data” texnologiyalarının əsas konsepsiyasını təşkil edir. Bu konsepsiya çox böyük sürətlə və müxtəlif mənbələrdən toplanan çox böyük həcmdə verilənləri daha səmərəli istifadə etmək, saxlamaq, analiz edərək ondan daha qiymətli informasiyanı əldə etmək ideyasını özündə əks etdirir. [1].

Həcm (volume). Həcm böyük verilənlərin ən əsas xarakteristikasıdır. Hazırda BV-nin miqyası terabaytlardan zetabaytlara qədər həcm ilə xarakterizə olunur [1]. Həcm problemi ilk növbədə saxlama problemi yaradır ki, bu da genişmiqyaslı saxlama və paylanmış emal tələb edir.

Sürət (velocity). Burada iki hal nəzərdə tutulur. Birinci, yeni verilənlər böyük sürətlə generasiya olunur, mövcudlar yenilənir və toplanır. İkincisi, həcm artdıqca, emal üçün də çox böyük sürət tələb olunur. Sürət zaman problemi kimi dəyərləndirilir və mövcud emal texnologiyalarının verilənləri real vaxtda analiz etmək imkanına malik olması ilə izah olunur [1].

Müxtəliflik (variety). Müxtəliflik BV-nin təbii özəlliklərindəndir. Məlumatların əksəriyyəti müxtəlif mənbələrdən (e-poçt, sosial şəbəkələr, veb-saytlar, sensorlar və s.) müxtəlif formatlarda daxil olur və müxtəlif indeksləşmə sxemi istifadə olunur. Bunları sadəcə olaraq, bir araya yığmaq və birgə emal etmək və analiz üçün uyğun şəklə salmaq çox çətin olur [1].

IBM verilənlərin həqiqiliyini əsas götürərək 4-cü "V" (*veracity*), Oracle isə BV-nin dəyərini qeyd etməklə 5-ci "V" (*value*)-ni daxil etmişdir [1]. Son zamanlar mütəxəssislər tərəfindən "v"-lərin sayı artmaqdadır.

III. BİG DATANIN ELMİ-NƏZƏRİ PROBLEMLƏRİ

Bu gün "Big data" adlandırılan informasiya bolluğu həqiqətən də mövcuddur. Ancaq bu insanlarda məyusluq deyil, tam əksinə, ona təbii xammal, resurs kimi baxmaq lazımdır. Çünki bu xam verilənlərdə elmi kəşflərə səbəb ola biləcək dərin bilik toplanmışdır. Ancaq bu resursdan maksimum istifadə etməklə cəmiyyət və biznes sahəsində dəyər yaratmaq üçün yeni nəsil analitik texnologiyalara ehtiyac vardır. Bu baxımdan, böyük verilənlər mövzusu istər elmi tədqiqatçılar, istər dövlət qurumlarında qərar qəbul edən şəxslər və siyasətçilər, istərsə də biznes cəmiyyətləri və nümayəndələri tərəfindən böyük diqqət çəkmişdir. Qeyd edildiyi kimi böyük verilənlər mövzusu yeni tədqiqat istiqamətinə çevrilmişdir. Nüfuzlu beynəlxalq təşkilatlar, elmi qurumlar tərəfindən çox böyük həcmdə informasiyanın emalının müxtəlif aspektlərinə həsr olunmuş çoxsaylı konfranslar (IEEE International Congress on Big Data, International Conference on Data Science and Advanced Analytics, IEEE International Conference on Big Data and Cloud Computing, IEEE International Conference on Cloud and Big Data Computing, IEEE International Conference on Multimedia Big Data və s.), simpoziumlar (IEEE International Symposium on Big Data & Visual Analytics, IEEE Symposium on Large Data Analysis and Visualization, IEEE/ACM International Symposium on Big Data Computing və s.), seminarlar (Workshop on Big Data Analytics in CPS: Enabling the Move From IoT to Real-time Control), forumlar keçirilməkdədir. Bu tədbirlərin əsas müzakirə mövzularını: böyük verilənlərin arxitekturu (*Big Data Architecture*), idarə edilməsi (*big data management*), modelləşdirilməsi (*Big Data Modeling*), analitikası (*Big Data Analytics*), alətləri (*Big Data Toolkits*), açıq platformalar (*Big Data Open Platforms*), "*Big Data"*xidmət kimi (*Big Data As a Service*), biznesin səmərəli idarə edilməsi (*Big Data in Business Performance Management*), e-dövlətdə və cəmiyyətdə böyük verilənlərin analitikası (*Big Data Analytics in e-Government and Society*), vizuallaşdırma (*Visualization*), təhlükəsizlik (*security*), böyük verilənlər üçün alqoritmlər və s. təşkil edir [1].

BV-nin problemlərinin həlli sahəsində dünyanın tanınmış elmi mərkəzləri (The Big Data Institute - University of Oxford, The Big Data Institute - University College London, Data Science İnstitute – İmperial College London, Cambridge Big Data University of Washington eScience Institute, Institute for Big Data Analytics - Dalhousie University, Institute for Data Science - University of Rochester, Big Data University – IBM) böyük verilənlərin arxitekturası, toplanması və emalı, saxlanılması, analitikası, təhlükəsizliyi, vizuallaşdırılması və s. istiqamətlərdə elmi-tədqiqat işləri aparırlar [1].

Elmi və populyar jurnalların: Nature (2008), Science (2011), Computer (2013) və s. xüsusi nömrələri bu mövzuya həsr olunmuşdur. Son illərdə işıq üzü görmüş böyük verilənlərin elmi-nəzəri problemlərini işıqlandıran "Journal of Big Data", "International Journal of Big Data" (IJBD), International Journal of Big Data Intelligence (IJBDI), Big Data Research, Big Data & Society, IEEE Transactions on Big Data və s. kimi akademik jurnallar nəşr olunmaqdadır.

Böyük verilənlərin analizi problemləri

Özündə son dərəcə faydalı informasiyanı daşıyan, adi relyasiya bazalarının emal edə bilmədiyi yüzlərlə terabayt və ekzabayt həcmində mətn, təsvirlər, audio-video və s. tip strukturlaşdırılmamış informasiyanın toplanması və idarə edilməsi, saxlanması, təhlükəsizliyi, axtarışı, analizi (*analitik hesabatların generasiyası və vizuallaşdırılması, proqnozlaşdırma*) və s. kimi məsələlərin həllində yeni elmi yanaşmalar, daha mükəmməl analiz üsulları tələb olunur [8].

BV-nin analizi onun əsas problemlərindəndir. Bu problem BV-nin xarakteristikaları, mövcud analiz model və metodları və verilənlərin emal sistemlərinin məhdudiyyətləri ilə bağlıdır [8]. Verilənlərin həcminin kəskin surətdə artması və real zamanda onların analizinə olan tələbat BV-nin ən əsas problemlərindən sayılan böyük verilənlərin analitikasının (*Big Data Analytics*) yaranmasına gətirib çıxardı. "Big Data Analytics" daha böyük və mürəkkəb massivlərə tətbiq edildiyindən kəşf edən analitika (*Discovery Analytics*) və izah edən analitika (*Exploratory Analytics*) terminlərindən də istifadə edilir. Necə adlandırılmasından asılı olmayaraq, mahiyyət eynidir – qərar qəbul edən şəxsləri müxtəlif proseslər haqqında məlumatlarla təmin edən əks əlaqəni yaratmaq [1].

Ümumiyyətlə, BV-dən ideyaların çıxarılması prosesi verilənlərin idarə edilməsi və analizi kimi iki alt prosesdən ibarətdir Məlumdur ki, analiz müxtəlif parametrlər, xarakteristikalar, hadisələr və s. arasındakı korrelyasiyanı tapmağa, təsnifatlandırma və analitik hesabatlar və bunun

əsasında proqnozların verilməsinə imkan verir. Bu aspektdən müasir texnologiyalar verilənlərdəki informasiyanın yeni biliklərə çevrilməsinə və ya biliklərin əldə edilməsinə imkan verməlidirlər. BV-in saxlanması, emalı və analizi üçün böyük hesablama gücü, miqyaslılığı təmin edən arxitektur yanaşmalar, vahid infrastruktur tələb olunur [1].

Əlbəttə verilənlərin çox da sadə olmayan analiz üsulları mövcuddur, lakin mətn, video və audio kimi strukturlaşdırılmamış informasiyaların analizi ciddi problemdir.

Mətn analizi (*text mining*) – "data mining" sinfindən olan üsulların köməyi ilə təbii dildə mətnlərdən əvvəlcədən məlum olmayan əlaqələri və korrelyasiyaları aşkarlamaqla bilikləri çıxarır [9]. Təsnifatlandırma (*classification*) və klasterləşdirmə (*clustering*), informasiyanın çıxarılması (*information extraction*), avtomatik referatlaşdırma (summarization) isə text mining-in həll etdiyi əsas problemlərdir. Mahiyyət etibarı ilə *text mining* mətnləri daha dərin analiz etmək üçün linqivistik, statistik üsulların, informasiya axtarışı, həmçinin maşın təlimi alqoritmlərini istifadə edir [10,11].

Video analitika – video axınının monitorinqi, analizi və onlardan faydalı informasiyanın çıxarılmasını özündə birləşdirir. Video analitika verilənlərin digər analiz növləri ilə müqayisədə başlanğıc mərhələsində olmasına baxmayaraq, onların real-vaxt rejimində emalı üçün iki yanaşma mövcuddur: "*server-based* " və "*edge-based"* [10].

Video məlumatların analizi hazırda tədqiqatçılar qarşısında duran çox çətin məsələ olaraq mövcuddur. Burada əsas problem kadrların tezliyi və təsvirlərin dəqiqliyinin azalması nəticəsində informasiya itkisinin yaranmasıdır 10].

Audio analitika – strukturlaşmamış audio verilənlərin analizi və onlardan faydalı məlumatın çıxarılması üsuludur. Bu üsullar müştərilərə göstərilən xidmətlərin keyfiyyətini yüksəltməyə, məxfilik və təhlükəsizlik kimi müxtəlif məsələlərin yerinə yetirilməsinə nəzarət etməyə və s. kömək edir. Onun əsas problemləri isə nitqin tanınması, küy və s. ilə bağlıdır 10].

Sosial media analitika – sosial media kanallarının strukturlaşdırılmış və strukurlaşdırılmamış verilənlərinin analizini həyata keçirir. Sosial medialarda məzmunlar çox vaxt böyük həcmli, küylü və dinamik olur. Ona görə də mətn, audio və video analitikada qeyd olunan problemlər, həmçinin, BV-nin ötürülməsi onun əsas problemlərindəndir [10,12].

Proqnostik analitika – tarixi və cari məlumatlar əsasında gələcək nəticələri proqnozlaşdırmağa imkan verir və müxtəlif statistik metodlara əsaslanır. Lakin böyük verilənlərin xarakterik xüsusiyyətlərindən irəli gələn verilənlərin müxtəlifliyi, küyün yaranması, yanlış korrelyasiya və təsadüfi endogenlik kimi problemlər ənənəvi statistik metodların tətbiqinə əngəl törədir və BV üçün yeni statistik metodların işlənilməsini tələb edir [3,10,13,14].

Müxtəlif tip verilənlərin inteqrasiyası və həcmi, miqyaslanması, təhlükəsizliyi, uyğunsuzluğu və s. analitikanın əsas problemləridir [15].

A. Böyük Verilənlərin Arxitekturası

Hal-hazırda BV-nin analitikası üçün geniş və hamı tərəfindən qəbul edilmiş arxitektura yoxdur. Böyük verilənlərin arxitekturasının əsas funksional komponemtlərinə faydalı verilənlərin çıxarılması (Data Extraction), verilənlər axınının emailı (Stream Processing), informasiyanın çıxarılması (Information Extraction), verilənlərin keyfiyyətinin/qeyri-müəyyənliyin idarə olunması (Manage Data Quality /Uncertainty), verilənlərin inteqrasiyası (Data Integration), verilənlərin analizi (Data Analysis), verilənlərin paylanması (Data Distribution), verilənlərin saxlanması (Data Storage), meta verilənlərin idarə olunması (Metadata Management), verilənlərin həyat dövrünün idarə olunması (Data Lifecycle Management), məxfilik (Privacy) daxildir [16].

BV-nin emal sisteminin enerji səmərəliliyinin qiymətləndirilməsi və optimallaşdırılması böyük elmi-təqiqat problemlərindəndir [8].

B. Böyük Verilənlərin emalı və saxlanması problemləri

Verilənlərin idarə edilməsinə informasiyanın əldə edilməsi, saxlanması, analiz üçün hazırlanması və emal prosesləri daxildir.

BV-nin emalında aşağıdakı yanaşmalar mövcuddur:

Package: paket emalı; **Stream Computing:** Real-vaxt rejimində daim yenilənən verilənlərin analitik emalını təmin edir və proqnozlar tərtib etmək, daha sürətli analiz və qərar qəbul etmək imkanı verir [15]; **Data-intensive computing:** terabayt və petabaytlarla verilənlərin emalında paralel hesablamaların aparılmasına yönəlmişdir. Petabytlarla verilənlər massivlərinin emalının zəruriliyi "data–intensive computing" yanaşmasını ortaya çıxarmışdır ki, bu "hesablamalar deyil, verilənlər daha əsas sərvətdir" anlamına gəlir [15,17,18].

BV emalı və saxlanması üsullarını, saxlanma qurgularını, verilənlərin saxlanması arxitekturasını, verilənlərə giriş mexanizmlərini köklü surətdə dəyişmişdir. Saxlanma qurğuları böyük həcmli verilənlərin əlyetərliyini və operativ analizini təmin etmək imkanına malik olmalıdır [15]. Hazırda saxlama məsələsinin həllində informasiyanın qurğular arasında miqrasıyasını həyata keçirmək imkanına malik olan DAS (Direct-Attach-Storage), NAS (Network Attached Storage), SAN (Storage Area Networks), HSM (Hierarchical Storage Management), ILM (Information Life-cycle Management) kimi texnologiyalar geniş istifadə olunur [1,8]. Lakin irimiqyaslı paylanmış sistemlər səviyyəsinə gəldikdə isə verilənlərin bu saxlanma arxitrkturalarının çatışmamazlıqları və məhdudiyyətləri ortaya çıxır. Son zamanlar isə saxlama qurğularının yaddaş tutumunun artması, çoxsaylı kompüterlərin (server, kompüter və s.) hesablama və yaddaş resurslarının klasterləşməsi və virtualllaşdırılmasını həyata

keçirməklə, verilənlərin emalı və yadda saxlanılmasına xidmət edən “grid” və “cloud computing” texnologiyalarının tətbiqi saxlama sahəsindəki problemləri demək olar ki, aradan qaldıra bilmişdir [1, 15]. Bulud hesablamaları (**cloud computing)** saxlanma, böyük hesablamaların aparılmasında son dərəcə müvəffəqiyyətli yanaşmalardandır. Digər tərəfdən bulud saxlaması verilənlərin tamlığına nəzarət baxımından verilənlərin təhlükəsizliyi problemi yaradır [15].

C. Böyük Verilənlərin Vizuallaşması Problemləri

BV-nin analizində əsas məsələlərdən biri də nəticələrin təqdim olunması – vizuallaşdırılmasıdır**.** Böyük ölçüyə və həcmə malik böyük verilənlərin vizuallaşdırılması xüsusilə çətindir. BV-nin xarakteriktikaları nəzərə alınmaqla onların vizuallaşması zamanı yaranan əsas problemlər aşağıdakılardır [19]:

- vizual küy (*visual noise*);
- böyük təsvirin anlaşılması (*larde image reception)*;
- informasiya itkisi (*information loss*);
- yüksək məhsuldarlıq tələbləri (*high perfomance requirements*);
- təsvirlərin dəyişməsinin yüksək sürəti (high rate of image change) və s.

Bu problemlərin aradan qaldırılması üçün yeni üsulların və texnologiyaların işlənməsi, eləcə də ixtisaslı kadrların olması əsas məsələlərdəndir.

D. Böyük verilənlərin təhlükəsizlik problemləri

BV meydana gəlməsiylə informasiya təhlükəsizliyi baxımından yeni problemlər yarandı. “Big data” texnologiyalarının tətbiqi mövcud təhlükəsizlik modellərinin köhnəldiyini üzə çıxarmışdır. 15 il əvvəl istifadə olunan təhlükəsizlik yanaşması bu gün üçün adekvat deyildir. Eyni zamanda böyük verilənlərin həcm, müxtəliflik və sürət kimi xarakteristikaları isə təhlükəsizlik və gizlilik problemini daha da kəskinləşdirir [1]. Geniş miqyaslı “bulud” infrastrukturu, verilənlərin mənbələrinin müxtəlifliyi, axın şəklində informasiyanın toplanması və böyük həcmdə informasiyanın “buludlarda” miqrasiyası təhlükəsizlik sistemlərinnin “zəif” cəhətlərini üzə çıxarmışdır. Belə ki, BV-nin genişlənməsi ilə ənənəvi təhlükəsizlik mexanizmləri kifayət deyildir. Eyni zamanda verilənlərin axını çox çevik və sürətli təhkükəsizlik həlləri tələb edir [15,17]. Böyük verilənlərin tətbiqiylə bağlı verilənlərin təhlükəsizlik problemləri bir neçə səbəblərə görə daha da mürəkkəbləşmişdir. Birincisi, böyük verilənlərin ölçüsü mövcud mühafizə yanaşmalarına münasibətdə çöx böyükdür. Bu da öz növbəsində təhlükəsizlik sahəsində işləri bir qədər də çətinləşdirirr. Digər tərəfdən böyük verilənlər paylanmış şəkildə saxlanılır və şəbəkələrdəki təhdidlər bu təhlükələri artıra bilərlər [15].

NƏTİCƏ

Biz böyük verilənlər erasına daxil olmuşuq. BV artıq elmi tədqiqat istiqamətinə çevrilmişdir. Elmi tədqiqatların əsas istiqamətlərini isə böyük verilənlərin konsepsiyasının əsasını təşkil edən, BV-nin həcm, sürət və mürəkkəblik kimi təbii xüsusiyyətlərindən doğan elmi tutumlu məsələlər təşkil edir. Məsələlərin həllində data mining sinfinin üsulları (assosiativ qaydalar, reqressiya, klassifikasiya, klasterizasiya və s.), süni neyron şəbəkələr, maşın təlimi, optimallaşma, o cümlədən genetik alqoritmlər, obrazların tanınması, prediktiv analitika, imitasiya modelləşməsi, statistik analiz, analitik verilənlərin vizuallaşdırılması kimi metodların tətbiqi vacibdir. Bunun üçün yüzlərlə terabayt və ekzabayt həcmində mətn, təsvirlər, audio-video və s. tip strukturlaşdırılmamış informasiyanın emalı və idarə edilməsi, saxlanması, təhlükəsizliyi, axtarışı, analizi və s. bağlı problemlər öz elmi həllini tapmalıdır.

ƏDƏBİYYAT

[1] Əliquliyev R.M., Hacırəhimova M. Ş. "Big Data" phenomenon: problemlər və imkanlar, İnformasiya texnologiyaları problemləri, 2014, №2, səh. 3-16.

[2] The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe/

[3] Fan J., Han F. & Liu H. Challenges of Big Data analysis, National Science Review, 2014, vol. 1, no 2, pp. 293–314.

[4] Diebold, F.X. "'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting", Discussion Read to the Eighth World Congress of the Econometric Society, 2000.

[5] NIST Big Data Working Group (NBD-WG). http://bigdatawg.nist.gov/home.php.

[6] Madden S. From Databases to Big Data, IEEE Internet Computing, 2012, vol.16, no 3, pp.4–6.

[7] Big data: The next frontier for innovation, competition, and productivity. Analyst report, McKinsey Global Institute, May 2011. http://www.mckinsey.com/

[8] Jina X., Benjamin W. Waha, Chenga X., Wanga Y. Significance and Challenges of Big Data Research, Big Data Research, 2015, Vol. 2, no 2, pp. 59–64.

[9] Alguliev R., Aliguliyev R., Hajirahimova M. Multi-document summarization model based on integer linear programming, Intelligent Control and Automation, 2010, vol.1, no.1, pp.105-111.

[10] Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics TedInternational Journal of Information Management, 2015, vol . 35, pp. 137–144.

[11] Jiang, J. Information extraction from text. In C. C. Aggarwal, & C. Zhai (Eds.), Mining text data (pp. 11–41). United States: Springer, 2012.

[12] Barbier, G., & Liu, H. Data mining in social media. In C. C. Aggarwal (Ed.),Social network data analytics (pp. 327–352). United States: Springer, 2011.

[13] Kambatla K., Kollias G., Kumar V., Grama A. Trends big data analytics, Parallel and Distributed Computing, 2014, vol.74, no.7, pp.2561-2573.

[14] Hajirahimova M. Sh., Aliyeva A.S., Review of statistical analysis methods of high-dimensional data, Eastern-European Journal of Enterprise Technologies, Kharkov, 2015, no 5, pp. 23-30. http://journals.uran.ua/eejet

[15] Philip C.L., Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, 2014, vol. 275, pp.314–347.

[16] Maier M, Towards a Big Data Reference Architecture, fevral, 2013 - www.win.tue.nl/~gfletche/Maier_MSc_thesis.pdf

[17] Labrinidis, A., & Jagadish, H. V. Challenges and opportunities with big data, .Proceedings of the VLDB Endowment, 2012, vol. 5, no12, pp. 2032–2033.

[18] Jagadish H.V. Big Data and Science: Myths and Reality // Big Data Research, 2015, vol. 2, no 2, pp. 49–52.

[19] Gorodov E., Gubarev V., Analytical Rewiew of Data Visalization Methods in Application to Big Data, Journal of Electrical and Computer Engineering, 2013, pp.7 - http://www.hindawi,com

An Efficient İncremental Clustering Algorithm for Very Large Data Sets

Adil M. Bagirov[1], Sona Taheri[2], Julien Ugon[3]
[1,2,3]Faculty of Science and Technology, Federation University Australia, Victoria, Australia
[1]*a.bagirov@federation.edu.au*

***Abstract*— An algorithm for solving the minimum sum of squares clustering problems in large data sets is introduced. This algorithm uses a special structure of the clustering problem such as difference of convex representations of its objective functions. The use of such a structure allows one to design a clustering algorithm which is efficient in large and very large data sets. The proposed algorithm is tested and compared with other clustering algorithms using large real world data sets.**

***Keywords*— Large data sets, cluster analysis, nonsmooth optimization, difference of convex representation.**

I. INTRODUCTION

Clustering is an unsupervised partitioning technique dealing with the problems of organizing a collection of patterns into clusters based on similarity. Most clustering algorithms are based on the hierarchical and partitional approaches. Algorithms based on the hierarchical approach generate a dendrogram representing the nested grouping of patterns and similarity levels at which groupings change [12]. Partitional clustering algorithms find the partition that optimizes a clustering criterion [12]. In this paper we develop a partitional clustering algorithm. More specifically, we develop an algorithm for solving the minimum sum-of-squares clustering (MSSC) problems.

The objective functions of the clustering problems, called cluster functions, can be represented as a difference of convex (DC) functions. There are several papers where the DC representation of the MSSC problems is used to design algorithms. In [10], the truncated codifferential method is applied to solve the MSSC using its DC representation. The branch and bound method was modified for such problems in [16] using their DC representation. In [1] an algorithm based on DC programming and DC Algorithms (DCA) is introduced. In [3], the authors use the hard combinatorial optimization model to formulate MSSC as a DC program and propose an algorithm based on DCA. Such an approach allows one to make simpler and less expensive computations in the resulting DCA. In [2], the DCA and a Gaussian kernel are applied to design an algorithm to solve the MSSC problem. All these algorithms are not always efficient for solving clustering problems in large data sets containing hundreds of thousands and more data points.

In this paper, we design an algorithm for solving the MSSC problem based on its DC representation. Results of numerical experiments on some real world data sets are reported and the proposed algorithm is compared with several other clustering algorithms. It is demonstrated that the proposed algorithm is especially efficient for solving the MSSC problems in very large data sets.

In what follows we denote by IR^n the n-dimensional Euclidean space with the inner product $\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i$ and the associated norm $\|u\| = \langle u, u\rangle^{1/2}, u, v \in IR^n$. $B_\epsilon(x) = \{y \in IR^n : \|y - x\| < \varepsilon\}$ is the open ball centered at x with the radius $\varepsilon > 0$.

II. DC PROGRAMMING APPROACH TO CLUSTERING PROBLEMS

In this section we give a nonsmooth optimization formulation of clustering problems and their DC representations.

Definition 1. *$f : IR^n \to IR$ is called a DC function if there exist convex functions $g, h: IR^n \to IR$ such that:*

$$f(x) = g(x) - h(x), x \in IR^n.$$

Here $g - h$ is called a DC decomposition of f while g and h are DC components of f. A function f is locally DC if for any $x_0 \in IR^n$, there exist $\varepsilon > 0$ such that f is DC on the ball $B_\epsilon(x_0)$. It is well known that every locally DC function is DC [11]. Note that a DC function has infinitely many DC decompositions.

An unconstrained DC program is an optimization problem of the form:

$$minimize f(x) = g(x) - h(x) \text{ subject to } x \in IR^n. \quad (1)$$

In cluster analysis we assume that we are given a finite set of points A in the $n-$dimensional space IR^n, that is $A = \{a^1, \dots, a^m\}$, where $ai \in IR^n$, $i = 1, \dots, m$. The hard unconstrained clustering problem is the distribution of the points of the set A into a given number k of disjoint subsets $A^j, j = 1, \dots, k$ such that:

1. $A^j = \emptyset_k$ and $A^j \cap A^l = \emptyset, j, l = 1, \dots, k, j = l.$
2. $A = \bigcup_{j=1}^{k} A^j$

The sets A^j, $j = 1, \ldots, k$ are called clusters and each cluster A^j can be identified by its center $x_j \in IR^n, j = 1, \ldots, k$.

The problem of finding these centers is called the k-clustering (or k-partition) problem. In order to formulate the clustering problem one needs to define the similarity (or dissimilarity) measure. Here the similarity measure is defined using the L_2 norm:

$$d_2(x, a) = \sum_{i=1}^{n} (x_i - a_i)^2 .$$

The nonsmooth optimization formulation of the MSSC problem is [7], [9]:

$$minimize\ f_k(x) \text{ subject to } x = (x^1, \ldots, x^k \in IR^{nk}), \quad (2)$$

where

$$f_k(x^1, \ldots, x^k) = \frac{1}{m} \sum_{a \in A} \min_{j=1,\ldots,k} d_2(x^j, a) \quad (3)$$

The objective function f_k in Problem (2) can be expressed as a DC function:

$$f_k(x) = f_{k1}(x) - f_{k2}(x), x = (x^1, \ldots, x^k \in IR^{nk}), \quad (4)$$

Where

$$f_{k1}(x) = \frac{1}{m} \sum_{a \in A} \sum_{j=1}^{k} d_2(x^j, a)$$

$$f_{k2}(x) = \frac{1}{m} \sum_{a \in A} \max_{j=1,\ldots,k} \sum_{s=1, s \neq j}^{k} d_2(x^s, a)$$

Since the function d_2 is convex in x the function f_{k1} as a sum of convex functions is also convex. The function f_{k2} is a sum of maxima of sum of convex functions. Since the sum of convex functions is convex, the functions under maximum are convex. Furthermore, since the maximum of a finite number of convex functions is also convex, the function f_{k2} is a sum of convex functions and therefore it is also convex.

Problem (2) is a global optimization problem, the objective function f_k in this problem has many local minimizers and only its global minimizers provide the best cluster structure of a data set with the least number of clusters. In general, conventional global optimization methods cannot be applied to solve this problem in large data sets. Therefore in such data sets heuristics and deterministic local search algorithms are the only choice. But the success of these algorithms heavily depends on the choice of starting cluster centers and the development of efficient procedures for generating starting clusters centers is crucial for the success of such algorithms. We apply an approach introduced in [15] to find starting cluster centers. This approach involves the solution of the so-called auxiliary clustering problem.

Assume that the solution $x^1, \ldots, x^{k-1}$, $k \geq 2$ to the $(k-1)$-clustering problem is known. Denote by r^a_{k-1} the distance between the data point $a \in A$ and the closest cluster center among $k-1$ centers $x^1, \ldots, x^{k-1}$:

$$r^a_{k-1} = min\{d_2(x^1, a), \ldots, d_2(x^{k-1}, a)\}. \quad (5)$$

The k-th auxiliary cluster function is defined as [5]:

$$\bar{f}_k(y) = \frac{1}{m} \sum_{a \in A} min\{r^a_{k-1}, d_2(y, a)\},\ y \in IR^n. \quad (6)$$

This function is nonsmooth, locally Lipschitz, directionally differentiable and as a sum of minima of convex functions it is, in general, nonconvex. It is obvious that $\bar{f}_k(y) = f_k(x^1, \ldots, x^{k-1}, y), \forall y \in IR^n$.

A problem:

$$\text{minimize } \bar{f}_k(y) \text{ subject to } y \in IR^n \quad (7)$$

is called *the k-th auxiliary clustering problem* [5]. The DC representation of the function $\bar{f}_k$ is as follows:

$$\bar{f}_k(y) = \bar{f}_{k1}(y) - \bar{f}_{k2}(y) \quad (8)$$

Where

$$\bar{f}_{k1}(y) = \frac{1}{m} \sum_{a \in A} \left(r^a_{k-1} + d_2(y, a)\right),$$

$$\bar{f}_{k2}(y) = \frac{1}{m} \sum_{a \in A} max\{r^a_{k-1}, d_2(y, a)\},$$

An algorithm for solving optimization problems for solving both Problems (2) and (7) is described in [8]. This algorithm is based on DC representations of both clustering and auxiliary clustering functions.

III. INCREMENTAL ALGORITHM

In this section we present an incremental algorithm for solving Problems (2) and (7) using the DC approach. An important part of this algorithm is a procedure for finding

starting points for the l-th cluster center where $1 \leq l \leq k$. This procedure was described in detail in [15].

Algorithm 1 An incremental clustering algorithm.

1: (Initialization). Compute the center $x^1 \in IR^n$ of the set A. Set $l := 1$.

2: (Stopping criterion). Set $l := l + 1$. If $l > k$ then stop. The k-partition problem has been solved.

3: (Computation of a set of starting points for the auxiliary clustering problem). Apply the procedure from [15] to find the set $S_1 \subset IR^n$ of starting points for solving the auxiliary clustering problem (7) for $k = l$.

4: (Computation of a set of starting points for the l-th cluster center). Apply the optimization algorithm to solve Problem (7) starting from each point $y \in S_1$. This algorithm generates a set $S_2 \subset IR^n$ of starting points for the l-th cluster center.

5: (Computation of a set of cluster centers). For each $\bar{y} \in S_2$ apply the optimization to solve Problem (2) starting from the point $(x^1, \dots, x^{l-1}, \bar{y})$ and find a solution $(\hat{y}^1, \dots, \hat{y}^l)$. Denote by $S_3 \subset IR^{nl}$ a set of all such solutions.

6: (Computation of the best solution). Compute

$$f_l^{min} = min\{f_l(\hat{y}^1, \dots, \hat{y}^l) : (\hat{y}^1, \dots, \hat{y}^l) \in S_3\}$$

and the collection of cluster centers $(\bar{y}^1, \dots, \bar{y}^l)$ such that $f_l = (\bar{y}^1, \dots, \bar{y}^l) = f_l^{min}$.

7: (Solution to the l-partition problem). Set $x^j := \bar{y}^j, j = 1, \dots, l$ as a solution to the l-th partition problem and go to Step 2.

Since the clustering Algorithm 1 applies the optimization based on the DC representation to solve clustering problems it is called the *DCClust* algorithm. It is easy to see that this algorithm in addition to the k-partition problem solves also all intermediate l-partition problems where $l = 1, \dots, k-1$. Steps 4 and 5 are the most time-consuming steps of this algorithm as the optimization is applied repeatedly.

TABLE 1. THE BRIEF DESCRIPTION OF DATA SETS

Data sets	Number of instances	Number of attributes
Gas Sensor Array Drift	13910	128
Skin Segmentation	245057	3
3D Road Network	434874	3

We compare the DCClust with the following algorithms:

1. The global k-means algorithm (GKM) [14].
2. The Multi-start modified global k-means algorithm (MSMGKM) [15].
3. The version of the Algorithm 1 where the optimization algorithm is replaced by the DCA [4] (MS-DCA).

All these algorithms are based on the incremental approach. The DCClust algorithm contains a special procedure to generate starting cluster centers (Step 3) which is described in detail in [6], [15]. To design the version of Algorithm 1 with the DCA in Steps 4 and 5 the optimization algorithm is replaced by the DCA.

IV. NUMERICAL RESULTS

To test the DCClust algorithm and compare it with other three clustering algorithms numerical experiments with a number of real-world data sets have been carried out. Algorithms were implemented in Fortran 95 and compiled using the *gfortran* compiler. Computational results were obtained on a PC with the CPU Intel(R) Core(TM) i5-3470S 2.90 GHz and RAM 8 GB. Three data sets have been used in numerical experiments. Their brief description is given in Table I. The detailed description can be found in [13]. All data sets contain only numeric features and they do not have missing values.

We computed up to 25 clusters in all data sets. The CPU time used by algorithms is limited to 20 hours. Since all algorithms computes clusters incrementally we present results with the maximum number of clusters obtained by an algorithm during this time limit. Results for cluster function values found by different algorithms are presented in Table II. In this table we use the following notation:

- k is the number of clusters;
- f_{best} (multiplied by the number shown after names of data sets) is the best known value of the cluster function (3) (multiplied by m) for the corresponding number of clusters;
- E_A is the error in % by an algorithm A which is calculated as follows:

$$E_A = \frac{\bar{f} - f_{best}}{f_{best}} \times 100\%$$

where $\bar{f}$ is the value of the clustering function obtained by an algorithm A;

- The sign "-" in tables shows that an algorithm failed to compute clusters in the given time frame.

Results presented in Table II show that all four algorithms are able to find global or near global solutions to clustering

TABLE 2. CLUSTER FUNCTION VALUES OBTAINED BY ALGORITHMS

k	*fbest*	*EDCClust*	*EMS−DCA*	*EMGKM*	*EGKM*
Gas Sensor Array Drift (×10_{13})					
2	79.11857	0.00	0.00	0.00	0.00
3	5.02412	0.00	0.00	0.00	0.00
5	3.22726	0.00	0.00	0.00	0.00
10	1.65524	0.00	0.00	0.00	0.00
12	1.40655	0.00	0.01	0.01	0.00
15	1.13801	0.35	0.00	0.00	0.36
20	0.87916	0.62	0.16	0.00	0.62
25	0.72348	0.47	0.00	0.00	0.16
Skin Segmentation (×10_9)					
2	1.32236	0.00	0.00	0.00	0.00
3	0.89362	0.00	0.00	0.00	0.00
5	0.50203	0.00	0.00	0.00	0.00
10	0.25122	0.00	0.00	0.00	0.00
12	0.21416	0.00	0.55	0.55	0.00
15	0.16964	0.18	0.19	0.18	0.00
20	0.12770	0.14	0.17	0.17	0.00
25	0.10299	0.00	0.00	0.00	0.00
3D Road Network (×10_6)					
2	49.13298	0.00	0.00	0.00	0.00
3	22.77818	0.00	0.00	0.00	0.00
5	8.82574	0.00	0.01	0.00	0.00
10	2.56710	0.00	0.21	-	-
12	1.84976	0.00	0.05	-	-
15	1.27072	0.00	0.26	-	-
20	0.80872	0.00	0.73	-	-
25	0.60334	1.93	0.00	-	-

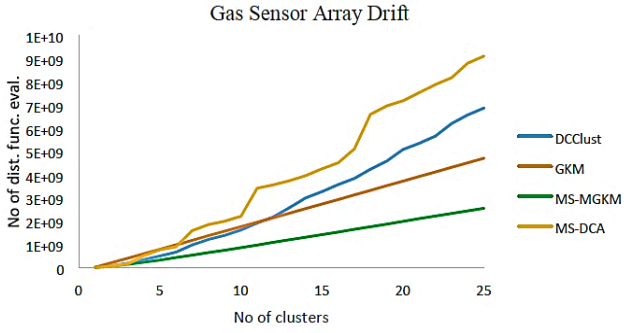
problems in large data sets, however the GKM and MSMGKM algorithms are not efficient for solving clustering problems within a given timeframe in data sets with hundreds of thousands of points.

Figures 1(a)-1(c) illustrate dependence of the number of distance function evaluations (N_d) on the number of clusters for four algorithms in all data sets. These figures demonstrate that the MS-MGKM algorithm requires the least number of distance function evaluations in two data sets. In the 3D Road Network data set this algorithm computed only 6 clusters within a 20 hours timeframe and the number N_d is similar to that of by the MS-DCA and DCClust algorithms. For the GKM algorithm N_d depends linearly on the number of clusters, however this algorithm requires significantly more distance function evaluations than other three algorithms. Comparison of the DCClust and the MS-DCA algorithms shows that the latter algorithm requires more distance function evaluations than the former algorithm.

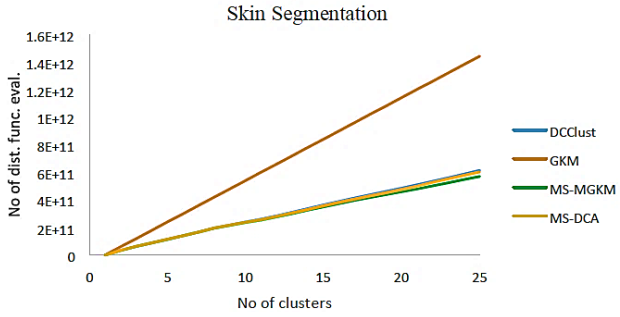
Figures 2(a)-2(c) illustrate dependence of the CPU time on the number of clusters for four algorithms in all data sets. It is obvious that the DCClust algorithm requires least CPU time among all four algorithms.

V. CONCLUSION

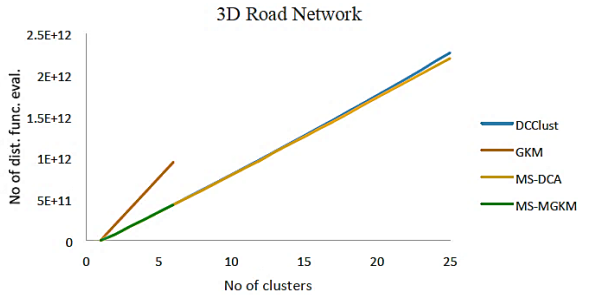
In this paper an algorithm for solving the minimum sum-of-squares clustering problems is designed using their DC representation. The algorithm is an incremental algorithm and computes clusters gradually starting from one cluster which



(a)



(b)



(c)

Figure 1. The number of distance function calculations vs the number of clusters.

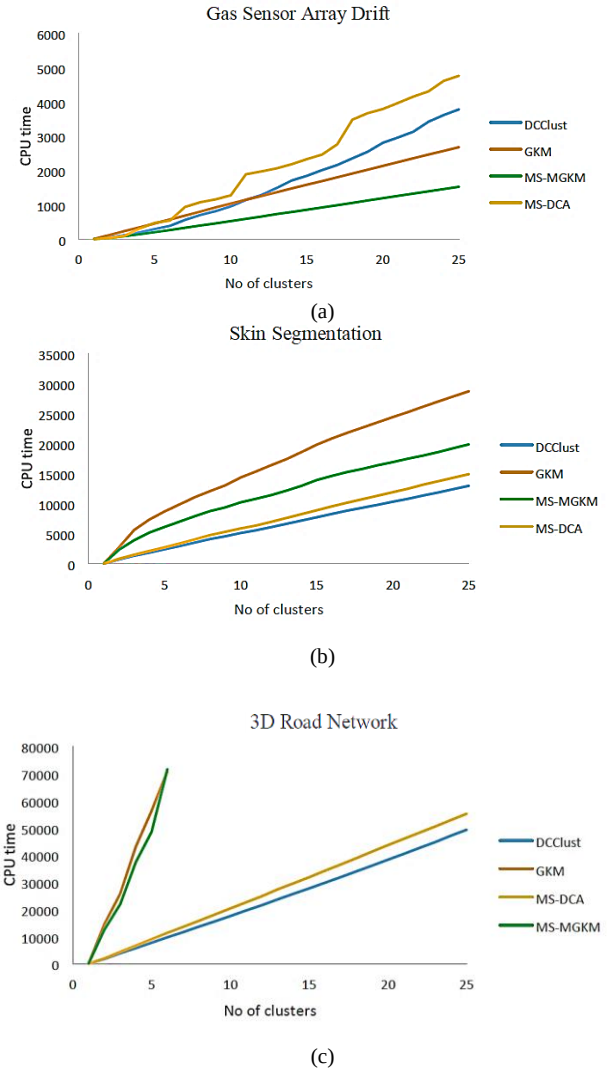


Figure 2. The CPU time vs the number of clusters.

is the whole data set. The proposed algorithm is tested using real world data sets with the number of data points ranging from tens of thousands to hundreds of thousands and compared with other clustering algorithms. Results clearly demonstrate that the use of the DC representation of clustering problems allows one to significantly improve ability of incremental algorithms to solve clustering problems in very large data sets in a reasonable time. Furthermore, the use of nonsmooth optimization algorithms can increase this ability for even larger data sets.

ACKNOWLEDGMENT

This research by Dr. Adil Bagirov and Dr. Sona Taheri was supported under Australian Research Council's Discovery Projects funding scheme (Project No. DP140103213).

REFERENCES

[1] L.T.H. An, M.T. Belghiti, and P.D. Tao. A new efficient algorithm based on DC programming and DCA for clustering. *J. of Global Optim.*, 37(4):593–608, 2007.

[2] L.T.H. An, L.H. Minh, and P.D. Tao. New and efficient DCA based algorithms for minimum sum-of-squares clustering. *Pattern Recognition*, 47:388–401, 2014.

[3] L.T.H. An and P.D. Tao. Minimum sum-of-squares clustering by DC programming and DCA. In D.-S. Huang, K.-H. Jo, H.-H. Lee, H.-J. Kang, and V. Bevilacqua, editors, *Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence, ICIC 2009, LNAI-5755*, pages 327–340. Springer-Verlag, Berlin, Heidelberg.

[4] L.T.H. An and P.D. Tao. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46, 2005.

[5] A.M. Bagirov. Modified global k-means algorithm for minimum sumof-squares clustering problems. *Pattern Recognition*, 41(10):3192–3199, 2008.

[6] A.M. Bagirov, B. Ordin, G. Ozturk, and A.E. Xavier. An incremental clustering algorithm based on hyperbolic smoothing. *Computational Optimization and Applications*, 61:219–241, 2015.

[7] A.M. Bagirov, A.M. Rubinov, N.V. Soukhoroukova, and J. Yearwood. Unsupervised and supervised data classification via nonsmooth and global optimization. *Top*, 11:1–93, 2003.

[8] A.M. Bagirov, S. Taheri, and J. Ugon. Nonsmooth dc programming approach to the minimum sum-of-squares clustering problems. *Pattern Recognition, submitted*, 2015.

[9] A.M. Bagirov and J. Yearwood. A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems. *European Journal of Operational Research*, 170(2):578–596, 2006.

[10] V.F. Demyanov, A.M. Bagirov, and A.M. Rubinov. A method of truncated codifferential with application to some problems of cluster analysis. *Journal of Global Optimization*, 23(1):63–80, 2002.

[11] R. Horst and N.V. Thoai. DC programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[13] M. Lichman. UCI machine learning repository, http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences, 2013.

[14] A. Likas, N. Vlassis, and J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003.

[15] B. Ordin and A.M. Bagirov. A heuristic algorithm for solving the minimum sum-of-squares clustering problems. *Journal of Global Optimization*, 61:341–361, 2015.

[16] Hoang Tuy, A.M. Bagirov, and A.M. Rubinov. Clustering via DC optimization. In *Advances in Convex Analysis and Global Optimization*, pages 221–234. Springer, 2001.

Toplu Ulaşımda Akıllı Kart Verilerinin Analizi ile Yolcu Yoğunluğunun Tahmin Edilmesi*

Ahmet Can Diker[1], Efendi Nasiboğlu[2], Elvin Nasibov[3]
[1,2]Dokuz Eylül Üniversitesi, İzmir, Türkiye
[3]Ege Üniversitesi, İzmir, Türkiye
[1]*dikerahmetcan@gmail.com*, [2]*efendi.nasibov@deu.edu.tr*, [3]*elvin.nasibov@netsis.com.tr*

***Özet* — Gelişmiş ve gelişmekte olan şehirlerde nüfusun hızla artması ulaşım sorunlarını da beraberinde getirmektedir. Toplu ulaşımda kullanılan akıllı kartlardan (Smart Card) günlük milyonlarca veri toplanabilmekte ve bu verilerin analizi ile biniş-iniş sayıları tahmini yapılabilmektedir. Çalışmada İzmir şehri model olarak ele alınmış, yolculuk zinciri ve rasgele atama yöntemlerinden bahsedilmiştir. Bu yöntemlerin uygulanması sonucu yolcu yoğunluğunun tahmin edilmesine yönelik hesaplamalar ele alınmıştır.**

Anahtar Kelimeler— akıllı kart, veri analizi; yolculuk zinciri; yolcu yoğunluğu, toplu taşıma/ulaşım.

I. GİRİŞ

Şehir içi ulaşımında kullanılan akıllı kartlar, akıllı şehirlere doğru giden yolda en önemli adımlardan biridir. Gelişmiş şehirlerde her gün toplu ulaşım araçlarında akıllı kartlar ile yapılan işlem (transaction) sayısı milyonları bulmaktadır. Bu verilerin analiz edilmesi kısa ve uzun vadede yolcu alışkanlıklarının belirlenmesi için önemlidir. Depolanan verinin gün geçtikçe artması toplu ulaşımda yolcuların hareket desenlerini çıkarmayı mümkün kılsa da ortaya çıkan büyük verinin analizi de işlem maliyetini artırmakta ve daha etkin çözümler aranmaktadır.

Gelişmiş ve gelişmekte olan şehirlerde nüfusun hızla artması, yerel yönetimleri toplu ulaşım hizmet kalitesini artırmaya zorlamakta ve bunun bir sonucu olarak kısa ve uzun vadeli planlar yapılmaktadır. Toplu ulaşımda akıllı kart verileri 2000'li yılların başından itibaren ulaşım planlama amacıyla kullanılmaya başlanmıştır [1-3]. Bagchi ve arkadaşları, İngiltere'de iki ayrı otobüs firmasına ait akıllı kartların otobüslerde sadece binişlerde kart kullanıldığını belirtmişler ve bu sebeple iniş bilgisinin elde edilmesine ilişkin zorlukları dile getirmişlerdir [1]. Bagchi ve White, diğer bir çalışmada yolculuk alışkanlıklarını akıllı kart verilerinden elde ettiği istatistikleri kullanarak incelemişlerdir [3]. Morency ve arkadaşları Gatineau, Quebec'te 10 aylık biniş verileri için k-ortalamalar kümeleme yöntemini kullanarak yolcu alışkanlıklarını belirlemeye çalışmışlardır [4]. Bu ve benzeri çalışmaların odaklandığı ana nokta, toplu ulaşım sistemlerinin birçoğunda akıllı kartların sadece girişlerde kullanılması sebebiyle inişe ait bir bilginin mevcut olmamasıdır. Dolayısıyla iniş bilgisinin tahmin edilmesi gereklidir. İniş tahmininin yapılması ile biniş, iniş ve aktarma noktalarına (durak, istasyon, ya da iskele) ait istatistikleri değerlendirilerek başlangıç-bitiş (origin-destination) matrisleri oluşturula bilmektedir. İniş bilgisinin tahmininde yöntemlerin birçoğu yolculuk zinciri (trip-chaining) yöntemine dayanır. Yolculuk zinciri yöntemi yolcuların gün içindeki yaptıkları hareketleri ardışık olarak inceler ve bir önceki yolculuğun inişini bir sonraki biniş yerine göre tayin eder. Literatürde yolculuk zinciri mantığına dayanan farklı yöntemlerin geliştirildiği birçok çalışma vardır [1-8]. Zhao ve arkadaşları, raylı sisteme ve otobüse ait biniş verilerinden yolculuk zinciri yöntemi ile iniş tahminini gerçekleştirilmiştir [5]. Trépanier ve arkadaşları 2007 yılında Gatineau, Quebec'te otobüs biniş verilerinde yolculuk zinciri yöntemini veri tabanında programlayarak %66 doğruluk oranı ile iniş bilgisini tahmin etmiştir [6]. Nasiboğlu ve arkadaşları, İzmir toplu ulaşım sisteminde durakların birleşmesi ile oluşan bölgeler için yolculuk zincirini yöntemini kullanarak bölge bazında biniş-iniş yoğunluğunu tahmin etmişlerdir [7].

Munigaza & Palma (2012) Santiago, Chile de büyük bir ulaşım ağında biniş verilerinin yaklaşık %80'inin iniş yer ve zamanını tahmin etmişlerdir [8]. Yolculuk zinciri yöntemi varsayım tabanlı bir yöntemdir ve bu sebeple varsayımlar, ele alınan şehrin toplu ulaşım sisteminin karakterine uygun olarak oluşturulur. Yolculuk zinciri mantığı ile oluşturulmuş yöntemler, her ne kadar birçok şehir için farklı varsayımlar için uygulanmış olsa da bu yöntemlerin direkt olarak uygulanamayacağı durumlarda mevcuttur. İzmir toplu ulaşım sisteminde yolculuk zincirinin uygulanamadığı gün içinde tek biniş ve aynı akıllı kart ile aynı araç içerisinde birden fazla kez işlem yapıldığı durumlar mevcuttur. Bu çalışmada İzmir şehri için yolculuk zinciri yönteminin uygulanması ve bu yöntemin uygulanamadığı durumlar için önceden elde edilmiş hareket desenlerinin kullanılabileceği rasgele atama yöntemi önerilecektir. Ayrıca, önerilecek yöntemler ile tahmin edilen iniş bilgisi ile araç içi yolcu yoğunluğunun tahmin edilmesine yönelik hesaplamalarda ele alınacaktır. Çalışmanın II. Kısımda öncelikle akıllı kart verisine ilişkin bazı bilgiler verilecek ve kalanında yolculuk zinciri yöntemine ilişkin varsayımlara değinilecektir. III. Kısımda rasgele atama yöntemine ilişkin hesaplamalar ve yolcu yoğunluğuna ait hesaplamalar ele alınacaktır. Son kısımda ise yöntemlere ilişkin yorumlar yapılacaktır.

II. YOLCULUK ZİNCİRİ YÖNTEMİ İLE İNİŞ BİLGİSİNİN TAHMİN EDİLMESİ

Çalışmada ele alınan veriler İzmir toplu ulaşımında iki ana veri kaynağından elde edilmektedir. Bunlardan ilki araç, istasyon ve iskele girişlerinde akıllı kartların okutulması

*Bu çalışma TÜBİTAK tarafından 113E535 Nolu araştırma projesi kapsamında desteklenmektedir.

amacıyla kullanılan validatörler; diğeri ise otobüslerde bulunan, otobüslerin konum bilgisinin elde edilmesini sağlayan GPS alıcılarıdır. Akıllı kart verilerinde her bir işlem için kart numarası, ücret tipi (öğrenci/öğretmen, tam, engelli, 60 yaş üstü ve personel), aktarma durumu, biniş zamanı, aracın hat numarası, hattın yönü (Gidiş, Dönüş), araç türü (otobüs, metro ve vapur), aracın ID'si ve şoför ID'si gibi alanlar bulunmaktadır. Akıllı kartlardan elde edilen veriler bir gün için yaklaşık 1.5 milyon işlemden oluşmaktadır. Bunun haricinde GPS alıcılarından da her 12 saniyede bir otobüslerin yerine ilişkin konum verisi alınmaktadır. Bu veriler ise günlük yaklaşık 15 milyon civarındadır. Akıllı karta ait işlemlerde, yolcunun hangi durakta biniş yaptığının bilgisi bulunmamaktadır. Bu durumda biniş verileri ile GPS verileri eşleştirilerek yolcunun hangi durakta biniş yaptığı belirlenmektedir.

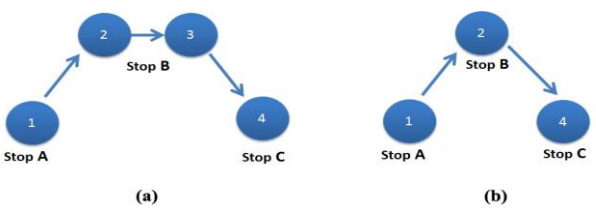
Çalışmada ele alınan toplu ulaşım sisteminde, biniş bilgisinin olması, fakat iniş bilgisinin olmaması sebebiyle iniş tahmininin yapılması gerekmektedir. İniş tahmini yapabilmek için geliştirilen yöntemlerinin temelinde yolcuların ardışık hareketlerini inceleyen yolculuk zinciri yaklaşımı kullanılmaktadır [6-8]. Bu yöntem varsayımlara dayanan bir yöntemdir ve iki ana varsayımdan oluşmaktadır. Bunlardan ilki: "bir önceki yolculuğun iniş durağı, bir sonraki binişin başladığı yerdir" düşüncesidir. İkinci varsayımın temelinde ise "yolcular gün sonunda güne başladıkları yere dönerler" düşüncesi yer almaktadır. Bu varsayımlar İzmir toplu ulaşım sistemi için aşağıdaki şekilde yazılabilir:

Varsayım 1. Yolcular, biniş yaptıkları hattın üzerinde, bir sonraki biniş yaptıkları durağa en yakın durakta inerler.

Varsayım 2. Yolcular, gün sonunda yolculuklarını ilk biniş yaptıkları durakta sonlandırırlar.

Varsayım 1'de sonraki binişe en yakın durağın bulunmasında yürüme mesafesinin kabul edilebilir olması gereklidir. Bu mesafenin kabul edilebilir olması amacıyla literatürde farklı çalışmalarda yürünebilecek mesafe kısıtı; Zhao ve arkadaşları tarafından 400 metre, Trépanier ve arkadaşları tarafından 2000 metre, Munizaga ve Palma tarafından 1000 metre olarak ele alınmıştır [5, 6, 8]. Bizim çalışmada yürünebilecek mesafe kısıtı 1000 metre olarak belirlenmiştir. Çalışmada İzmir toplu ulaşım sisteminde iniş tahminine yönelik bir yöntem önerileceğinden, bu sistemi yansıtan varsayımlar da burada ele alınacaktır. Çünkü, literatürdeki çok az sayıda çalışmada ele alınan 'günde tek biniş durumu' [9, 10] ve 'aynı kart ile aynı seferde birden fazla işlem yapılması durumu' gibi durumlar İzmir toplu ulaşım sisteminde görülmektedir. Günde tek biniş durumunda yolculuk zinciri, yolcunun aynı gün içinde ardışık hareketleri olmaması sebebiyle uygulanmamaktadır.

Şekil 1. Aynı kart ile aynı seferde birden çok işlem yapılması durumuna ilişkin örnek



Aynı kart ile aynı seferde birden çok işlem yapılması ise yolcunun kartını başka yolcular için kullanması ile gerçekleşmektedir. Örneğin bir yolcunun günlük 4 işlemi Şekil 1.a'daki gibi olsun. Bu durumda yolcu ilk binişini Stop A'dan yaptıktan sonra 2. ve 3. binişlerini Stop B'den yapmakta ve son olarak da Stop C'de biniş yaptığı görülmektedir. Yolculuk zinciri mantığına göre bu yolcunun Stop A'dan Stop B'ye, Stop B'den Stop B'ye ve Stop B'den Stop C'ye gittiği söylenebilir. Şekil 1.a'da yolcuya ait 2. biniş işleminin kendi tarafından yapıldığı, 3. binişte ise yolcunun kartını başkası için kullandığı düşünülebilir. Çalışmada bir hattın aynı seferi üzerinde aynı kart ile yapılan işlemlerden ilki yolcunun yolculuk zincirinde kalacak, diğerleri ise her biri ayrı tek binişler olacak şekilde işlem görecektir. Bu durumda, gerçek yolculuk zincirindeki gidişler sırasıyla Stop A'dan Stop B'ye, Stop B'den Stop C'ye şeklinde olacaktır (Şekil 1.b). Yukarıda ele alınan iki durumdaki verilere ait iniş tahmini III. Kısımda anlatılacak rasgele atama yöntemi ile yapılacaktır.

Yolculuk zinciri kullanılmasına rağmen özellikle Varsayım 1'deki yürüme mesafesi kriterinden dolayı iniş durağı tahmin edilemeyen biniş verileri bulunmaktadır. Bu veriler de yukarıda bahsedilen iki durum gibi rasgele atama yöntemi ile değerlendirilecektir.

III. RASGELE ATAMA YÖNTEMİ İLE İNİŞ TAHMİNİ

Yolculuk zinciri ile iniş durağı tahmininin yapılamadığı durumlarda biniş-iniş sayıları dengelenmediği için otobüs içi yoğunluk hakkında bir tahmin yapılamayacaktır. Bu sebeple tahmin edilemeyen veriler için Varsayım 1 ve Varsayım 2 kullanılarak tahmin edilen biniş-iniş bilgilerin kullanılması düşünülmüştür. Varsayım 1 ve Varsayım 2'nin tüm günleri içeren veri seti üzerinde uygulanması sonucunda varsayımlara uyan tüm biniş durağı, biniş yapılan hat ve iniş durağı şeklinde üçlüler olarak elde edilecektir. Bu üçlüler genel yolcu akış tahmini bilgisini barındırmaktadır. Yolcuların benzer akış desenini izleyebileceği düşüncesiyle genel yolcu akış desenine ait özet bilgilerin kullanılacağı rasgele atama yöntemi geliştirilmiştir. Bu yöntemde iniş tahmini yapılamayan bir işlem için; öncelikle biniş hat ve durak bilgisi alınır. Önceden Varsayım 1 ve Varsayım 2 ile elde edilmiş bilgiler içinden ele alınan hattın durağından biniş yapanlara ait ortalama iniş sayıları kullanılarak iniş yapılabilecek durakların olasılıkları bulunur. Bilgisayarda uniform dağılımla üretilen 0 ile 1 arasında rasgele sayıya göre iniş duraklarından hangisinin seçileceği belirlenir.

Aşağıdaki işaretlemeleri kabul edelim. S duraklar kümesinde s herhangi durağı belirtirken ($s \in S$) , L hatlar kümesinde l herhangi bir hattı göstersin ($l \in L$). Herhangi bir hattın üzerindeki durağın sıra numarası aşağıdaki ilişki gibi gösterilsin:

$$O(l,s) = \begin{cases} \geq 1, & s \text{ durağının } l \text{ hattındaki sıra numarası,} \\ 0, & l \text{ hattında } s \text{ durağı bulunmuyor.} \end{cases} \quad (1)$$

$S(l)$; l hattının geçtiği duraklar kümesi olsun. Açıktır ki:

$$S(l) = \{s \in S:\ O(l,s) > 0\} \quad (2)$$

sağlanmaktadır. Burada l hattının toplam durak sayısı $|S(l)|$ ile işaretlenecektir.

s_i durağından l hattıyla s_j durağına yapılan yolculuğu $0 < O(l, s_i) < O(l, s_j)$ koşulu sağlanmak üzere $< l, s_i, s_j >$ üçlüsü (deseni) şeklinde gösterelim. Ayrıca, herhangi bir d gününde l hattıyla s_i durağından s_j durağına yolculuk yapanların toplam sayısını $MC_d(l, s_i, s_j)$ ile gösterelim. Ele alınacak rasgele atama yönteminde, $< l, s_i, s_j >$ günlük toplamlarının ortalaması $\overline{MC}$ kullanılacak ve bu ortalamalar aşağıdaki gibi yazılacaktır:

$$\overline{MC}(l, s_i, s_j) = \frac{\sum_{d=1}^{D} MC_d(l, s_i, s_j)}{D}, \qquad (3)$$

burada $d = 1,2, \dots, D$, D veri analizinde kullanılan gün sayısıdır.

L hattının duraklarını $(s_{l_1}, s_{l_2}, \dots, s_{l_{|S(l)|}})$ olarak işaretleyelim. Hareket sayıları ortalamalarının ($\overline{MC}$) olasılıkları aşağıdaki gibi hesaplanır:

$$P\left(l, s_{l_i}, s_{l_j}\right) = \frac{\overline{MC}(l, s_{l_i}, s_{l_j})}{\sum_{k=i+1}^{|S(l)|} \overline{MC}(l, s_{l_i}, s_{l_k})}, \quad j = i+1, \dots, |S(l)|, \quad (4)$$

burada $P\left(l, s_{l_i}, s_{l_j}\right)$, l hattının i. sıradaki durağından j. sıradaki durağına yapılacak yolculuğun olasılığını belirtmektedir.

Bilgisayarda rasgele sayılar üreticisini kullanarak iniş duraklarını tahmin etmek için $u \epsilon (0,1]$ aralığında uniform üretilen rasgele sayı kullanılacaktır. İniş durağı,

$$\sum_{k=i}^{j} P(l, s_{l_i}, s_{l_k}) < u \leq \sum_{k=i+1}^{j+1} P(l, s_{l_i}, s_{l_k}), \qquad (5)$$

$$j = i, \dots, |S(l)| - 1,$$

koşulunu sağlayan $s^* = s_{l_{j+1}}$ durağı olarak gösterilir. Burada $P(l, s_{l_i}, s_{l_i}) = 0$ olduğu kabul edilmiştir.

Rasgele atama yöntemi uygulanarak iniş durağı önceki yöntemle tahmin edilemeyen veriler tahmin edilebilir. l hattına s_{l_i} durağından binen ve s_{l_j}durağından inenlerin tahmini sayısı $X(l, s_{l_i}, s_{l_j})$ matrisinde saklansın. Bu matris kullanılarak herhangi bir duraktan otobüs çıktığı anda o otobüs içindeki yolcu yoğunluğu tahmin edilebilir.

l hattının s_{l_i} durağından biniş yapan yolcuların sayısı $B(s_{l_i})$ ile gösterilsin ve aşağıdaki gibi hesaplansın:

$$B(s_{l_i}) = \sum_{k=i+1}^{l_{|S(l)|}} X(l, s_{l_i}, s_{l_k}). \qquad (6)$$

burada l hattının son durağında biniş yapılamayacağından $B\left(s_{l_{|S(l)|}}\right) = 0$ olduğu kabul edilir.

l hattının s_{l_i} durağında iniş yapanların toplam sayısı da $A(s_{l_i})$ ile gösterilsin ve aşağıdaki gibi tanımlansın:

$$A(s_{l_i}) = \sum_{k=1}^{i-1} X(l, s_{l_k}, s_{l_i}). \qquad (7)$$

burada l hattının ilk durağında iniş yapılamayacağından $A(s_{l_1}) = 0$ olarak kabul edilmiştir.

Denklem (6) ve Denklem (7) kullanılarak l hattının aracı s_{l_i} durağından ayrıldığında araç içindeki yolcu sayısı aşağıdaki gibi hesaplanır:

$$C(s_{l_i}) = \sum_{k=1}^{i} \left(B(s_{l_i}) - A(s_{l_i})\right). \qquad (8)$$

SONUÇ VE YORUMLAR

Toplu ulaşımda akıllı kart verileri kısa ve uzun vadeli planlar yapmak için oldukça elverişlidir. Özellikle nüfusu kalabalık şehirlerde toplu ulaşım verileri günlük milyonları bulmakta ve bu sebeple büyük hacimli verilerinde analizini zorlaştırmaktadır. 4.5 milyon nüfuslu Türkiye'nin 3. Büyükşehri olan İzmir'in toplu ulaşım sisteminin model olarak ele alındığı bu çalışmada, yolcuların hareket desenlerinin elde edilmesi için yolculuk zinciri ve rasgele atama yöntemi birlikte önerilmiştir. Herhangi bir hattın herhangi bir durağından bir başka durağına seyahat eden yolcuların sayılarının bu yöntemleri ile tahmin edilerek araç içi yolcu yoğunluğunun hesaplanması mümkündür. Araç içi yoğunluğun tahmin edilmesi, ele alınan hattın otobüs kapasite ve sefer sayısı planlamasına yardımcı olacağı düşünülmektedir. Ayrıca yapılan bu ve benzeri çalışmaların, son yıllarda popüler bir konu olan akıllı şehir uygulamalarına da katkı sunacağı düşünülmektedir.

REFERENCES

[1] M. Bagchi, S. D. Gleave, and P. R. White, "Use of public transport smart card data for understanding travel behavior," European Transport Conference, 2003.

[2] M. Bagchi and P. R. White, "What role for smart-card data from bus systems?", Proceedings of the ICE – Municipal Engineer, 157 (1), 39-46, 2004.

[3] M. Bagchi and P. R. White, "The potential of public transport smart card data," Transport Policy, 12, 464-474, 2005.

[4] C. Morency, M. Trépanier, and B. Agard, "Analysing the variability of transit users behaviour with smart card data," Proceedings of the IEEE ITSC 2006.

[5] J. Zhao, A. Rahbee, and N. H. M. Wilson, "Estimating a rail Passenger Trip origin-destination matrix using automatic data collection systems," Computer-Aided Civil and Infrastructure Engineering, 22 (5), 376-387, 2007.

[6] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," Journal of Intelligent Transportation Systems, 11(1), 1-14, 2007.

[7] E. Nasiboğlu, Ü. Kuvvetli, M. Özkılçık, and U. Eliiyi "Origin-Destination matrix generation using smart card data: A case study for Izmir," Proc. IV International Conference "Problems of Cybernetics and Informatics" (PCI'2012), 188-191, 2012.

[8] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile," Transportation Research Part C, 24, 9-18, 2012.

[9] M. A. Munigaza, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smart card," Transportation Research Part C, 44, 70-79, 2014.

[10] W. Wang, J. P. Attanucci, and N. H. M. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," Journal of Public Transportation, 14 (4), 131-150, 2011.

Big Data and Fuzzy based Techniques in Analysis of Social Networks

Ali Abbasov[1], Shahnaz Shahbazova[2]
[1]Institute of Information Technology of ANAS, Baku, Azerbaijan
[2]Azerbaijan Technical University, Baku, Azerbaijan
[1]*abbasov@mincom.gov.az*, [2]*shahbazova@gmail.com*

Abstract - **Social networks – a phenomenon of an early 21st century – is perceived as a source of data generated by users about users themselves and things and dealings related to them. Because of that social networks are treated as an object of many academic, corporate and industrial research activities leading to a better understanding of individuals' behavior, (dis)likes and needs, as well as events and issues important for them. A very active involvement of individuals in social network means that every day millions of new pieces of information is generated. Analysis of this vast amount of available data requires methods and approaches taken from the domain of big data.**
The theory of fuzzy sets and systems, introduced in 1965, provides the researchers with techniques that are able to cope with imprecise information expressed linguistically. This theory constitutes a basis for designing and developing methodologies of processing data that are able to identify and understand views and judgments expressed in a unique, human way – the core of information generated by the users of social networks.
The paper tries to recognize a few important example of extracting value from social network data that can benefit from application of fuzzy set and systems methodology.

***Keywords - fuzzy sets; social network; data-driven manufacturing; statistical approaches; graph based approaches;* summarization; sentiment analysis; event detection.**

I. INTRODUCTION

More and more often social networks, or shall we say data generated by its users, is an object of research activities being conducted by variety of organization and corporations in order to extract information and knowledge about multiple aspects characterizing activities, behavior, as well as interests and likes of users and group of users. We can find multiple examples of corporations and agencies putting enormous effort to analyze and understand data that is generated by actions, interactions, and conversations involving users, as well as by the users' views and opinions on almost everything what happens in their lives and surroundings.

Varieties of techniques are used to process data generated by users in social networks: statistical approaches, graph based approaches, and many others [8], [13], [20], [21], [22], [26]. However, a human nature is present in the social networks. This means that the networks are human-like – full of imprecise relations and connections between individuals, vague terms, groups and individuals with indefinite descriptions and characteristics of interests. It seems that many aspects of social networks resemble the ones of their users.

In the light of these statements, we would like to state that techniques of processing social networks of users and groups should reflect such human facets. These techniques should be based on a human-like methodology, and the theory of fuzzy sets and systems [29] is suitable for such a purpose. Its ability to deal with ambiguous data and facts, its ability to describe things in a human manner, and its ability to handle imprecision and ambiguity make fuzzy sets [11] one of the best tools for analyzing social network [17].

II. ANALYSIS OF SOCIAL NETWORKS: TARGETED AREA

The nature of data generated by users of social networks allows us to identify multiple areas of human life, as well as industrial and corporate activities that could benefit from analyzing social network data. In the following subsections we describe just a few of them in the wide category of issues contributing to improvements in: quality of human life, manufacturing, and corporate visions and goals.

II.I QUALITY OF LIFE

The user's ability to observe and quickly react to different events – would they be positive or not – mean that analysis of data can be an important element of sophisticated and intelligent **Disaster Management** systems. An early detection of disasters, for example earthquakes or floods, would enable quick interventions that increase changes of minimizing or even mitigating the effects of disasters. Tools that allow visualization of social media data would allow for better understanding of dynamics of investigated calamities. They would allow for close, real-time monitoring of disasters and their progress, and also estimating effectiveness of preventing actions and provided help. This is closely related to another very important area that could benefit form data analysis: **Health Care**. In this case, application of big data techniques would have an enormous impact on coping with diseases – better

understanding of their origin, better knowledge about their spreading, as well as better understanding of their impact on human behavior and actions. Systems for detecting diseases and tracking them could be of great importance.

Education is yet another domain where preforming big data analysis on social network data could play an important role. For example, it is already known that analysis of posts generated by pupils and students can lead to detecting bullying and even preventing tragedies caused by such a behavior. Development of systems monitoring and analyzing variety of issues related to ways of study as well as problems and issues encountered by students would mean creation of a better environment for education. Such an environment would increase effectiveness of education systems and contribute to better-educated societies.

Analysis of social network data could also lead to a more effective **Criminal Justice** system. More information extracted from social data that is related to actions violating laws and by-laws would definitely change the way law enforcement agencies work. Ability to detect undesired behaviors as well as presence of dangerous individuals would lead to safer communities.

II.II. MANUFACTURING AND INDUSTRIAL ACTIVITIES

Facebook [32], Twitter [34], Pinterest [33] and other social sites continue to generate and provide an uninterrupted stream of data. Development of methods and approaches that analyze this data from a perspective of the users' opinions on different products and services will allow companies to identify the customers' preferences, as well as their needs and likes. This would lead to a so-called **data-driven manufacturing** – a scenario where existing and potential customers influence what is being manufactured. Analysis of the users' data would also help identifying weaknesses and strengths of manufactured goods. Social media sentiment investigations could determine if users intent to purchase specific products or if they dislike these products. All this would provide insight that can be explored and acted upon. Such analyses could also assess consumers' interest in a product before it is launched.

Introduction of tools and systems analyzing social network data would affect variety of **service and utility** companies. Such systems could change the way companies provide services to the users, and how these systems could react to changes in the demand and needs. For example, General Electric (GE) is about to release a system that uses analysis of social media to support estimation of potential problems in an electrical grid – it is called Grid IQ [35]. Many aspects of individuals' lives are being discussed on the forum of social networks therefore any company that provides services would benefit a lot from social network data. Another interesting aspect is related to a so-called geo-tagging where many aspects of collected data is location sensitive – and this alone will bring valuable information to be analyzed.

II.III CORPORATE VISIONS AND GOALS

Understanding of the users' needs and their attitude to multiple services and products offered by companies are key elements of building corporate strategies and plans for future. Also here, vast amounts of data collected during the users' activities could provide corporations with valuable information. In such areas like **advertising and marketing** any indications regarding the users' moods, attitudes and opinions expressed directly or indirectly would be able to change policies and strategies of companies. The analysis would provide companies with indicators of their social presence, their ranking, and popularity among users.

One of the most interesting and intriguing aspects of analysis of social network data could be related to strategies and operations of **insurance and financial** companies. Information – in a form of opinions, facts or evaluations – that describe the users' behaviors, patterns and rules of actions would be of great interest to insurance companies. They could adopt and customize their policies and offer insurance packages to variety of customers trying to fit their specific needs. Financial companies could use social media data to improve returns on investment – analysis of the users' discussions and posts from the point of few of sentiments and attitudes towards investment and associated with it expectations can lead to identification of trends, patterns and motivations of the users. All this would translate into better strategies and policies of investment polices of companies.

III. DISCUSSIONS AND CONCLUSION

The above section provides a just few examples of areas that can benefit from processing, analyzing and modeling social network data. At this stage we would like to foresee how synthesis of big data methods with fuzzy technologies could contribute to analysis of social networks.

Most of the work dedicated to analysis of network data is targeting structural information of networks. The interconnection and relations between actors have been the main source of information [1], [2], [3], [5], [9], [15]. Some works [4], [6], [7], [14], [19], [23], [27], [28] show an attempt to use additional information that brings different aspects of analysis. It is well known that each of the nodes/actors as well as connections/relation is associated with supplementary information. It seems very reasonable that including that information in analysis procedure and methods would increase the scope of analysis and allow looking and discovering new insights regarding the users and their behavior.

A very important aspect of analysis of network data is to look into posts, comments, tweets, or any form of text generated by the users. They contain enormous amounts of information about users: what is being currently discussed, what are moods among people, what they like, what types of things invoke positive responses, what type of things are perceived as negative. It seems critical to be armed with the ability to find answers to these questions – this would result in detecting

trends, popular and/or important topics, determine users' requirements and expectations regarding variety of items and events. In such a context, we would like to state that application of fuzzy-based approaches will enhance the capabilities of data analytical methods, will enable deeper and more semantic oriented analysis, and what is also essential will make the obtained outcomes more human-like. Fuzzy methods targeting **summarization** of texts [16], **sentiment analysis** [24], [25], **trust** inference and propagation [10], [12], and **event detection** [30][31][18], are just a few examples of research topics that constitute important issues that can be addressed with, and will benefit from application of fuzzy-based techniques and methods.

ACKNOWLEDGMENT

This scientific work was supported by Grant a Science Development Foundation under the President of the Republic of Azerbaijan, Ministry of Communications and High Technologies of the Republic of Azerbaijan in order to support projects aimed at the development of ICT for grants for 2013 for the 2nd joint ICT Contest (SDF-MCI-MQM-2 / ICT-2-2013-7 (13)).

REFERENCES

[1] Abbasov A.M. Information Boom: New Trends and Expectations, SPRİNGER, Series Title: Soft Computing: State of the Art Theory and Novel Applications (Studies in Fuzziness and Soft Computing, Editors: Yager R.R., Abbasov A.M., Reformat M.Z. and Shahbazova Sh.N. Springer; 2012, pp.1-12.

[2] S.P. Borgatti, Centrality and network flow, Social Networks, vol. 27, no. 1, pp. 55-71, 2005.

[3] J. Boyd and M. Everett, Relations, residuals, regular interiors, and relative regular equivalence, Social Networks, vol. 21, no. 2, pp. 147-165, 1999.

[4] M. Brunelli, M. Fedrizzi, and M. Fedrizzi, Fuzzy m-ary adjacency relations in social network analysis: Optimization and consensus evaluation, Information Fusion, vol. 17, pp. 36-45, 2014.

[5] Y. Cao, J. Cao, and M. Li, Distributed Data Distribution Mechanism in Social Network Based on Fuzzy Clustering, F. Sun et al. (eds.), Foundations and Applications of Intelligent Systems, Advances in Intelligent Systems and Computing 213, Springer-Verlag Berlin Heidelberg, pp. 603-620, 2014.

[6] T. Casasús-Estellés and R. R. Yager, Fuzzy Concepts in Small Worlds and the Identification of Leaders in Social Networks, A. Laurent et al. (Eds.): IPMU 2014, Part II, CCIS 443, Springer International Publishing Switzerland, pp. 37-45, 2014.

[7] S. Elkosantini and D. Gien, A dynamic model for the behavior of an operator in a company, in Proceedings of the 12th IFAC Symposium on Information Control Problems in Manufacturing, France, vol. 2, pp. 187-192, 2006.

[8] S. Elkosantini and D. Gien, Human Behavior And Social Network Simulation: Fuzzy Sets/Logic And Agents- Based Approach, in Proceedings of the 2007 Spring Simulation Multi-conference SpringSim '07, vol. 1, 102-109, 2007.

[9] L. Freeman, The Development of Social Network Analysis, Empirical Press, Vancouver, 2006.

[10] F. Hao, G. Min, M. Lin, C. Luo, and L. T. Yang, IEEEMobiFuzzyTrust: An Efficient Fuzzy Trust Inference Mechanism in Mobile Social Networks, IEEE Transactions on Parallel and Distibuted Systems, vol. 25, no. 11, pp. 2944-2955, 2014.

[11] R. Hannemanand and M. Riddle, Introduction to Social Network Methods, University of California, Riverside, 2005.

[12] S. Kim and S.Han, The Method of Inferring Trust in Web-based Social Network using Fuzzy Logic, in Proceedings of the International Workshop on Machine Intelligence Research, pp. 140-144, 2009.

[13] G. J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Pretience Hall, 1995.

[14] P.N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models, Social Networks, vol. 31, 204-213, 2009.

[15] M. J. Lanham, G. P. Morgan, and K. M. Carley, Social Network Modeling and Agent-Based Simulation in Support of Crisis De-Escalation, IEEE Transactions on Systems, Man, and Cybernetics, vol. 44, no. 1, pp. 103-110, 2014.

[16] J. Liu, Fuzzy modularity and fuzzy community structure in networks, The European Physical Journal B, vol. 77, no. 4, pp. 547-557, 2010.

[17] 40 X.H. Liu, Y.T. Li, F.R. Wei, and M. Zhou, Graph-based Multi-tweet Summarization Using Social Signals, in Proceedings of COLING 2012, pp. 1699-1714, 2012.

[18] T. Matuszka, Z. Vincellér, and S. Laki, On a Keyword-Lifecycle Model for Real-time Event Detection in Social Network Data, in Proceedings of 4th IEEE International Conference on Cognitive Infocommunications, pp. 453-458, 2013.

[19] W. Pedrycz and S.-M. Chen (eds.), Social Networks: A Framework of Computational Intelligence, Studies in Computational Intelligence 526, Springer International Publishing Switzerland, 2014.

[20] M. Z. Reformat and R. R. Yager, Using Tagging in Social Networks to Find Groups of Compatible Users, in Proceedings of Join IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), Edmonton, Canada, June 24-28, 2013, pp. 697-702.

[21] Shahbazova Sh.N. Application of Fuzzy Sets for Control of Student Knowledge // Applied and Computational Mathematics, An International Journal, Volume 10, N 1, 2011. Special Issue on Fuzzy Set Theory and Applications, p.195-208.

[22] J. Scott, Social Network Analysis. A Handbook, London, Sage, 2000.

[23] T.A.B. Snijders and C. Baerveldt, A multilevel network study of the effects of delinquent behavior on friendship evolution, Journal of Mathematical Sociology, vol. 27, pp. 123-151, 2003.

[24] T.A.B. Snijders, Statistical Models for Social Networks, Annu. Rev. Soc. 2011.

[25] G. Stakias, M. Psoras, and M. Glykas, Fuzzy Cognitive Maps in Social and Business Network Analysis, M. Glykas (Ed.), Business Process Management, SCI 444, Springer-Verlag Berlin Heidelberg, pp. 241–279, 2013.

[26] D.N. Trung, J. J. Jung, L. A. Vu, and A. Kiss, Towards Modeling Fuzzy Propagation for Sentiment Analysis in Online Social Networks: a Case study on TweetScope, in Proceedings of 4th IEEE International Conference on Cognitive Infocommunications, pp. 331-337, 2013.

[27] D.N. Trung and J.J. Jung, Sentiment Analysis Based on Fuzzy Propagation in Online Social Networks: a Case study on TweetScop, Computer Science and Information Systems, vol. 11, no. 1, pp. 215-228, 2014.

[28] F. Vega-Redondo, Complex Social Networks, Cambridge University Press, 2007.

[29] R.R. Yager, Intelligent Social Network Analysis using Granular Computing, International Journal of Intelligent Systems, vol. 23, pp. 1196-1219, 2008.

[30] R. R. Yager and M. Z. Reformat, Looking for Like-Minded Individuals in Social Networks Using Tagging and Fuzzy Sets, IEEE Transactions on Fuzzy Sets, vol. 21, no. 4, pp. 672-687, 2013.

[31] L.A. Zadeh, Fuzzy Sets, Information and Control. Vol. 8, pp. 338-353, 1965.

[32] https://www.facebook.com (accessed Nov 7, 2015)

[33] https://www.pinterest.com (accessed Nov 7, 2015)

[34] https://twitter.com (accessed Nov 7, 2015)

[35] http://www.gegridsolutions.com/demandopt/Catalog/GridIQ.htm (accessed Nov 7, 2015)

Introduction to Data Science with Apache Zeppelin

Khalid Huseynov
NFLabs, Seoul, South Korea
khalidhnv@nflabs.com

***Abstract* — Apache Zeppelin is a collaborative data analytics and visualization tool for distributed, general-purpose data processing systems such as Apache Spark, Apache Flink, etc. It is a modern web-based tool for the data scientists to collaborate over large-scale data exploration and visualization projects. Its notebook style interpreter enables collaborative analysis sessions sharing between users.**

Keywords — big data; visualization; notebook analytics; open source

I. INTRODUCTION

Data science refers to an interdisciplinary branch of computer science that facilitates extraction of knowledge or insights from large amounts of structured or unstructured data [1, 2]. This process of data discovery can be divided into multiple steps as shown in Figure 1 [3]. Thus, this is an iterative process of data collection, cleaning, analysis, visualization and decision making.

Apache Zeppelin is an open source project for simplifying big data analytics with web-based notebooks that enable interactive data analytics. It has multiple language backends and Apache Spark [4] integration, thus allowing to address various analytic tasks inside notebooks. Moreover it allows interactive visualization and collaboration with ready insights into data. Thus, Apache Zeppelin can cover whole data discovery flow inside a single Zeppelin notebook.

Apache Zeppelin is currently an incubating project under Apache Software Foundation (ASF) [5] with all the following copyright, development process, and community decision making implications. It has more than 70 contributors and its first release of 0.5.0 version happened in July of 2015. Although in an early stage, Apache Zeppelin is growing its user base from industry as well as research communities.

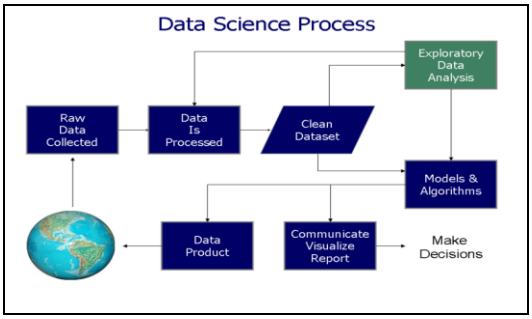


Figure 1 Data Science Process

Supported by NFLabs – Seoul, South Korea (http://www.nflabs.com)

Furthermore, multiple existing solutions address this problem in various ways. For example, Jupyter (IPython) notebook [6] is one of the existing notebook-based solutions that implements python interface with additional computational features known altogether as IPython. Another project known as Spark notebook [7] is forked from Scala notebook project [8] and almost entirely refactored for Massive Dataset Analysis using Apache Spark.

II. ARCHİTECTURE

A. *Multiple Backends*

Apache Zeppelin has pluggable backend architecture in terms of its interpreters. New interpreters can be implemented and plugged in via its *Interpreter* interface. Figure 2 showcases some of the existing interpreters supported by Zeppelin.

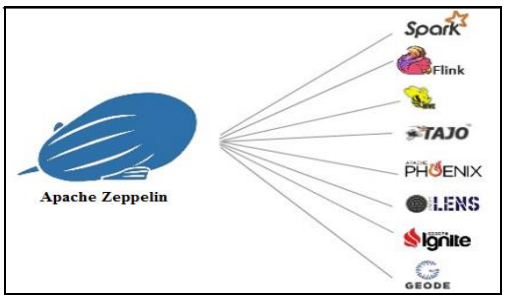


Figure 2 Various backend interpreters

Thus all of the backend frameworks implemented by interpreters can be used inside Zeppelin. Moreover, one Zeppelin notebook may contain different paragraphs each using its own language/interpreter.

B. *Display System*

Since Zeppelin notebooks are web-based, Zeppelin has its *webapp* front-end and back-end server. This architecture is shown in Figure 3.

Zeppelin *webapp* client communicates with back-end server using HTTP REST and *websocket* APIs. Further, server communicates with the interpreter processes. For example, Spark interpreter is just one of the interpreter processes running. According to clients query, server calls corresponding interpreter with a query to interpret the code. Note that interpreter initialization is lazy and isn't started unless client submitted query for it.

Currently, Zeppelin *webapp* front-end supports visualization by the following four different means: pure text, html, tables, and *angular* objects. Angular objects allow Zeppelin to build interactive visualizations inside the notebook.

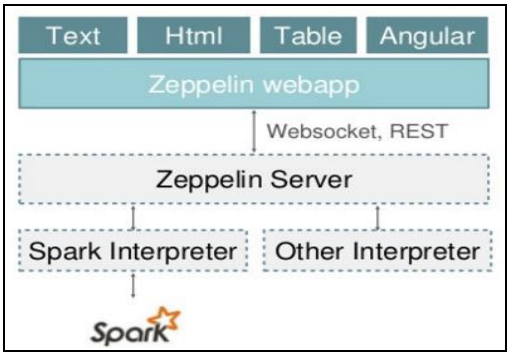


Figure 3 Apache Zeppelin Architecture

III. ZEPPELİN DEMO ANALYTİCS

In this section we will get into real analytics showcase using Apache Zeppelin. After installation [9], Zeppelin includes default showcase notebook with analytics on banking dataset [10].

A. *Prepare Data*

In Zeppelin we can download data using simple *shell* script as if typing in *terminal.* This is realized by the means of *shell* interpreter. Thus, the following script in Figure 4 can download our dataset.

```
%sh
rm ~/bank.zip
rm -rf  ~/data
cd ~
wget http://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip
mkdir data
unzip bank.zip -d data
rm bank.zip

hadoop fs -put  ~/data/bank-full.csv .
hadoop fs -ls -h bank-full.csv
```

Figure 4 Prepare data

The first line with *%sh* implies that shell interpreter is going to be used. Lines 2~3 clean up folder in case previous data exists there. Then we can download data using *wget* command followed by unzipping and putting data into HDFS [11]. Now our data is ready for further processing.

B. *Loading Data into Spark*

Once our dataset is in HDFS we can access it and load into Spark specific RDD format [12]. Figure 5 represents the script written in default Scala Spark interpreter in Zeppelin.

The script starts with creation of Spark context. Further, the input file is mentioned and Bank class representing an object from dataset is created. Then according to structure of Bank class we can load whole dataset using similar template. As you can see, every element of the dataset has five fields: age, job, marital status, education, balance. Finally, this table is registered under "bank" name.

```
//import sys.process._
// sc is an existing SparkContext.
val sqlContext = new org.apache.spark.sql.SQLContext(sc)


val bankText = sc.textFile("bank-full.csv")

case class Bank(age: Integer, job: String, marital: String, education: String, balance: Integer)

val bank = bankText.map(s => s.split(";")).filter(s => s(0) != "\"age\"").map(
    s => Bank(s(0).toInt,
            s(1).replaceAll("\"", ""),
            s(2).replaceAll("\"", ""),
            s(3).replaceAll("\"", ""),
            s(5).replaceAll("\"", "").toInt
        )
)
// toDF() works only in spark 1.3.0.
// For spark 1.1.x and spark 1.2.x,
// use below instead:
// bank.registerTempTable("bank"
bank.toDF().registerTempTable("bank")
```

Figure 5 Loading Data into Spark

C. *Quering Data*

Once we loaded dataset into structured table, we can do various kinds of queries using SQL interpreter of Zeppelin. Suppose we want to get statistics about bank users who are under 30. SQL query from Figure 6 can be used in that case.

```
%sql
select age, count(1) value
from bank
where age < 30
group by age
order by age
```

Figure 6 Query 1 - distribution under 30

First line with *%sql* refers to the usage of SQL interpreter. The results of this query are shown in Figure 7.

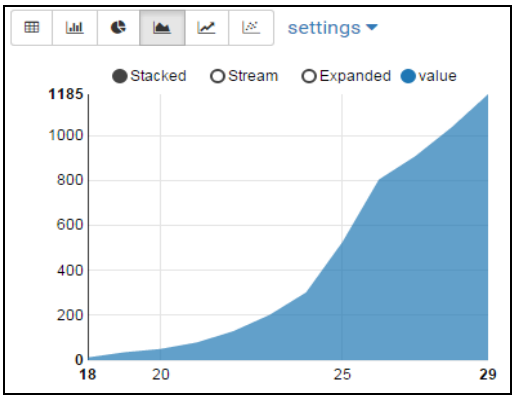


Figure 7 Results of Query 1.

This is *graph* view of the results. Zeppelin also provides *table*, *histogram*, *pie chart* as well as other views. This view is pluggable, thus new types of visualizations can be added by the users.

Next we can query users by *age* and *marital* status as shown in Figure 8.

```
%sql
select age, count(1) value
from bank
where marital="${marital=single,single|divorced|married
}"
group by age
order by age
```

marital

single

Figure 8 Query 2 - by age and marital status

As you can see, the field for marital status is obtained from dropdown menu box. The results of this query are shown in Figure 9.

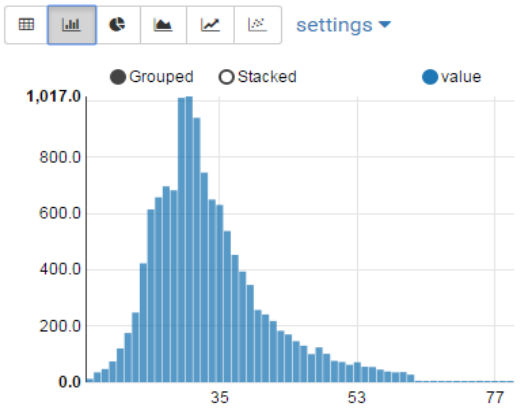


Figure 9 Results of Query 2

These results are shown in *histogram* view. Note that the view can interactively change using *settings* menu. For more insights into Zeppelin workflow you can install it from [9] or visit sharing platform for it discussed in next section.

IV. ZEPPELİNHUB - ZEPPELİN COLLABORATİON PLATFORM

Once you have Zeppelin notebooks, you may need to share or collaborate on them with other parties. For this purpose, ZeppelinHub [13] platform is created. It’s currently in private beta testing state and accepting user applications for beta stage. The collaboration on analytic results is specifically useful inside large companies for decision making in marketing or future company development strategies.

CONCLUSİON AND FUTURE WORK

After going through the data discovery workflow with Apache Zeppelin, we saw Zeppelin addressing all the stages of data discovery process inside a single notebook. Apache Zeppelin supports various back-end interpreters with various front-end visualizations. Moreover, ZeppelinHub lets Zeppelin notebooks to be shared and collaborated on.

REFERENCES

[1] V. Dhar. "Data science and prediction". Communications of the ACM 56 (12): 64. 2013. Vol.56, №12, pp.64, 2013

[2] Jeff Leek. "The key word in "Data Science" is not Data, it is Science". Simply Statistics. 2013

[3] Wiki reference: https://en.wikipedia.org/wiki/Data_science

[4] Apache Spark: http://spark.apache.org/

[5] Apache Software Foundation: http://www.apache.org/

[6] Jupyter (IPython) notebook: http://ipython.org/notebook.html

[7] Spark notebook: https://github.com/andypetrella/spark-notebook

[8] Scala notebook: https://github.com/Bridgewater/scala-notebook

[9] Apache Zeppelin: https://zeppelin.incubator.apache.org/download.html

[10] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

[11] Shvachko, et al. "The hadoop distributed file system." Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE, 2010.

[12] Zaharia, et al. "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing." Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012.

[13] ZeppelinHub – Zeppelin notebooks collaboration platform: https://www.zeppelinhub.com/

Об Одной Модели Анализа Данных Большого Сетевого Трафика

Рамиз Шыхалиев
Институт Информационных Технологий НАНА, Баку, Азербайджан
ramiz@science.az

***Аннотация* — сегодня сложность конфигурации компьютерных сетей (КС) продолжает расти. Кроме того, объем трафика, передаваемого по этим сетям, также увеличивается. При этом анализ трафика является перспективным методом для обеспечения эффективности работы и безопасности КС. В статье исследованы возможности применения Big Data-технологий для анализа больших сетевых трафиков. В результате анализа существующих методов анализа Big Data предложена модель анализа большого сетевого трафика КС.**

Ключевые слова — компьютернве сети, мониторинг, большой сетевой трафик, анализ сетевого трафика, Big Data- технологии, методы анализа Big Data.

I. ВВЕДЕНИЕ

Сегодня компьютерные сети играют фундаментальную роль в социально-экономической инфраструктуре общества. Вместе с тем их масштаб и сложность постоянно растут. А также растет и количество сервисов, предоставляемых КС, особенно в сети Интернет, которая стала очень популярной. Широко стали использоваться CDN (Content Delivery Networks) и cloud-сервисы, а также значительно вырос объем потокового видео и т.д. В результате этого намного вырос объем общего трафика КС. В таких условиях для обеспечения эффективности работы и безопасности КС необходим постоянный мониторинг. Потому что постоянный мониторинг даст необходимую информацию, которая позволит планировать и управлять КС, а также обеспечить их безопасность.

Обычно мониторинг трафика КС осуществляется централизованно, что намного усложняет их постоянный мониторинг. Это прежде всего связано с необходимостью анализа очень большого объема гетерогенного и высокоскоростного трафика, а также постоянного обновления данных мониторинга. При этом анализ слишком большого объема данных требует соответствующих ресурсов хранения и вычисления. Однако, несмотря на высокую производительность традиционных методов и приложений анализа данных, они становятся бесполезными.

Целью статьи является решение указанной проблемы с помощью технологий Big Data. Термином Big Data обозначаются большие и сложные наборы данных, которые трудно обрабатывать с помощью традиционных методов или приложений [1]. Big Data определяется тремя характеристиками данных: объемом, многообразием и скоростью, и когда их значения становятся очень большими, то современные методы и приложения не могут справиться с хранением и обработкой данных. В этом контексте анализ сетевого трафика является проблемой Big Data. При этом для анализа сетевого трафика требуются высокоэффективные Big Data-методы и приложения, которые позволили бы решить проблему в реальном масштабе времени. Такой поход к анализу сетевого трафика позволил бы заблаговременно предупредить о предстоящих неисправностях и проблемах, а также об угрозах безопасности КС. Вместе с тем использование методов Big Data для анализа сетевого трафика позволит эффективно осуществлять сбор всевозможных данных о сетевом трафике и в полной мере оценить состояние всей КС.

II. СЕТЕВОЙ ТРАФИК КАК BIG DATA

Исследования сетевого трафика показали, что он представляет собой сложный динамический процесс и является суперпозицией многих потоков с множественными взаимосвязанными характеристиками, которые генерируются различными протоколами. Во-первых, это трафики, связанные с управлением КС (например, трафик инициализации клиентов, серверный трафик и т.д.), которые генерируются периодически. Во-вторых, это трафики сетевых сервисов, приложений (например, DNS, FTP, запросы WINS, ARP, сеанс NetBIOS, HTTP, P2P, SMTP, POP3, Telnet и т.д.) и протоколов, которые составляют основную часть сетевого трафика КС [2]. При этом для того, чтобы проанализировать сетевой трафик КС с помощью методов Big Data, необходимо определить, что данные сетевого трафика удовлетворяют характеристикам Big Data. Потому что для анализа не всех данных могут потребоваться методы Big Data и с помощью традиционных методов анализа может быть проведен достаточно эффективный анализ. Так как сегодня нет единого мнения по принципиальному вопросу о том, насколько большими должны быть данные, чтобы квалифицировать их как Big Data. Поэтому прежде чем проанализировать большой сетевой трафик, необходимо определить его характеристики с точки зрения Big Data. То есть, при каких значениях характеристик объема, многообразия и скорости данные сетевого трафика можно

считать Big Data. Это очень важная задача, решение которой позволит создать эффективные Big Data-модели для анализа больших сетевых трафиков, так как определение значений этих характеристик даст возможность выбирать эффективные Big Data-технологии.

Обычно при мониторинге КС для централизованного сбора и анализа потока данных сетевого трафика используются высокопроизводительные серверы с большой памятью. Однако при мониторинге крупных КС, например общегосударственных, КС приходится иметь дело с тера- или петабайтами информации. А также при вирусных заражениях (вспышке сетевых червей) или DDoS (Distributed denial of service) атаке появляется необходимость быстро обработать большой объем данных. В таких случаях, чтобы проанализировать трафик, за короткое время невозможно вычислить статистику трафика из большого потока данных. Для решения этой проблемы, то есть для уменьшения объема постоянно поступающего потока данных трафика, традиционно используется метод выборки или агрегации [3, 4]. Однако при таких подходах необходимо заранее знать характеристики трафика.

III. BIG DATA-МЕТОДЫ АНАЛИЗА ДАННЫХ

Сегодня в мире Big Data-технологии привлекают очень большое внимание и в этой области имеется множество исследований и разработок. К ним можно отнести исследования и разработки в области хранения и обработки данных в большом масштабе, такие, как облачное вычисление (Cloud computing) [5], MapReduce, Hadoop [6], а также методы анализа данных – методы машинного обучения и интеллектуального анализа данных (Data Mining). Например, компании Google, Yahoo, Amazon, Facebook разработали и используют платформы кластерных файловых систем и облачных вычислений. Google разработал модель программирования MapReduce для ранжирования веб-страниц и анализа веб-журналов, которая поддерживает распределенные вычисления и имеет две функции, такие, как отображение (map) и уменьшение (reduce) размеров больших наборов данных до кластеров [7]. В фирме Google работают тысячи машин для MapReduce, чтобы обработать большие наборы веб-данных. После того как Google объявила о разработке модели MapReduce, фирма Yahoo выпустила систему Hadoop [8] для платформы облачных вычислений, которая может легко обрабатывать очень большие файлы с потоковой моделью доступа. А компания Amazon предоставляет сервисы облачных вычислений на основе Hadoop, такие, как Elastic Compute Cloud (EC2) или простой сервис хранения (Simple Storage Service (S3)) [9]. Сегодня Facebook также использует Hadoop для анализа данных веб-журналов социальной сети [10].

Сегодня в литературе имеется ряд работ, посвященных применению указанных Big Data-технологий для мониторинга КС. С помощью этих технологий из огромного количества сетевых данных может быть получена полезная информация, которую раньше без таких технологий невозможно было получить. В работе [11] авторы предлагают метод анализа потока интернет-трафика на основе программного обеспечения MapReduce в рамках платформы облачных вычислений. В работе [12] авторы представляют систему мониторинга сетевого трафика на основе Hadoop, который выполняет IP, TCP, HTTP и NetFlow анализ терабайтов интернет-трафика. В работе [13] автор обсуждает проблемы классификации Big Data-данных с использованием методов геометрического представления-обучения и современных Big Data-технологий. В частности, автор рассматривает вопросы комбинирования методов обучения с учителем, методов представления-обучения, методов постоянного машинного обучения (machine lifelong learning) и Big Data-технологий (например, Hadoop, Hive и Cloud) для решения задач классификации сетевого трафика.

IV. МОДЕЛЬ АНАЛИЗА БОЛЬШОГО СЕТЕВОГО ТРАФИКА

На основе анализа приведенных в предыдущем разделе методов Big Data можно сделать вывод о том, что для решения нашей задачи более подходящей технологией является облачное вычисление. Использование технологии облачного вычисления позволит построить большую и гибкую сетевую инфраструктуру хранения данных, которая может адаптивно изменяться в зависимости от требований обработки Big Data.

В качестве модели анализа большого сетевого трафика предлагается интегрированная модель, которая состоит из трех блоков (рис. 1): системы регистрации (записи) сетевого трафика (Network Traffic Recording System (NTRS); системы хранения облачных вычислений (Cloud Computing Storage System (CCSS)) и системы машинного обучения (Machine Learning System (MLS)). Такая интегрированная модель повысит эффективность обработки большого сетевого трафика.

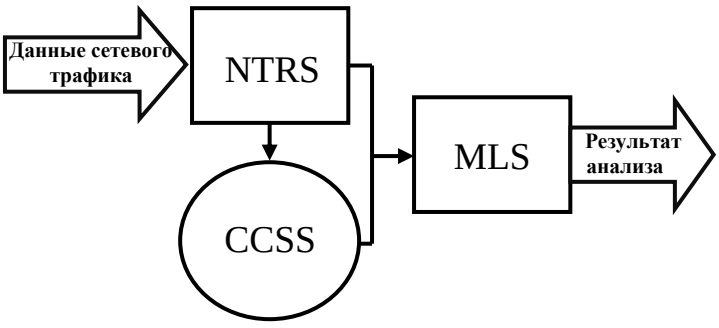


Рис.1. Модель анализа большого сетевого трафика

В предложенной модели блок NTRS, для реализации которого в литературе имеются различные методы [14, 15], получает данные трафика непосредственно из сети или из лог-файлов и записывает их в CCSS. В зависимости от поставленной задачи мониторинга блок MLS анализирует

данные, получаемые из блока CCSS или непосредственно из блока NTRS. В MLS в качестве метода обучения предлагается использовать Machine Lifelong Learning (ML3) [16, 17, 18] метод.

ЗАКЛЮЧЕНИЕ

В условиях постоянного роста объема и сложности трафика очень трудно обеспечить эффективность управления и безопасности КС. При этом постоянный мониторинг, одной из основных функций которого является анализ сетевого трафика, даст необходимую информацию, которая позволит планировать и управлять КС, а также обеспечивать безопасность. Однако в условиях чрезмерно большого объема сетевого трафика решать эту задачу традиционными методами анализа становится все труднее.

Для решения этой задачи в статье предложено использовать технологии Big Data. Для этого была выбрана модель анализа большого сетевого трафика, в которой интегрируются системы регистрации (записи) сетевого трафика (Network Traffic Recording System (NTRS), хранения облачных вычислений (Cloud Computing Storage System (CCSS)) и машинного обучения (Machine Learning System (MLS)).

На наш взгляд, предложенная модель позволит эффективно осуществлять сбор всевозможных данных сетевого трафика, анализировать и в результате этого в полном мере оценивать состояние всей КС.

ЛИТЕРАТУРА

[1] Zikopoulos P. C., Eaton C., et al., Understanding big data – Analytics for enterprise class Hadoop and streaming data, McGraw-Hill, 2012.

[2] Шыхалиев Р.Г. Анализ и классификация сетевого трафика компьютерных сетей // İnformasiya texnologiyaları problemləri, №2, 2010, с. 15–23.

[3] Hohn N. and Veitch D. Inverting sampled traffic, In Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, 2003, pp. 222–233.

[4] Duffield N., Lund C. and Thorup M., Properties and prediction of flow statistics from sampled packet streams, Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment. New York,

[5] Carlin S and Curran K.. Cloud Computing Technologies. International Journal of Cloud Computing and Services Science (IJ-CLOSER) 1.2: 59–65, 2012.

[6] Hadoop, http://hadoop.apache.org/

[7] Dean J. and Ghemawat S., MapReduce: Simplified Data Processing on Large Cluster, OSDI, 2004. NY, USA: ACM, 2002, pp. 159–171.

[8] https://developer.yahoo.com/hadoop/

[9] http://wiki.apache.org/hadoop/AmazonEC2

[10] http://borthakur.com/ftp/hadoopmicrosoft.pdf

[11] Youngseok Lee, Wonchul Kang, Hyeongu Son, An Internet Traffic Analysis Method with MapReduce, Proceedings of the Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP, 19-23 April 2010, pp. 357 – 361.

[12] Yeonhee Lee, and Youngseok Lee, Toward Scalable Internet Traffic Measurement and Analysis with Hadoop, ACM SIGCOMM Computer Communication Review, vol. 43, num. 1, 2013, pp. 6–13.

[13] Suthaharan Shan, Big data classification: problems and challenges in network intrusion prediction with machine learning, ACM SIGMETRICS Performance Evaluation Review, vol. 41 issue 4, 2014, pp. 70–73.

[14] Papadogiannakis A., Polychronakis M., Markatos Evangelos P. Long-term Raw Network Traffic Recording using Fixed-size Storage, Proceedings of the 2010 IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2010, pp. 101–110.

[15] Stefan Kornexl, Vern Paxson, et al., Building a time machine for efficient recording and retrieval of high-volume network traffic, Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, 2005, pp. 23–23.

[16] Qiang Y., Big data, lifelong machine learning and transfer learning, Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 505–506.

[17] Daniel L. Silver and Ryan Poirier, Requirements for Machine Lifelong Learning, Proceedings of the 2nd international work-conference on the Interplay Between Natural and Artificial Computation, Spain, June 18-21, 2007, Part I, pp. 313–319.

[18] Daniel L. Silver, Qiang Y and Lianghao Li, Lifelong Machine Learning Systems: Beyond Learning Algorithms, AAAI Spring Symposium, 2013, pp. 49–55.

Böyük Verilənlər Analizinin Problemləri

Rəna Qasımova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
depart1@iit.ab.az

***Xülasə — Big Data* mənbələrinin əsas hissəsini təşkil edən strukturlaşdırılmamış və qismən strukturlaşdırılmış böyük verilənlərin anazi ciddi problemlərdəndir. Məqalədə böyük verilənlər problemləri araşdırılır və onların analizi üçün bəzi metodlar təhlil olunur.**

Açar sözlər — verilənlər xəzinəsi, bulud, VBİS,, big data, big data analytics, OLAP, Fast Data, Deep Insight.

I. GİRİŞ

Zaman keçdikcə kompüter proqramları özünün çoxcəhətliyi ilə real dünyaya daha da yaxınlaşır. Emal olunmamış verilənlərin həcminin artması onların real vaxtda analizinin zəruriliyi ilə birlikdə *Big Data Analytics* məsələsini effektiv həll etməyə imkan verən alətlərin yaradılması və tətbiqini zəruri edir. Bunun nəticəsidir ki, son zamanlar böyük həcmdə informasiya massivlərinin toplanması, eyni zamanda onların çox böyük sürətlə artması, həm akademik mühitdə, həm də informasiya texnologiyaları (İT) sənayesində daha çox diqqət cəlb etməyə başlamışdır. Miniwatts Marketing Group analitik şirkətinin apardığı statistik hesabata əsasən 2015-ci ilin birinci rübündə 3 milyarddan çox insan, yəni planetin əhalisinin 42,4%-i İnternetə qoşulmuş və mobil rabitə abunəçilərinin sayı 7,1 milyarda çatmışdır. 2020-ci ilə İnternetə qoşulan qurğuların sayının 50 milyard olacağı gözlənilir. 2012-ci ildə dünyada rəqəmsal informasiyanın həcmi 2,7 zetabayt olmuşdur. 2015-ci ildə bu həcmin üç dəfə və növbəti hər il üçün 40% artması proqnozlaşdırılır [1-5].

Rəqəmsal informasiyanın belə sürətlə artımı, verilənlərin müxtəlifliyi, onların ötürülmə sürətinin yüksək artımı çoxsaylı problemlərin yaranmasına səbəb olur. Qeyd edildiyi kimi, artıq böyük verilənlərin (Big Data) saxlanılması, real-vaxt rejimində emalı, analizi və idarə edilməsi problemlər yaratmışdır. Bununla belə, böyük verilənlər (BV) problemi hələ ilkin araşdırmalar səviyyəsindədir, yəni bu sahə hələ də tam olaraq təhlil olunmayıb. Aparılan tədqiqatlar “Big Data” anlayışını, onun mahiyyətini, müxtəlif xarakteristikalarını təsnifləndirməyi, BV-nin mənbələrini, bu texnologiyanın imkanlarını, problemlərini, təhlükəsizlik məsələlərini hərtəfli tədqiq etməyə imkan verir. Tədqiqatlar göstərir ki, BV-in emalı və analizi mükəmməl analitik texnologiyalar və alətlər tələb edir [6-8].

II. BİG DATA ANALİTİKA

Yüzlərlə terabayt və ekzabayt həcmində böyük verilənlərin mövcud metodologiyalarla və ya alətlərlə toplanması, idarə edilməsi, saxlanması və onlardan faydalı informasiyanın əldə edilməsi ciddi problemdir. Həm strukturlaşdırılmış, həm də strukturlaşdırılmamış informasiya ilə işləmək, daha dərin intellektul analiz aparmaq və analizin nəticələrini vizuallaşdırılmaq BV-nin analitikasının əsas məsələlərindəndir.

Verilənlərin həcminin artması və real zamanda onların analizinə olan tələbat BV-nin ən əsas problemlərindən sayılan böyük verilənlərin analitikası (Big Data Analytics) istiqamətinin yaranmasına gətirib çıxarmışdır [9-11]. Big Data Analytics müxtəlif tipli verilənlərdən ibarət olan böyük verilənlər yığımının öyrənilməsi prosesidir. Yəni, böyük verilənlərdən gizli qanunauyğunluqları, naməlum korrelyasiyaları və digər faydalı işgüzar informasiyanı aşkarlamaq üçündür. Analitik verilənlər daha səmərəli marketinqə, yeni gəlir almaq imkanlarına, müştərilərə xidmət keyfiyyətinin yaxşılaşmasına, işin effektivliyinin artmasına, təşkilatların rəqabət və digər biznes üstünlüklərinə gətirib çıxara bilər.

Bu istiqaməti digər tətbiqlərdən fərqləndirən böyük həcm, sürət və mürəkkəblik kimi xarakteristikalar uyğun texnologiyalar tələb edir. Buna görə də, bu gün Big Data Analytics sahəsində əsas istehsalçılar xüsusi proqram-aparat sistemlərini təklif edirlər: SAP HANA, Oracle Big Data Appliance, Oracle Exadata Database Machine, Oracle Exalytics Business Intelligence Machine, Teradata Extreme Performance Appliance, NetApp E-Series Storage Technology, IBM Netezza Data Appliance, EMC Greenplum, HP Converged Infrastructure əsasında Vertica Analytics Platform. Bununla yanaşı kiçik və yeni başlayan şirkətlərin də böyük həcmli verilənləri səmərəli emal edən proqram-aparat alətləri vardır. Onlara Cloudera, DataStax, Northscale, Splunk, Palantir, Factual, Kognitio, Datameer, TellApart, Paraccel, Hortonworks aiddir [12-14].

Verilənlər informasiya almaq üçün emal edilir, bu informasiyanın həcmi o qədər olmalıdır ki, insan onu biliyə çevirə bilsin. Həcm böyük verilənlərin ən əsas xarakteristikasıdır. Həcmindən asılı olaraq BV üç qrupa bölünür [15-18]:

- Tez (sürətli) verilənlər (Fast Data) – onların həcmi terabaytlarla ölçülür;
- Böyük analitika (Big Analytics) – onların həcmi petabaytlarla ölçülür;
- Dərinə nüfuzetmə (Deep Insight) – onların həcmi ekzabaytlarla və zetabaytlarla ölçülür.

Qruplar bir-birindən yalnız verilənlərin həcminə görə deyil, həm də onların keyfiyyətli emalına görə fərqlənirlər. Statistik rəqəmlərdə əks olunan verilənlərin həcmi bir daha

mütəxəssisləri bu sahədə yeni metod və alətlərin işlənməsinə sövq edir.

Fast Data üçün emal yeni biliklərin alınmasını nəzərdə tutmur, onun nəticələri aprior biliklərlə əlaqələndirilir, bu və ya digər proseslərin necə keçməsinə nəzarət edir, eyni zamanda baş verənləri daha yaxşı və ətraflı şəkildə görməyə, hansısa hipotezləri təsdiq və ya inkar etməyə imkan verir. Yalnız mövcud texnologiyaların kiçik bir hissəsi Fast Data məsələlərinin həllinə yarayır. Bunlara xəzinələrlə işləyən bəzi texnologiyaları göstərmək olar, məsələn, Greenplum, Netezza, Oracle Exadata, Teradata, Verica tipli verilənlər bazasının idarəetmə sistemləri (VBİS) və s. Bu texnologiyaların iş sürəti verilənlər həcminin artımı ilə sinxron artır.

Big Analytics vasitələri ilə həll olunan məsələlər kəmiyyət və keyfiyyətcə çox fərqlənirlər. Uyğun texnologiyalar isə yeni biliklərin alınmasına və faydalı məlumatların əldə olunmasına kömək etməklə, verilənlərdə olan informasiyanı yeni biliyə çevirirlər. Başqa sözlə qərarın seçilməsində süni intellekt texnologiyaları nəzərdə tutulmur, analitik sistem “müəllimlə təlim” prinsipi üzrə qurulur və onun bütün analitik potensialı təlim prosesində ona tətbiq edilir. Belə analitikanın klassik nümunələri MATLAB, SAS, Revolution R, Apache Hive, SciPy Apache və Mahout məhsullarıdır [19-21].

Yüksək səviyyə, *Deep Insight* müəllimsiz təlimi (unsupervised learning) və analitikanın müasir metodlarının istifadəsini, həmçinin müxtəlif vizuallaşma üsullarını nəzərdə tutur. Bu səviyyədə aprior bilik və qanunauyğunluqların aşkarlanması mümkündür.

Keyfiyyət baxımından Big Data Analytics proqramları nəinki yeni texnologiyalar, həm də yeni düşüncə tərzi tələb edir. Analitikaya ilkin verilənlərin hazırlanması vasitələrindən, vizuallaşmadan və nəticələri insana təqdim edən digər texnologiyalarından ayrıca baxılır. Hətta The Data Warehousing Institute kimi təşkilatın analitikaya baxışları başqadır. Təşkilatın məlumatına görə hazırda müəssisələrin 38%-i idarəetmə praktikasında Advanced Analytics vasitələrindən istifadə imkanlarını tədqiq edirlər, 50%-i isə yaxın üç il ərzində bunu etməyi nəzərdə tuturlar. Qeyd etmək lazımdır ki, müəssisələrdə bu sahəyə belə maraq biznes sahəsindən çoxlu arqument gətirməklə əsaslandırılır. Belə ki, müəssisələrə yeni şəraitdə daha təkmil idarəetmə sistemi tələb olunur. Onun yaradılmasına əks əlaqə qurulmasından, yəni qərar qəbulunu dəstəkləyən sistemlərdən başlamaq tələb olunur ki, bunun da nəticəsində gələcəkdə qərar qəbulunu avtomatlaşdırmaq mümkün olacaqdır.

Texnoloji obyektlərin avtomatlaşdırılmış idarəetmə sistemlərinin yaradılması problemi heç də yeni deyildir. Bu verilənlər bazası sahəsindəki çoxdan mövcud olan mövzudur ki, paylanmış verilənlər bazası, resursların birgə istifadəsi arxitekturunun yaranması da bu məsələnin həllinə xidmət etmişdir [22].

Analiz üçün yeni vasitələr verilənlər mənbələrinin çox olmasına, həmçinin müxtəlif formatlarda (strukturlaşdırılmış, strukturlaşdırılmamış, qismən strukturlaşdırılmış) olmasına, həm də müxtəlif idekləşmə sxemlərindən (relyasion, çoxölçülü, noSQL) istifadə edilməsinə görə tələb olunur. Belə ki, böyük həcmli verilənləri toplamaq, birgə emal etmək və analiz üçün uyğun şəklə salmaq çox çətin olur.

Ənənəvi üsullarla verilənlərlə işləmək artıq mümkün olmur. Big Data Analytics daha böyük və mürəkkəb massivlərə tətbiq edildiyindən Discovery Analytics və Exploratory Analytics terminlərindən də istifadə edilir. Necə adlandırılmasından asılı olmayaraq, mahiyyət eynidir – qərar qəbul edən şəxsləri müxtəlif proseslər haqqında məlumatlarla təmin edən əks əlaqəni yaratmaq tələb olunur [23-26].

Müasir İT faktorları: böyük verilənlər, analitika və bulud texnologiyalarını bu gün bir-birindən ayrı təsəvvür etmək mümkün deyildir. Çoxsəviyyəli saxlama sistemlərinə artan tələbat, bulud texnologiyalarının real olaraq mövcudluğu həm də BV-nin analitikasına marağı artırmışdır. Bulud texnologiyaları böyük hesablamaların aparılmasında son dərəcə müvəffəqiyyətli yanaşmalardandır, buludlarda saxlamaq və buludlarda hesablamalarsız BV-lə işləmək mümkün deyildir. Burada böyük həcmli rəqəmsal informasiya IaaS (Infrastructure as a service), PaaS (Platform as a service), SaaS (Software as a service) “bulud” xidmətləri vasitəsi ilə mərkəzləşdirilmiş qaydada idarə olunur və saxlanılır. İT sahəsindəki nəhəng şirkətlər yeni nəsil saxlama sistemlərində məhz miqyaslama aspektlərinə və verilənlərin çoxsəviyyəli saxlanmasına böyük diqqət ayırmışdır [27, 28].

Praktika göstərir ki, bu gün analitik məsələlərin yerinə yetirilməsi üçün sistemləri çox yükləmək tələb olunur. Lakin biznes tələb edir ki, bütün servis, əlavələr və verilənlər həmişə əlçatan olmalıdır. Bundan başqa, hazırda analitik tədqiqatların nəticələrinə tələbat çox yüksəkdir. Çünki, savadlı, düzgün və vaxtında keçirilən analitik proseslər bütövlükdə biznes işinin nəticələrini əhəmiyyətli artırmağa imkan verir.

III. BÖYÜK VERİLƏNLƏRİN ANALİZİ PROBLEMLƏRİ

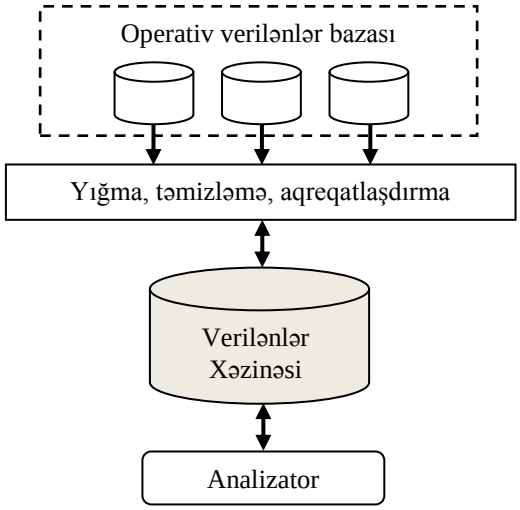
Bu gün BV-nin analizi üçün VX (Data Warehouse) texnologiyalarının yanaşma və metodlarından istifadə etməyə cəhd göstərilir. Bununla yanaşı ənənəvi metodların bəzi xüsusiyyətləri BV-nin emalı spesifikasına zidd ola bilər. Verilənlərin operativ və analitik emalı məsələlərinin əhəmiyyətli fərqi verilənlər bazaları texnologiyalarının inkişafının ilk vaxtlarında yaranmışdı. VX termini 70-ci illərdə Bill İnmon tərəfindən təklif edilmiş, lakin bu texnologiyalara maraq 20 il keçəndən sonra, belə sistemlərə real tələbat yarandıqda və lazımi hesablama gücləri əlçatan olduqda baş vermişdi [29].

VX-də verilənlərin emalı mərhələləri verilənlərin *toplanmasından, təmizləmədən, yükləmədən, analizdən* və nəhayət analizin nəticələrinin *təqdimatından* ibarətdir. Bu mərhələlərin hər birində verilənlər üzərində xüsusi əməliyyatlar yerinə yetirilir. Qeyd etmək lazımdır ki, əgər BV-nin analizi üçün VX texnologiyalarının tətbiqinə cəhd olsa, onda yalnız alqoritmlərin analizinə deyil, həm də verilənlərlə işin bütün mərhələlərinə diqqət yetirmək lazımdır.

Verilənlərin toplanması. VX-nin yaradılmasında məqsəd ayrı-ayrı verilənlərin emal sistemlərində yığılmış verilənlərin təmizlənməsi, razılaşdırılması, inteqrasiya edilməsi və analizdə istifadə üçün rahat formaya gətirilməsidir (şəkil 1).

VX-də nəzərdə tutulur ki, informasiya operativ informasiya bazasından çıxarılır, lazımi şəklə çevrilir, yoxlanılır və yalnız bundan sonra sistemə yüklənilir. Yəni, VX-də verilənlərin hazırlanması texnologiyası bir-biri ilə əlaqəli üç mərhələdən ibarətdir [30]:

1. Verilənlərin toplanması (Data Acquisition).
2. Verilənlərin təmizlənməsi (Data Cleaning).
3. Verilənlərin aqreqatlaşdırılması (Data Consolidation).

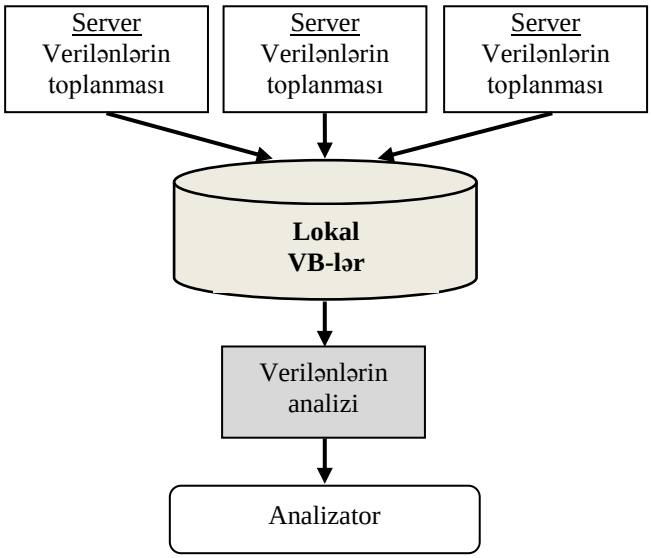


Şəkil 1. Ənənəvi analizdə verilənlərin xəzinəyə toplanması

Sadalanan əməliyyatlar dövriliklə yerinə yetirilir. Lakin BV-lərlə iş zamanı müxtəlif mənbələrdən girişə daxil olan verilənlərin analizini təmin etmək üçün belə dövrilük həmişə mümkün deyil. İnformasiyanın toplanması ilə onun analiz üçün əlçatanlığı arasındakı müddət VX-nin qurulması zamanı əməliyyatların yerinə yetirilməsinə lazım olan vaxtdan kiçik ola bilər. Belə məsələyə nümunə kimi sosial şəbəkələrdə tez yayılan informasiyanın və onun mənbələrinin müəyyən edilməsi, fəal istifadəçilərin təyin edilməsi, neqativ ifadələrin aşkarlanması və ya konfidensial informasiyanın sızması faktlarının aşkarlanması məqsədilə monitorinqi göstərilə bilər [31].

Məlumdur ki, bütün bu hadisələr daha tez aşkarlanmalı və neytrallaşdırılmalıdır. Lakin burada verilənlərin qeyri-formal təsviri movcuddur ki, onların da emalı üçün mətnlərin intellektual analizi alqoritmləri tələb olunur. Məsələn, sosial şəbəkələrin monitorinqi məsələsində girişə istifadəçilərin ifadələri, şərhlər, qoyulmuş qiymətlər, fotoşəkillər və s. verilir. Aydındır ki, məşhur sosial şəbəkələrin monitorinqini çoxlu sayda istifadəçilərlə keçirmək olar. Həm də aydındır ki, istifadəçilərin çox olması və onların yüksək fəallığı səbəbindən bu məsələnin həlli qeyri-formal şəkildə təsvir edilən böyük həcmli verilənlərin yığım və emalını nəzərdə tutur [32]. VX-də verilənlərin ilkin emalı zamanı (məsələn, uyğunsuzluqların axtarışında) xəzinənin əvvəl yığılmış tərkibinin (verilənlərinn) istifadə edilməsi nəzərdə tutulur ki, bu isə BV-lərlə iş zamanı çətin yerinə yetirilir. Yəni, problem ondadır ki, bu verilənlər həmişə paylanır, həm də təkcə analiz üçün deyil, yiğim üçün rahat olur. Məsələn, əgər söhbət telekommunikasiya sistemlərindən gedirsə, onda verilənlər regional serverlərdə toplanır (şəkil 2).

Beləliklə, aparılan təhlillər göstərir ki, ənənəvi VX-lərdə bütün verilənlər həmişə vahid məntiqi blokdan keçir, burada onlar konvertasiya olunur, yoxlanılır, təmizlənir, yüklənir və bu əməliyyatların yerinə yetirilməsi vaxtı nadir hallarda bütün qalan sistemlər üçün vacib olur. Lakin BV-nin emalı zamanı belə vahid blok ola bilməz. Qeyd etmək lazımdır ki, hələ ki, intensiv girişli verilənlər axınına malik məsələlər çox deyil və paylanmış, lakin məntiqi vahid sistem şəklində yığım, təmizləmə, çevirmə və yükləmə bloku reallaşdırmaq mümkündür.



Şəkil 2. Böyük verilənlərin analizi

Böyük verilənlərin analizi. Böyük həcmli verilənləri müxtəlif məqsədlər üçün operativ analiz etmək həmişə idarə edənlər üçün ciddi problem yaradır. Elmi-texniki ədəbiyyatın analizi göstərir ki, tətbiq olunan məhsullar, eləcə də digər analitik sistemlər hələ də tam olaraq bütün tələblərə cavab vermir. Gartner, McKinsey Global İnstitutu, IDC (International Data Corporation) və s. kimi kompaniyaların son tədqiqatlarında BV böyüyən, dinamik inkişaf edən sahə kimi təqdim edilir. McKinsey Global İnstitutunun analitik hesabatında BV-lərə tətbiq edilən analiz metodlarının adları verilmişdir [5]:

– ***Data Mining sinfinin metodları*** - assosiativ qaydaların öyrənilməsi (*association rule learning*), təsnifat, klaster analizi, reqressiya analizi və s.;

– ***onlayn analitik emal (OLAP)***;

– ***kraudsorsinq*** *(crowdsourcing)* - əmək münasibətləri olmadan, ictimai müqavilə əsasında cəlb edilən bir qrup şəxsin gücü ilə verilənlərin emalı;

– ***verilənlərin birləşməsi və inteqrasiyası*** (*data fusion and integration*) - dərin analizi təmin etmək üçün müxtəlif mənbələrdən müxtəlif verilənləri inteqrasiya etməyə imkan verən texniki vasitələrdir. Buna nümunə olaraq siqnalların rəqəmsal emalı və təbii dilin emalını (tonal analiz daxil olmaqla) göstərirlər;

– ***maşin təlimi*** - bura müəllmlə və müəllimsiz təlim, həmçinin kollektiv təlim (Ensemble learning), baza modelləri əsasında kompleks proqnozlar almaq üçün statistik analiz və ya maşın təlimi əsasında qurulan modellərdən istifadə daxildir;

– ***süni neyron şəbəkələri, şəbəkə analizi, optimallaşdırma,*** o cümlədən ***genetik alqoritmlər*;**

– ***obrazların tanınması*;**

– ***prediktiv analitika*;**

– ***imitasiya modelləşdirilməsi*;**

– ***fəza verilənlərinin analizi*** (*Spatial analysis*) - verilənlərdə topoloji, həndəsi və coğrafi informasiyadan istifadə edən metodlar;

– ***statistik analiz,*** buna nümunə A/B testləşməsi və zaman sıralarının analizi göstərilir;

– ***analitik verilənlərin vizuallaşdırılması*** - nəticələrin alınması, sonrakı analiz, ilkin verilənlər kimi istifadə üçün informasiyanın şəkillər, diaqramlar, interaktiv imkanlardan və animasiyadan istifadə etməklə təqdimatı.

BV üzərində çoxölçülü əməliyyat aparmaq üçün klassik B-ağaclardan başlamış bütün mövcud metodların tətbiqi yararlıdır. Bu gün ənənəvi VX-lər verilənlərin analizi alətlərinin təxminən eyni dəstini təqdim edirlər [33-35]. BV-lərdə sadalanan metodları işləməyə imkan verən, operativ yaddaşda terabaytlarla informasiyanı analitik emal edən məhsullar da yaranmışdır, məsələn, SAP HANA, Greenplum Chorus, Oracle Exalytics, Oracle Exadata, Aster Data nCluster. Bundan başqa, Netezza, Teradata, Greenplum və s. şirkətlərinin ənənəvi relyasion verilənlərinin idarə edilməsi sistemi əsasında terabaytlar və ekzabaytlarla verilənləri səmərəli emal edən proqram-aparat alətləri vardır. Belə həllərin potensial imkanlarını başa düşmək üçün həmin metodların əsasında olan alqoritmlərə baxmaq, həmçinin onların BV-nin emalına açar olan mümkün paralel hesablamalar üçün yolları analiz etmək lazımdır. Bu zaman verilənlərin konkret paylanmış emalı texnologiyalarına bağlanmamalı, yalnız BV üçün xarakterik olan əsas parametrləri (sürət, müxtəliflik, həcm və s.) nəzərə almaq lazımdır. İstənilən VBİS-lərin iş sürətinə təsir edən mühüm faktor giriş/çıxış əməliyyatlarının sayı və yaradılmış indekslərin effektivliyidir.

Beləliklə, BV-lərlə işləyərkən məsələnin həllində ilk addım kimi emal olunacaq verilənlərin formatına, analizin tipinə, tətbiq olunan emal metodlarına, eyni zamanda məqsədli sistemin alacağı, yükləyəcəyi, emal və analiz edəcəyi, saxlayacağı verilənlərin mənbələrinə görə təsnifatlandırmaq tövsiyə olunur.

NƏTİCƏ

Verilənlərin həcminin artması, onların real vaxtda analizinin zəruriliyi Big Data Analytics məsələsini effektiv həll etməyə imkan verən alətlərin yaradılması və tətbiqini tələb edir. Big Data Analytics vasitələrini və biznes-analitika texnologiyalarını əsasən İT nəhəngləri istifadə edirlər. Bu da biznes-analitikanın müəssisələrdə istifadəsinin aşağı mədəniyyətindən və biznes-istifadəçi tərəfindən mövcud analiz metodlarının qavrama çətinliyindən irəli gəlir. Bunu nəzərə alaraq bəzi şirkətlərin (məsələn, Information Builders) analitikləri tərəfindən real vaxt rejimində istənilən mənbədən daxil olan verilənlərlə işləməyə imkan verən, istifadədə ən sadə qiymətləndirilən məhsul təklif edirlər.

Big Data sadə düşüncə deyil, gələn texniki inqilabın simvoludur. BV-lərlə analitik iş zəruriliyi İT-sənayesini tamam dəyişəcək yeni proqram və aparat platformalarının yaranmasını stimullaşdırır. Artıq bu gün böyük həcmli verilənlərin analizi üçün ən qabaqcıl metodlar təmin edilir: süni neyron şəbəkələri – bioloji neyron şəbəkəlrinin təşkili və funksiyası prinsipi üzrə qurulmuş modellər, prediktiv analitika, statistika, Natural Language Processing metodları və s. Həmçinin kraudsorsinq, A/B testləşmə, sentiment analiz və ekspertləri cəlbedən metodlardan istifadə edilir.

Nəticələrin vizuallaşdırılması üçün məlum metodlar, məsələn, bulud teqləri və ən yeni Clustergram, History Flow və Spatial Information Flow istifadə edilir. BV texnologiyası tərəfindən Google File System, Cassandra, HBase, Lustre və ZFS paylanmış fayl sistemləri, MapReduce və Hadoop proqram konstruksiyaları və digər qərarlar dəstəklənir. Ekspertlərin qiymətlərinə görə, BV-nin təsiri altında ən çox transformasiyaya istehsalat, səhiyyə, ticarət, inzibati idarəetmə uğrayacaqdır.

ƏDƏBİYYAT

[1] Miniwatts Marketing Group, Worldwide Internet Market Research, www.miniwatts.com

[2] The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe

[3] Worldwide Big Data Technology and Services 2013-2017 Forecast, http://www.idc.com

[4] Data Science Central, The online resource for Big Data practitioners, www.datasciencecentral.com

[5] Big data: The next frontier for innovation, competition, and productivity. Analyst report, McKinsey Global Institute, May 2011. http://www.mckinsey.com

[6] Madden S. From Databases to Big Data, IEEE Internet Computing, 2012, vol.16, No. 3, p. 4-6.

[7] What is big data? - Bringing big data to the enterprise, 2013. http://www-01.ibm.com

[8] Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001. http://blogs.gartner.com

[9] Черняк Леонид. Вычисления с акцентом на данные, Открытые системы, 2008, № 8, с. 36-39.

[10] InfoSphere Platform: Big Data Analytics, 2013, http://www-01.ibm.com/software

[11] Jacobs A. The pathologies of big data, Communications of the ACM. 2009, vol. 52. No. 8, p. 36-44.

[12] Вахрамеев Кирилл. СУБД для анализа Больших Данных, Открытые системы, 2011, № 10, с. 26-29.

[13] Babu S., Herodotou H. Massively Parallel Databases and MapReduce Systems3, Foundations and Trends in Databases, 2013, vol.5, No.1, p. 1-104.

[14] Vignesh Prajapati, Big Data Analytics with R and Hadoop, Publisher: Packt Publishing Ltd, 2013, p. 238.

[15] Черняк Леонид. Большие Данные – новая теория и практика, Открытые системы, 2011, № 10, с. 18-25.

[16] Черняк Леонид. Свежий взгляд на Большие Данные, Открытые системы, 2013, № 7, с. 48-51.

[17] Krish Krishnan. Data Warehousing in the Age of Big Data. 1st Edition, Morgan Kaufmann Publishers Inc. San Francisco, USA, 2013, p. 370.

[18] Билл Фрэнкс. Укрощение больших данных. Как извлекать знания из массивов информации с помощью глубокой аналитики, пер. с англ. Андрея Баранова, М.: Манн, Иванов и Фербер, 2014, 352 с.

[19] Big Data - What Is It? 2013, http://www.sas.com/big-data

[20] MathWorks, http://www.mathworks.com/discovery/big-data-matlab.html

[21] Hadoop Distributed File System. http://hadoop.apache.org/docs

[22] Witt D., Gray J. Parallel Database Systems: The Future of High Performance Database Systems, Communications of the ACM, 1992, vol. 35, No. 6, p. 85-98.

[23] Билл Фрэнкс. Укрощение больших данных. Как извлекать знания из массивов информации с помощью глубокой аналитики, пер. с англ. Андрея Баранова, М.: Манн, Иванов и Фербер, 2014, 352 с.

[24] Селезнев К. Проблемы анализа больших данных, Открытые системы, 2012, №7, с.25-29.

[25] Gudivada V.N., Rao D., Raghavan V.V. NoSQL Systems for Big Data Management, Proceedings of the 2014 IEEE World Congress on Services (SERVICES '14), USA, 2014, p. 190-197.

[26] Майер-Шенбергер Виктор, Кукьер Кеннет. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим, пер. с англ. Инны Гайдюк, М.: Манн, Иванов и Фербер, 2013, 240 с.

[27] Əliquliyev R.M, Hacırəhimova M.Ş. Big data fenomeni: problemlər və imkanlar, İnformasiya Texnologiyaları Problemləri, 2014, №2, səh. 3-16.

[28] Marcos D. Assunção, Rodrigo N., Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya. Big Data computing and clouds, Journal of Parallel and Distributed Computing, 2015, vol. 79, No. C, p. 3-15.

[29] Inmon W. H. “Building the Data Warehouse,” 3rd Edition, John Wiley & Sons, Inc., New York, 2002, p. 412.

[30] Əliquliyev R.M., Qasımova R.T., Ələkbərova İ.Y. Qərarların qəbul edilməsini dəstəkləyən müasir konsepsiyalar haqqında, AMEA-nın Xəbərləri, fizika-riyaziyyat və texnika elmləri seriyası, 2005, №2, s. 70-75.

[31] Tonkin E.L., Pfeiffer H.D. Zombies Walk Among Us: Cross-Platform Data Mining for Event Monitoring, Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW '13), USA, 2013, p. 452-459.

[32] Krishna Kumar K.P., Geethakumari G. A taxonomy for modelling and analysis of diffusion of (mis)information in social networks, International Journal of Communication Networks and Distributed Systems, Switzerland, 2014, vol. 13, No.2, p. 119-143.

[33] Alguliev R.M., Gasimova R.T. Identification of Categorical Registration Data of Domain Names in Data Warehouse Construction Task, Intelligent Control and Automation, 2013, vol. 4, No. 2, p. 227-234.

[34] Alguliev R.M., Gasimova R.T. On a approach for intellectual analysis of registration data of domain names, International Journal of Ubiquitous Computing and Internationalization, 2011, vol. 3, No. 1, p. 27-30.

[35] Ordonez C. Can we analyze big data inside a DBMS?, Proceedings of the sixteenth international workshop on Data warehousing and OLAP (DOLAP '13), USA, 2013, p. 85-92.

Viki-Mühitdə Big Data Problemləri və Onların Həlli Yolları

İradə Ələkbərova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
airada.09@gmail.com

***Xülasə*– Məqalədə viki-mühitdə toplanan və durmadan artan böyük həcmli verilənlərin yaratdığı problemlər analiz edilmişdir. Viki-mühitdəki Big Data ilə işləmək üçün zəruri şərtlər göstərilmiş, problemin həllində bəzi mövcud yanaşmalardan birgə istifadə təklif edilmişdir.**

Açar sözlər– viki-mühit, viki-texnologiya, Vikipediya, big data, xromoqram, intellektual analiz, Map-Reduce.

I. GİRİŞ

Müasir dövrdə Big Data (BD) problemləri İnternet şəbəkəsinin genişlənməsi və yayılması ilə daha da aktual olmuşdur [1]. İnternetin nəhəng layihələrindən olan Facebook, Twitter kimi sosial şəbəkələrdə, Google, Yahoo, Yandex kimi axtarış sistemlərində, Wikipedia, WikiaMapia, WikiTravel kimi açıq ensiklopediyalarda, müxtəlif forumlarda, bloqlarda və s. sistemlərdə hər saniyə artan informasiyanın saxlanması, strukturlaşdırılması və emalı ilə bağlı problemlər bu gün İKT mütəxəssislərini düşündürməkdədir. Yuxarıda adı çəkilən layihələrdə kontent əsasən İnternet istifadəçiləri tərəfindən daxil edildiyi üçün virtual məkanda informasiyanın sürətlə artmasının qarşısını almaq mümkün deyil.

Viki-texnologiyaları ilə idarə olunan Vikipediya virtual ensiklopediyası və onun törəmə layihələri (Vikikitab, Vikimənbə, Commons, Vikinövlər və s.) ümumilikdə, viki-mühit təşkil edir və onun nəhəng məlumat bazası hər saniyə yeni məqalə, xəbər, kitab, foto, audio və video-fayllarla zənginləşir [2, 3]. Məqalədə əsas məqsəd viki-mühitdə BD ilə bağlı problemləri müəyyən etmək, Vikipediyanın verilənlər bazasında (VB) toplanmış məlumatlar əsasında bir çox məsələlərin həllinin mümkünlüyünü və verilənlərin səmərəli emalı üçün düzgün həll yolunu göstərməkdir.

II. VİKİ-MÜHİTDƏ BÖYÜK VERİLƏNLƏRLƏ BAĞLI PROBLEMLƏR

Viki-mühit özündə milyonlarla veb-səhifələri birləşdirir. 2015-cü ilin oktyabr ayına olan məlumata görə, yalnız Vikipediya layihəsi 58 milyondan artıq qeydiyyatdan keçmiş istifadəçiyə və 36 milyona yaxın ensiklopedik məqaləyə [4], 30 milyona yaxın şəkil, audio və video-fayla malikdir [5]. 4 milyondan artıq məqaləyə malik ingilis dilindəki Vikipediyada yalnız məqalələrin ümumi həcmi 1,7 terabaytdır [6]. Viki-mühitdə toplanan bütün kontent isə onlarla terabayt informasiya təşkil edir və bu informasiya durmadan artmaqdadır. Göndərilən sorğuları, yaradılan şablon və bot-proqramları da əlavə etsək, informasiyanın həddən artıq çox olmasını təsəvvür etmək çətin deyil. İstifadəçilərin çox olması və yüksək aktivliyi qoyulan məsələlərin həllində BD texnologiyalarının aktuallığını önə çəkir. BD-də olduğu kimi, viki-mühitdə də problemlər, əsasən verilənlərin həddən artıq çox olması və strukturlaşmaması ilə bağlıdır. Əsas problemlər aşağadakılardır:

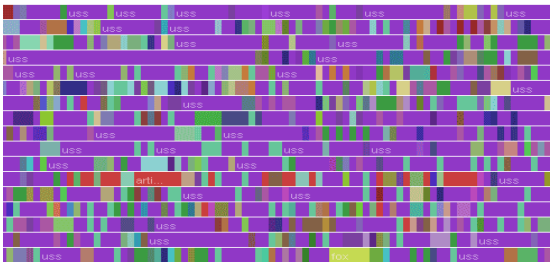
- *Verilənlərin ölçüsü.* Vikipediyanın VB-də verilənlərin həcmi baytlarla göstərilir. Hər bir viki-səhifədə, kateqoriyada, faylda, dilə və mövzuya görə ayrılmış layihələrdə toplanan verilənlərin həcmi onların BD kimi analiz olunub-olunmamasını müəyyən edir. Bu baxımdan, böyük həcmdə verilənlərin saxlanması üçün xüsusi şərtlər tələb olunur.
- *Verilənlərin müxtəlifliyi.* Viki-kontent müxtəlif tipdə ola bilər. Viki-səhifələrə mətn, səs, şəkil və video-fayllar yerləşdirmək mümkün olduğuna və viki-texnologiyaları əksər formatları dəstəklədiyinə görə verilənlər müxtəlif tipdə və strukturda toplanırlar. Bu verilənləri bir yerə toplayıb emal etmək mümkün deyil və analiz üçün əvvəlcə onları yararlı vəziyyətə gətirmək tələb olunur.
- *Verilənlərin sürəti.* Viki-serverlər brauzerdən daxil edilən bütün verilənləri (sorgular, mediafayllar, proqram kodları, mətn və s.) emal edərək yadda saxlayırlar. Nəzərə almaq lazımdır ki, Vikipediyaya daxil edilən hər bir simvol generasiya olunaraq bazada saxlanılır və veb-səhifələrdən silinsə belə, VB-da qalır və istənilən zaman istifadəçi tərəfindən bərpa edilir. Saniyədə milyonlarla informasiya saxlayan Vikipediyanın VB üçün heç bir informasiyanın silinməməsi zaman keçdikcə verilənlərin həddən artıq çoxalması deməkdir. Verilənlərin emalında sürət nə qədər çox olarsa, interaktivlik şərtləri daha yaxşı ödənilər.
- *Verilənlərin dəyişgənliyi.* Viki-səhifələr açıqdır və hər an birbaşa brauzerdən dəyişdirilir. Baxılan səhifə yenilənmədən sonra artıq tam başqa strukturda və dizaynda təqdim oluna bilər. Səhifənin bütün köhnə versiyaları (yaranma tarixi, şərhlər və müəllif göstərilməklə) bazada saxlanılır. Bu faktorlar analiz zamanı verilənlərin idarə olunmasında müəyyən problemlər yaradır.
- *Mürəkkəblik.* Verilənlər müxtəlif mənbələrdən daxil olurlar. Viki-səhifələr arasında mürəkkəb semantika mövcuddur. Səhifələr arasında linklər müxtəlif tipdə olurlar: layihədaxili linklər, dil faktoruna görə

(intervikilər) linklər, kateqoriyalara görə linklər, digər viki-layihələrə linklər və nəhayət, xarici layihələrə (İnternetin müxtəlif veb-saytlarına) linklər. Bu linklər hər an dəyişdirilə bilər. Belə bir mürəkkəblik vikimetrik tədqiqatları da mürəkkəbləşdirir.

Verilənlərin ənənəvi emalı zamanı onlar yoxlanılır, müəyyən formata gətirilir və yalnız ondan sonra sistemə yüklənir. Bu cür ardıcıllıq viki-mühitdəki BD məsələlərində özünü doğrultmur. Böyük verilənlərin emalı prosesi verilənlər axınının vizuallaşdırılması, verilənlərin intellektual analizi, BD əsasında situasiyanın təsviri, analiz üçün müasir aparat və proqram təminatı və s. məsələlərin həllini nəzərdə tutur.

III. PROBLEMİN HƏLLİ ÜÇÜN BƏZİ YANAŞMALAR

IBM şirkəti Vikipediyada istifadəçilərin fəaliyyətini və hadisələr arasında qanunauyğunluqları müəyyən etmək üçün xüsusi alqoritm işləmişdir. 2007-ci ildən başlayaraq tətbiq edilən alqoritm verilənlərin vizuallaşdırılmasına əsaslanır. Xromoqram şəklində təsvir və xromoqramdakı rənglər mətndəki sözlərə verilən rənglərdən asılıdır. Xromoqramdan istifadə etməklə istifadəçinin hansı mövzuya maraq göstərməsini, vandalizm hallarını, konfliktli situasiyaları və digər maraqlı məsələləri analiz etmək mümkündür [7]. Məsələn, aşağıda göstərilən xromoqram hərbi-dəniz qüvvələrinin tarixi ilə bağlı məqalələrdə edilən düzəlişlər əsasında təşkil olunmuşdur. Xromoqramda üstünlük təşkil edən bənövşəyi rəng hərbi-dəniz gəmilərinin adlarını göstərir. Yəni, xromoqram göstərir ki, istifadəçi müxtəlif mövzulara müraciət etsə də, hərbi gəmilərə olan maraq üstünlük təşkil edir.



Şəkil 1. Viki-mühitdə istifadəçinin maraq dairəsini göstərən xromoqram

Vikipediyada böyük verilənlərin analizini nəzərdə tutan digər sistem SGI UV 2 sistemidir [8]. SGI şirkəti və İllinoys Universitetinin əməkdaşı K. Liptau (Kalev H. Leetaru) tərəfindən yaradılan bu sistem Vikipediyada tammətinli kontentlərin zaman və məkana görə xronoloji kartoqrafiyasını yaradır və axtarışını həyata keçirir. Sistem ingilis dilindəki Vikipediyada toplanmış verilənləri analiz etməklə dünya tarixinin inkişaf mərhələlərini müəyyən etmiş və müasir tarixin vizual təsvirini yaratmışdır. Sistem həmçinin, Vikipediyadakı verilənlərdən istifadə etməklə son 2 əsrdə dünyanın necə vizuallaşmasını göstərmişdir. Sistem viki-səhifələrin məzmununu müəyyən etmiş, səhifələr arasındakı semantik əlaqələri dəqiqləşdirmiş və bu əlaqələr əsasında böyük şəbəkə yaratmışdır (şəkil 2).

Şəkil 2. Dünyanın SGI UV 2 sistemi vasitəsi ilə qurulmuş 1900-cü ildə vizual tarixi xəritəsi

SGI şirkətinin marketinq üzrə direktoru Frans Aman (Franz Aman) SGI UV 2 sistemini Google Earth ilə müqayisə edərək bildirmişdir: “Google Earth-də xəritənin ölçülərini azaltmaqla ümumi vəziyyətə baxmaq bizim xoşumuza gəlir və bu konsepsiyanı Vikipediyadakı böyük verilənlərə tətbiq etməklə, biz ümumi təsvir əldə etmişik” [8]. Sistemdə verilənlərin emalı əsas yaddaşda toplanan verilənlərin intellektual analizi (in-memory data-mining) vasitəsi ilə həyata keçirilir. İntellektual analiz alqoritmləri, ilk növbədə, mətnin xüsusiyətini, məzmununu, səhifələr arasındakı əlaqələri üzə çıxarır. Nəticələr metaverilənlərdə saxlanılır və sonrakı analizlərdə istifadə olunur.

Bəzi yanaşmalarda BD problemləri ənənəvi 3V (volume, variety, velocity) [9] ilə deyil 3C parametrləri ilə göstərilir: cardinality (həddən artıq böyüklük), continuity (fasiləsizlik) və complexity (mürəkkəblik) [10]. Müəlliflərin fikrincə məsələnin həllində riyazi və statistik analiz üsullarını 3C parametrlərinə görə qurmaq daha sadədir. Nəzərə almaq lazımdır ki, 3V şərtindəki verilənlərin sürəti Vikipediyada bir çox hallarda özünü doğrultmur. Belə ki, viki-mühitdə kontentin daxil edilməsində və dəyişdirilməsində fərqli fasilələrin yaranması ehtimalı vardır.

IV. PROBLEMİN HƏLLİ YOLLARI

BD probleminin həlli ilə bağlı bəzi mövcud yanaşmaların analizi belə deməyə əsas verir ki, problemin həllində elə optimal həll yolunu seçmək lazımdır ki, alqoritm əvvəlki təcrübələrdən topladığı verilənlərdən istifadə edərək gələcəkdə viki-mühitdə hadisələrin inkişafını qabaqcadan avtomatik göstərsin. Belə ki, ənənəvi verilənlər xəzinəsi ilə işləyərkən müəyyən məsələlərin həlli zamanı, məsələn, verilənlər arasında uyğunsuzluğun təyin edilməsi və s. məsələlərdə əvvəlcədən verilənlər xəzinəsində toplanmış və aqreqatlaşdırılmış verilənlərdən istifadə olunur. BD ilə işləyərkən məsələnin bu cür həlli problemlər yaradır. İnformasiyanın emal prosesinə hazırlanması üçün verilənlər əvvəlcədən hazırlanmış serverlərə köçürülməlidirlər. Viki-mühitdəki BD bu yanaşmanı da mümkünsüz edir. Digər problem ondadır ki, viki-mühitdə verilənlər paylanmış şəkildədir və kontent yaratmaqla məşğul olan istifadəçilərin istəyinə uyğun şəkildə strukturlaşdırılır və formata gətirilir. Məsələn, verilənlər regional serverlərdə toplanır və hər bir serverdəki verilənlərin strukturlaşdırılması müxtəlif olur.

Verilənlərin analizi baxımından bu verilənləri əraziyə görə deyil, zamana görə serverlərə paylamaq daha düzgün olardı və hər bir server müəyyən zaman intervalına cavab verərdi. Lakin viki-serverlərin imkanları buna yol vermir. Bunları nəzərə alsaq viki-mühitdəki BD-nin analizi üçün verilənləri aşağıdakı növlərə bölmək olar:

- Yaranma zamanına görə (köhnə verilənlər/yeni dəyişdirilmiş verilənlər);
- Əldə olunma üsuluna görə (kompüterlərdən/mobil telefonlardan daxil edilən verilənlər);
- Tipinə görə (mətn/video/səs/şəkil tipli verilənlər).

BD ilə işləyərkən verilənlər axını girişə verilir və bu zaman çox sürətli analiz tələb olunur. Məsələn, viki-mühitdə gizli sosial şəbəkələrin aşkarlanması, konfliktli situasiyaların qarşısının alınması, viki-səhifələrin və istifadəçilərin monitorinqinin aparılması, sürətlə yayılan informasiyanın və onun mənbəyinin müəyyən edilməsi və s. kimi məsələlər tez bir zamanda həll olunmalıdır. Lakin informasiya həddən artıq çox olduğundan biz burada sürətli emal imkanına malik intellektual analiz alqoritmlərinə ehtiyac görürük. Alqoritmlər həm də viki-səhifələrdəki mətnləri, istifadəçilərin müzakirələrini, daxil etdikləri şəkil, video və audio-faylları emal etməlidirlər. Odur ki, viki-mühitdə toplanmış BD ilə işləmək üçün aşağıdakı şərtlər ödənməlidir:

1. Viki-mühitdəki verilənlərin analizində istifadə olunacaq metod müəyyən edilməlidir;
2. Qısa zamanda böyük verilənlərin analizini həyata keçirə biləcək çox güclü intellektual sistem olmalıdır.

BD-dan əldə olunmuş massivlər heç də həmişə təsadüfi seçmədən əldə olunmur. Bəzən statistik verilənlər daha böyük əhəmiyyət kəsb edir. Burada əsas problem ondan ibarətdir ki, böyük, mürəkkəb, geterogen verilənlərin statistik analizi yalan və uyğun olmayan dəyişənlərin və alqoritmlərin seçilməsinə səbəb ola bilər. Odur ki, böyük verilənlərdən istifadə zamanı əhatəlilikdə, seçimdə, ölçmədə və əldə olunan cavablarda yanlışlıq ola bilməsi riskləri nəzərə alınmalıdır [11].

Böyük verilənlərin analizində yalnız proqram deyil, həm də xüsusi aparat təminatına tələbin olmasını nəzərə alaraq bu sahədə yanaşmaların xüsusi kompüter modelləşmələri ilə birgə tədqiqi məsələsi önə çəkilməlidir. Məsələn, buna misal olaraq agent modelləşdirməsini göstərmək olar. Agent modelləşdirmə desentralizə olunmuş verilənlərin davranışını tədqiq edir və bu verilənlərin davranışı bütün sistemin davranışını təyin edir [12]. Eyni zamanda faktor və klaster analiz kimi çoxölçülü analiz metodlarından istifadə etməklə verilənlərin gizli stukturunu tədqiq etmək mümkündür [13]. Böyük verilənlərin klasterləşdirilməsində Hadoop Distributed File Systems (HDFS) [14] modelindən, Hive database və ZooKeeper metodlarından [15], Cloud texnologiyalarından [16] birgə istifadə daha səmərəli nəticə əldə etməyə imkan verə bilər.

HDFS verilənlərin xüsusi klasterlərdə saxlanması və təşkilini təmin edir. ZooKeeper – yüksək əlyetərliyə malik əlaqələndirmə servisidir. Paylanmış əlavələrin qurulmasında istifadə olunur. Hive – paylanmış verilənlər xəzinəsidir. Hive xüsusi sorgu dili olan HiveQL dilində işləyir və HDFS-də toplanmış verilənləri idarə edir. HiveQL sorğuları avtomatik olaraq Map-Reduce məsələlərinə tətbiq edilirlər.

BD problemlərinin həlli ilə bağlı yanaşmaların qısa təhlili göstərir ki, verilənlərin paylanmış şəkildə emalı böyük verilnlərin tez bir zamanda analizini aparmağa kömək edə bilər. Böyük verilənlərin səmərəli emalı paylanmış fayl sistemləri texnologiyalarına əsaslanır. Paylanmış sistemlərdə verilənlər bir fayl sistemində deyil, bir neçə serverlərdə yerləşən verilənlər xəzinəsində saxlanılır və indeksasiya edilir. Bu baxımdan, böyük verilənlərin emalında paralel və paylanmış verilənlər bazası idarəetmə sistemlərindən (VBİS) istifadə və verilənləri yerləşdikləri qovşaqlarda emal etmək daha düzgündür. Bu halda paralel VBİS-lərin iş keyfiyyəti verilənlərin düzgün paylanmasından asılıdır. Analitik VB üçün verilənlərin paylanması sorğuların yerinə yetirilməsi zamanı qovşaqlar arasında informasiya mübadilələrinin sayının minimum olmasına gətirib çıxaracaqdır.

Nəzərə almaq lazımdır ki, ənənəvi verilənlər xəzinəsində bütün verilənlər onların konvertasiyası, təmizlənməsi, yoxlanması və yüklənməsi işlərinə cavabdeh olan bir məntiqi blokdan verilirsə, BD-nin emalı zamanı bir məntiqi blokdan istifadə səmərəli deyildir. Ənənəvi verilənlər xəzinəsində istifadə olunan alətlər: çoxölçülü analiz (OLAP), reqressiya, təsnifatlandırma, klasterləşmə və qanunauyğunluqların axtarışıdır. Bu gün SAP HANA, Greenplum Chorus, Aster Data Cluster kimi sistemlər bu metodları BD-nin analizi üçün tətbiq etməyə imkan yaradır. BD ilə işləmək üçün ən populyar texnologiyalara misal olaraq 2004-cü ildə Google şirkəti tərəfindən təklif olunan Map-Reduce modelini və Yahoo şirkəti tərəfindən təklif olunan açıq kodlu Hadoop proqramını göstərmək olar. Bu sistemlərdə verilənləri cədvəllərə yığmaq və ciddi iyerarxiyaya tabe olmaq lazım gəlmir. Analizdən öncə verilənləri növlərə bölmək kifayətdir. Map-Reduce modeli verilənlərin düzgün paylanması işində səmərəliliyi təmin edir. Map-Reduce – paylanmış verilənlərin hesablanması modelidir. Böyük verilənlər üzərində paralel hesablamalar üçün istifadə olunur. Giriş verilənləri ilkin emal üçün nəzərdə tutulmuş xəritə (Map) işçi qovşaqlara (individual nodes) paylanmış verilənlərin harada yerləşməsini göstərir, emalda istifadə olunan alqoritmlər və nəticə haqqında informasiya saxlayır. Əməliyyatların sonunda nəticələr toplanır (Reduced). Məsələn, alqoritm yekun cəmi tapmaq üçün paralel olaraq paylanmış fayl sisteminin hər bir qovşaqda aralıq cəmləri hesablayacaq və sonrakı mərhələdə bu aralıq qiymətləri toplayacaqdır.

Verilənlərin effektiv paylanması üçün verilmiş qrafın müəyyən xüsusiyyətli altqraflara ayrılması məsələlərinə də baxmaq lazımdır. Burada əsas şərt altqrafları birləşdirən qovşaqların sayının minimum olmasıdır. Bu, "NP-completeness" məsələsidir. Yəni, məsələni həll etmək üçün onu NP (non-deterministic polynomial) sinifindəki məsələlərə bölmək daha düzgündür. Əgər bölünmüş məsələlərdən hər hansı biri üçün həll alqoritmi tapılarsa, o zaman NP sinifindən olan digər məsələlərin həllində də bu alqoritmdən istifadə

etmək mümkündür. Məsələnin həllində intellektual analiz metodlarından və hibrid alqoritmlərdən istifadə daha səmərəli nəticələr əldə etməyə imkan yaradar. BD-nin emalı üçün yalnız verilənlərin paylanmış emalını həyata keçirən konkret texnologiyalardan istifadə ilə yekunlaşmaq olmaz. Emal zamanı BD üçün xarakterik olan vacib parametrlər də nəzərə alınmalıdır. Məsələn, şəbəkədə qarşılıqlı əlaqənin intensivliyi, verilənlərin həcmi və s. Viki-mühitdə də, digər sahələrdə olduğu kimi BD problemləri yeni suallar yaradır ki, onlara cavab vermək lazım gəlir. Məsələn, BD ilə bağlı yanaşmaların və modellərin verifikasiyası və validasiysı vacib məsələlərdəndir. Bu ondan irəli gəlir ki, verilənlər həddən artıq böyük olduqda xətaların ölçülməsi və ölçmə zamanı yaranan xətaların sayı qarşılıqlı əlaqəlidir.

NƏTİCƏ

Tədqiqat nəticəsində məlum olmuşdur ki, viki-mühitdə böyük verilənlərin toplanması bu verilənlərin emalında və saxlanmasında problemlər yaradır və məsələlərin həllində yeni yanaşmaların tətbiqini tələb edir. Viki-mühitdəki böyük verilənlərin vizuallaşdırılmasına aid mövcud sistemlər göstərdi ki, Vikipediyada toplanmış verilənlər özlərində BD-nın xüsusiyyətlərini saxlayırlar: ölçüləri böyükdür – onların səmərəli strukturlaşdırılması və saxlanması üçün xüsusi şərtlər tələb olunur, müxtəlifdirlər – onları ənənəvi üsullarla bir VB-yə toplayıb emal etmək mümkün deyil. Həmçinin viki-səhifələr arasında mürəkkəb semantika mövcuddur ki, bu da onların emalını çətinləşdirir. Viki-mühitdəki verilənlərin müxtəlifliyi və dəyişkənliyi onların analizi və təsvirində kompüter qrafikasının yeni texnologiyalarından istifadəni qaçılmaz edir.

BD ilə bağlı məsələlərin həllində araşdırılan yanaşmalar belə deməyə əsas verir ki, verilənlərin paralel və paylanmış şəkildə emalı daha səmərəli nəticə əldə etməyə imkan yaradar.

ƏDƏBİYYAT

[1]. R.M. Əliquliyev, Hacırəhimova M.Ş. “Big Data” fenomeni: problemlər və imkanlar // İnformasiya texnologiyaları problemləri. 2014, №2, s. 3–16.

[2]. I.Y. Alakbarova. Some Approaches to the Development of Information Influence and Hidden Communications Detection Systems in Wiki-Environment / Proceedings of the 4th International Conference “Problems of Cybernetics and Informatics” (PCİ), vol. I, Sept. 12-14, 2012, pp. 119-120

[3]. Р.М. Алгулиев, Р.М. Алыгулиев, И.Я. Алекперова. Викиметрические исследования: современное состояние и песпективы // “Телекоммуникации”, М. 2014, №5, стр. 15-31

[4]. https://meta.wikimedia.org/wiki/List_of_Wikipedias

[5]. https://commons.wikimedia.org/wiki/Main_Page

[6]. https://www.openhub.net/p/mediawiki

[7]. M. Wattenberg , F. B.Viégas, Hollenbach K. Visualizing Activity on Wikipedia with Chromograms // Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction, 10 Sept. 2007, Berlin, vol. 4663, pp. 272–287

[8]. http://www.sgi.com/go/wikipedia/

[9]. R. Alguliyev, Y. Imamverdiyev. Big Data: Big Promises for Information Security / Proceedings of the 8th International Conference on Application of Information and Communication Technologies (AICT), 15-17 Oct. 2014, IEEE, Astana, pp. 1–4

[10]. M. Hilbert. Big Data for Development: From Information- to Knowledge Societies, 2013, Available at SSRN: http://ssrn.com/abstract=2205145

[11]. M. Couper. Is the sky falling? New technology, changing media, and the future of surveys // Survey Research Methods, 2013, vol. 7, no. 3, pp. 145–156.

[12]. S. Suthaharan. Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning // ACM Sigmetrics Performance Evaluation Review archive, 2014, vol. 41, no 4, pp. 70–73

[13]. R.M. Alguliev, R.M. Aliguliyev and I.Ya. Alekperova. Cluster approach to the efficient use of multimedia resources in information warfare in Wikimedia // Automatic Control and Computer Sciences, 2014, vol.48, no 2, pp. 97–108

[14]. A. Thusoo, J.S. Sarma, N. Jain, Z. Shao, P. Chakka, N Zhang, A. Antony, H. Liu and R. Murthy. Hive – A Petabyte Scale Data Warehouse Using Hadoop // Proceedings of the 26th International Conference on Data Engineering (ICDE), 2010 IEEE, 1-6 March, 2010, Long Beach, pp. 996 – 1005

[15]. P. Hunt, M. Konar, F.P. Junqueira and B. Reed. ZooKeeper: wait-free coordination for internet-scale systems // Proceedings of the 2010 USENIX conference on USENIX annual technical conference. Berkeley, CA, USA: USENIX Association, 2010, pp. 11–12

[16]. S. Carlin and K. Curran. Cloud Computing Technologies // International Journal of Cloud Computing and Services Science (IJ-CLOSER), 2012, vol. 1, no. 2, pp. 59-65

Böyük Verilənlərdən Verilənlər Haqqında Elmə: Fənlərarası Perspektiv

Məkrufə Hacırəhimova[1], Hicran Gözəlova[2]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*makrufa@science.az*, [2]*gozalova@yandex.com*

***Xülasə* — “Big data” konsepsiyasının populyarlaşması ilə *«data scientist»* daha cəlbedici və perspektivli ixtisasa çevrilmişdir. Məqalədə verilənlər haqqında elmin terminoloji, tarixi aspektləri, mahiyyəti şərh olunmuş, bir fənn kimi bu sahədəki mövcud vəziyyət analiz olunmuşdur. Məqalədə həmçinin verilənlər haqqında elmin fənlərarası perspektivləri araşdırılmışdır.**

Açar sözlər — böyük verilənlər; verilənlər haqqında elm; verilənlər alimi.

I. GİRİŞ

Böyük verilənlər (BV) problemi yeni problem deyil. Texniki və texnoloji inkişaf ilə əlaqədar olaraq verilənlərin sürətlə artması və bu verilənlərin emalı problemləri hələ keçən əsrin 40-cı illərinə təsadüf edir [1]. Şübhəsiz o vaxt ilə müqayisədə, son on ildə verilənlərin həcmi kəskin surətdə artmışdır. Bunu IDC, Gartner və s. kimi analitik şirkətlərin hesabatları da təsdiq edir [2].

İnformasiya texnologiyalarının səhiyyə, astronomiya, bioinformatika, nəqliyyat, dövlət idarəçiliyi və s. kimi sahələrdə geniş tətbiqi verilənlər axınının artmasına gətirib çıxarmış, dünyada informasiya bolluğu yaranmışdır. Bunun məntiqi nəticəsi kimi verilənlərin mənimsənilməsi və istifadəsində yeni eranı əks etdirən böyük verilənlər (*ing. big data*) termini meydana çıxmışdır [3].

Verilənlərin həcminin böyük sürətlə artması həm informasiya istehlakçılarını, həm də informasiya emalı texnologiyaları istehsalçılarını bu verilənləri real-vaxt ərzində emal edə bilməməsi problemi ilə üz-üzə qoymuşdur. Problem müxtəlif mənbələrdən və müxtəlif formatda (*strukturlaşdırılmış və strukturlaşdırılmamış*) fasiləsiz olaraq generasiya olunan verilənlərin ənənəvi üsul və alətlərin köməyi ilə emal edilməməsindədir. Bəzən “new oil”, bəzən “data mining” və s. kimi dəyərləndirilən BV-nin həqiqi qiyməti qarşılıqlı əlaqəli verilənlər dəstinin analizi nəticəsində biznes tendensiyaları proqnozlaşdırmaq, insan səhhətindəki təhlükələri analiz etmək, terror təhlükələrini müəyyən etmək, axtarışı optimallaşdırılmaq, bioloji və ekoloji tədqiqatlarda və s. dərin korrelyasiyaları və prinsipləri tapmağa imkan verir. Yəni, yeni nəzəriyyələr, yeni üsullar, analitik alətlər alimlərə və biznes nümayəndələrinə BV-də aşkar şəkildə olmayan biliklərin aşkarlanmasında kömək edə bilər. Bu, verilənlər haqqında elmin əsasını təşkil edir [3,2]. Verilənlər haqqında elm (*ing. data science*) bəzən də datalogiya adlandırılan (*ing. datalogy*) bu akademik sahə informatikanın bir bölməsi olmaqla rəqəmsal verilənlərin emalı, analizi və təqdim olunması problemlərini öyrənir [4,5,6]. Verilənlər haqqında elm verilənlərdən əlavə biliklərin və qiymətli informasiyanın əldə olunmasıdır. O, riyaziyyat, statistika, obrazların tanınması, biliklər bazası, maşın təlimi və s. kimi fənlərarası (*multidisciplinary*) yanaşmaları təklif edir. Bu, verilənlərin həyat tsiklinin idarə edilməsi, verilənlərin məxfiliyinin təminatı, verilənlərin hesablanmasının elastikliyi, böyük verilənlərin məkan və zaman xarakterləri və sosial aspektlərinə əsaslanan yeni elmi-tədqiqat sahələrinin yaranmasına səbəb olur. “Big data” konsepsiyasının populyarlaşması ilə o həm elm, həm də biznes strukturları kimi maraqlı tərəflərin diqqət mərkəzinə çevrilmişdir. Belə ki, BV-dən biliklərin aşkarlanmasında yeni keyfiyyətdə kadrlara ehtiyac böyükdür. Bu da “big data” erasında təhsil müəssisələrinin qarşısında duran əsas məsələlərdəndir [3-7]. Verilənlər haqqında elmlə məşğul olan mütəxəssislər verilənlər alimi (*ing. data scientist*) adlanırlar [8]. Onlar bir qayda olaraq, riyaziyyat, statistika və informatikanın əsas elementləri ilə işləmə qabiliyyətinə, həm də müəyyən bir elm sahəsində böyük təcrübəyə malik olurlar.

“Big data” texnologiyaları və verilənlər haqqında elm iqtisadiyyat, biznes, maliyyə, səhiyyə, fizika, astronomiya, geologiya, biologiya, sosial və humanitar elmlər və s. daxil olmaqla müxtəlif sahələrdə tədqiqatların aparılmasına güclü təsir edir. Multidissiplinar baxımdan verilənlər haqqında elm yeni yaranmış sahə kimi bir çox fənləri əhatə edir ki, bu da ona dıgər fənlərlə müqayisədə üstünlük qazandırmağa kömək edir. İşdə də əsas məqsəd akademik fənn kimi “data science” sahəsində mövcud vəziyyəti və perspektivləri araşdırmaqdır.

II. VERİLƏNLƏR HAQQINDA ELM

“O´Reilly Radar” jurnalında nəşr olunan “Verilənlər haqqında elm nədir?” (What is Data Science?) məqaləsinin müəllifi Mayk Lukidis yazırdı: “Gələcək, verilənləri məhsula çevirə biləcək insan və kompaniyalara məxsus olacaq”. Bu deyim, məşhur bir kəlamı yada salır: “Kim informasiyaya malikdirsə, o da dünyaya sahibdir” [9]. Bu gün bu aforizmə bir az düzəliş etmək olar: “Dünyanı verilənlərə və onların analizi texnologiyalarına malik olanlar idarə edir”. İnformasiyanın alınması üsullarına sahib olmaq verilənlər haqqında elm adlanır. “Data science” termin kimi ilk dəfə 1960-cı illərdə kompüter elmlərinin pionerlərindən sayılan Peter Naur tərəfindən “Datalogy, the science of data and of data processes

and its place in education” adlı məruzəsində səsləndirilmiş [10], fənn kimi isə keçən əsrin 60-cı illərindən formalaşmağa başlamışdır. Bu 1966-cı ildə Elm və Texnika üçün Məlumat Komitəsinin (CODATA – International Council for science: Committee on Data for Science and Technology) təsis edilməsinə təsadüf edir [3,4]. 1974-cü ildə Peter Naur “Concise Survey of Computer Methods in Sweden and the United States” kitabında verilənlərin müasir emal metodlarının icmalını vermiş və verilənlər haqqında elmi rəqəmsal verilənlərin həyat dövrünü (yarandığı andan başqa bilik sahələrinə təqdim olunmaq üçün edilən dəyişikliklərə qədər) öyrənən bir fənn kimi təyin etmişdir. Ancaq, termin 1990-cı illərdə geniş istifadə olunmağa başlamış və 2000-ci illərin əvvəllərindən hamı tərəfindən qəbul edilmişdir. 2001-ci ildə Purdue Universitetinin professoru, statistika, verilənlərin vizuallaşdırılması, maşın təlimi sahəsində tanınmış mütəxəssis Uilyam Klivlend (*William S. Cleveland*) statistik tədqiqatların texniki aspektlərinin inkişaf planı (Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics) adlı tədqiqatında diqqəti “data science” fənninin tədrisi məsələsinə yönəltmişdir [11]. O, planda “data science”-i ayrıca akademik fənn kimi irəli sürmüşdür. Bu da “data scientist” adlanan kadrların hazırlanmasına marağı stimullaşdırmışdır. Təsadüfi deyildir ki, təxminən həmin vaxtdan da CODATA-nın “Data Science Journal”-ı (2002-ci ildən) nəşr olunmağa başlanmışdır.

Burada “data science” ilə bilavasitə əlaqədar bəzi faktları da diqqətə çatdırmaq lazımdır. 1977-ci ildə Statistik Hesablamalar üçün Beynəlxalq Assosiasiyanın (The International Association for Statistical Computing - Section of the ISI) ənənəvi statistik metodlarla müasir kompüter texnologiyaları və ekspertlərin nəticələrini əlaqələndirərək verilənləri informasiyaya və biliyə çevirmək üçün xüsusi seksiya yaradılmışdır [6].

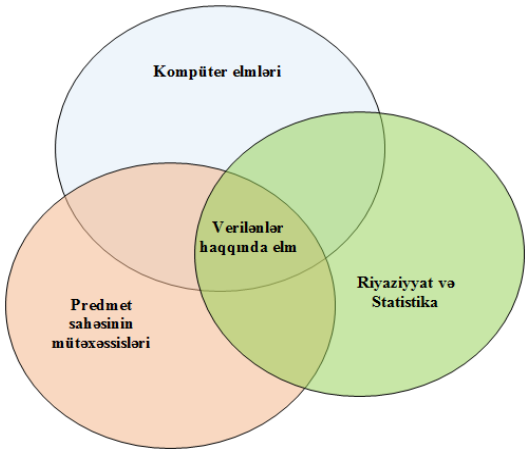
1989-cu ildə Qriqoriy Piatetski-Şapiro verilənlər bazasında biliklərin aşkarlanması (Knowledge Discovery in Databases - KDD) cəmiyyətini yaratmış və 1995-ci ildən hər il keçirilən ACM SIGKDD konfransına çevrilmişdir [6] .

1996-cı ildə verilənlər elmi Təsnifatlandırılmış Cəmiyyətlərin Beynəlxalq Federasiyasının (International Federation of Classification Societies) üzvləri tərəfindən Tokioda, ilk dəfə konfransın adına çıxarılaraq “Data science, classification, and related methods” adlı konfransı keçirilmişdir [6].

1997-ci ildə IBM şirkəti tərəfindən “Biliklərin Aşkarlanması və Verilənlərin İntellektual Analizi” (Knowledge Discovery and Data Mining) jurnalı işıq üzü gördü [12].

III. FƏNLƏRARASI PERSPEKTİV

“Data science” termini hərfi mənada “verilənlər haqqında elm” kimi tərcümə olunmamalıdır [3]. Çünki, ingilis dilində “science” sözü ancaq “elm” deyil, həm də “bacarıq”, “məharət”, “qabiliyyət” deməkdir. Deməli, bu bilik və bacarıqlara əsaslanan elmdir. Verilənlər haqqında elm ənənəvi informatikadan başlayaraq riyaziyyata qədər müxtəlif sahələr üzrə bacarıq və vərdiş tələb edir. Bu elmin əsas məqsədi verilənlərdə qanunauyğunluqların tapılması, verilənlərdən ümumiləşdirilmiş halda biliklərin aşkarlanmasıdır. Mayk Lukides verilənlər haqqında elmi verilənləri yalnız informasiyaya deyil, həmçinin məhsula çevirdiyini iddia edir [9]. Bu sahədə vacib olan bacarıqlar adətən, mütəxəssisə lazım olan ümumi fənlərlə informasiya texnologiyalarındakı təcrübələrin və riyazi statistika biliklərinin kəsişməsi – Venn diaqramı [4] vasitəsilə izah olunur (şək.1) .



Şəkil 1. Venn diaqramı

Şəkildən də göründüyü kimi “data science” bir çox elm sahələrinin kəsişməsidir. “Big data” texnologiyalarının tətbiq sahələri artdıqca bu kəsişmələrin sayı da artır. Maşın təlimi, data mining və text mining, süni intellekt, təbii dilin emalı, prediktiv modelləşmə, vizuallaşdırma, təhlükəsizlik və s. kimi elm sahələri daxil olmaqla daha geniş miqyas almaqdadır.

BV-nin potensialından tam istifadə etmək üçün yalnız verilənlər, alətlər və infrastruktura deyil, “nəhəng verilənlər dağını aşmağa” kömək edən çoxlu sayda bacarıq da tələb olunur [13]. Bu sahə ilə məşğul olmaq üçün “*data scientist*” adlanan yüksək ixtisaslı mütəxəssislərə ehtiyac vardır. Bu mütəxəssislər böyük verilənlərdə qanunauyğunluqların aşkarlanmasında təşkilatlarda aparıcı rola malik olacaqlar.

McKinsey şirkəti 2018-ci ilədək ABŞ-da 440-490 min “*data scientist”* ixtisaslı mütəxəssisə tələbatın olacağını proqnozlaşdırır. Bu tip proqnozlar bu sahədə təhsil proqramlarının yaranması məsələsini bir qədər də sürətləndirmişdir [14] .

ABŞ-ın bir çox ştatlarında və digər inkişaf etmiş dövlətlərin universitetlərində “data science” bir elm sahəsi kimi bakalavr, magistr və doktorluq pilləsində tədris edilir [15]. 2013-cü ildən başlayaraq Dandi Universitetində (Şotlandiya), Oklend Universitetində (Yeni Zelandiya), London İmperial Kollecində, Cənubi Kaliforniya Universitetində, Vaşinqton, Berkli, Nyu-York Universitetlərində və s. verilənlər haqqında proqramlar tədris olunmaqdadır [15]. Belə proqramlar böyük həcmli verilənlərin potensialından maksimum istifadə etmək üçün

hesablama modelləri, modelləşdirmənin və proqnozlaşdırmanın riyazi metodları, arxitektura, müasir proqramlaşdırma metodları, verilənlərin əldə edilməsi, saxlanması və analizi fundamental hazırlığı təmin etməyi nəzərdə tutur. Hazırda bu universitetlərin sayı getdikcə artmaqdadır. Məsələn, Rusiyada Ali İqtisadiyyat Məktəbində, Türkiyədə Sabançı Universitetində və s. magistr pilləsi üzrə proqram tədris olunur. “Data scientist” ixtisası son illərdə daha cəlbedici və perspektivli ixtisaslardan biri kimi təqdim edılir.

[15]-də təqdim olunmuş siyahıda adı çəkilən universitetlər xüsusi olaraq böyük verilənlərin analizinə yönəlmiş və bu sahədə magistr pilləsinə yiyələnməyə nəzərdə tutulmuş fənlər tədris edən ali məktəblərdir. Bu təhsil proqramını öyrənməklə məzunlar böyük verilənləri sərbəst şəkildə idarə edə biləcək, ixtisasyönümlü müəssisələr tərəfindən arzuolunan mütəxəssisə çevriləcəklər. Yaxşı mütəxəssis yalnız biznes problemlərinin həlli ilə deyil, müəssisəyə ən çox gəlir gətirən sahəni də düzgün seçməyi bacaracaqdır. Tədris olunan əsas fənlər proqramlaşdırma, alqoritm, verilənlər bazası, riyaziyyat, statistika, verilənlərin analizi, verilənlərin kompüter emalı, paylanmış sistemlər, böyük verilənlər alətləri və verilənlərin vizuallaşdırılmasıdır.

Qeyd edildiyi kimi, “big data” paradiqmasının mövcud olması ilə verilənlər haqqında elmə maraq daha da artmışdır. Verilənlər haqqında elm üzrə IEEE International Conference on Data Science and Advanced Analytics, 2011-ci ildən O´Reilly şirkətinin Strata konfransları, EMC-in sammitləri keçirilir.

Bütün bu sadalananlar “data science”-in böyük perspektivlərindən xəbər verir. Böyük verilənləri idarə etmək üçün “verilənlər alimləri”nin əsas bacarığı kod yazmaqdan başlayaraq, analiz etmək və hamının anladığı tərzdə verilənləri vizuallaşdırmaqdır. Müəssisənin müvəffəqiyyəti onlardan birbaşa asılıdır, onların yaratdıqları alətlər isə müəssisə hüdudlarını aşır. Verilənlər alimləri, məsələn, marketinq sahəsində bazarın seqmentlərə bölünməsi, modelləşdirmə, sosial medianın analizi məsələlərini, yaxud maliyyə və sığorta şirkətlərində müxtəlif tətbiqi məsələləri həll edirlər. Şirkətlər, “data scientist” adı altında ayrı-ayrı komandalar və ya şöbələr yaradır ki, onların da tərkibində dörd tip mütəxəssis olur. Bunlar, hər biri biznesdə müxtəlif səviyyəlli bacarıqlara – kompüter, statistik və proqramlaşdırma, riyazi biliklərə malik olan “data business person”, “data creative”, “data developer”, “data researcher” qrupundan təşkil olunur. IBM şirkətinin vitse prezidenti Anjul Bhambri “data scientist” haqqında demişdir: “Böyük verilənlər mütəxəssisləri – verilənlərə baxan və tendensiyaları təyin edənlərdir. Onlar, müəssisəyə dəyişiklik gətirmək üçün öyrənməyi üstün tutan dirçəliş dövrünün şəxsiyyətləridir” [14].

NƏTİCƏ

Verilənləri biliyin əsas inqridenti hesab etmək olar. Bu baxımdan verilənləri analiz etmək, məlum olmayan bilikləri aşkarlamaq və təqdim etmək vacibdir. Ancaq verilənlərdən biliklərin əldə olunmasında yüksək bilik və bacarığa malik mütəxəssislərə ehtiyac yaranır. Verilənlər artdıqca onları emal etmək üçün informatika və tətbiqi riyaziyyat, statistika və s. kimi elmlərin kəsişməsindən alınan “data science” akademik fənn meydana gəlmişdir. O, bir sıra Amerika və Avropa dövlətlərinin aparıcı universitetlərində müxtəlif səviyyələrdə (*bakalavr*, *magistr*, *doktorantura*) tədris olunur. Elmdə fənlərarası istiqamətlərin çoxalması, həmçinin “big data” texnologiyalarının imkanlarından istifadə edən elm sahələrinin sayı artdıqca “data science” daha çox sahənin mütəxəssislərini özündə birləşdirəcəkdir.

ƏDƏBİYYAT

[1] Черняк Л. Серьезно о технологиях для Больших Данных, http://www.osp.ru/os/2014/01/13039646/

[2] The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe.

[3] Əliquliyev R.M., Hacırəhimova M. Ş. "Big Data" phenomenoni: problemlər və imkanlar, İnformasiya texnologiyaları problemləri, 2014, №2, səh. 3-16

[4] https://ru.wikipedia.org/wiki/Hayp,_Петер

[5] Naur P. «Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences», 1974.

[6] http://whatsthebigdata.com/2012/04/26/a-very-short-history-of-data-science/

[7] http://www.datasciencecentral.com/profiles/blogs/17-analytic-disciplines-compared

[8] https://anjulbhambhri.wordpress.com/2012/04/27/new-and-exciting-career-opportunities-for-it-professionals-data-scientist/

[9] https://beta.oreilly.com/ideas/what-is-data-science

[10] Naur P. Datalogy, the science of data and of data processes and its place in education, *Proc. IFIP Congress*, Edinburgh, Scotland,. Amsterdam: North-Holland, 1968, pp. 48-52.

[11] Cleveland W.S., Statistics Research, Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics, 2001. http://www.stat.purdue.edu/~wsc/papers/datascience.pdf

[12] http://researcher.ibm.com/researcher/view_group_pubs.php?grp=144

[13] http://www.ibmbigdatahub.com/blog/going-beyond-data-science-toward-analytics-ecosystem-part-1

[14] Big data: The next frontier for innovation, competition, and productivity. Analyst report, McKinsey Global Institute, May 2011. http://www.mckinsey.com/

http://datascience.community/colleges

Bulud Test Mühitlərinin Böyük Verilənlərin Emalında Reallaşma İmkanlarının Tədqiqi və Analizi

Fərqanə Abdullayeva
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
farqana@iit.ab.az

***Xülasə* — Məqalə bulud test mühitlərinin böyük verilənlərin emalında reallaşma imkanlarının tədqiqi məsələsinə həsr olunmuşdur. Bu məqsədlə, bulud simulyasiya proqramlarının, ticari buludların, bulud test mühitlərinin məqsədləri araşdırılmış, onların böyük verilənlərin emalında tətbiqi imkanlarının dərindən analizi aparılmışdır.**

Açar sözlər — bulud texnologiyaları; test mühiti; böyük verilənlər; CloudSim; CloudAnalyst; GreenCloud; OpenCirrus; Google Apps

I. GİRİŞ

Son zamanlar bulud texnologiyaları təşkilat və korporasiyaların infrastrukturunda yeni texnologiya kimi meydana gəlmişdir və elmin bir sıra müxtəlif tədqiqat sahələrinə ciddi təsir göstərmişdir. Bulud xidmətlərinin uğurla həyata keçəcəyinə zəmanəti təmin etmək məqsədi ilə bu sahədə yaradılmış modellər istifadəçiyə servis xidməti şəklində təklif edilməmişdən əvvəl test edilməlidir.

Bulud texnologiyalarında tədqiqatların aparılmasına maneə törədən başlıca amil eksperimentlərin yoxlanmasını təmin edən tədqiqat platformasının yaradılmasındadır.

Bulud texnologiyaları əsasən kommersiya məqsədi daşıyır və kommersiya məqsədli buludlar təbiətcə tədqiqat platforması kimi olduqca realdır. Lakin onlar tədqiqatçıların etibarlı eksperimentlərinin aparılması üçün onlara kifayət qədər idarəetmə hüququ vermir. Bundan əlavə, simulyasiya, riyazi modelləşdirmə və ya kiçik prototiplərdən istifadə etməklə aparılan tədqiqatlar nəticəsində yaradılmış bulud məhsullarının irimiqyaslı real buludlarda tətbiqi hər zaman uğurlu alınmır.

Sadalanan problemləri aradan qaldırmaq üçün bulud sistemlərini dəstəkləyən test mühitlərinin (testbeds) yaradılması istiqamətində aktiv şəkildə işlər aparılır [1]. Bu növ platformaların yaradılması həmçinin ABŞ-ın Milli Elm Fondunun (National Science Foundation, NSF) da diqqət mərkəzindədir. Belə ki, NSF-in 2012-ci ildə elan etdiyi çağırışların başlıca mövzularını məhz elə bulud texnologiyaları üçün test mühitlərinin yaradılması təşkil edirdi [2], 2014-cü ildə elan etdiyi çağırış bulud texnologiyaları üçün “Chameleon” v’ "CloudLab" adlı test mühitlərinin yaradılması ilə bağlı idi [3].

Təqdim olunan məqalədə böyük verilənlərin emalında bulud test mühitlərinin mövcud problemlərinin müəyyən edilməsi istiqamətində araşdırmalar aparılmışdır. Bulud simulyatorlarının, test mühitlərininin böyük verilənlərin emalında yararlı olmasını təmin etmək üçün təklif və tövsiyələr verilir.

II. BULUD SİMULYASİYA PROQRAMLARI

Hazırki dövrdə bulud sistemlərinin təbiətinin öyrənilməsinə həsr olunmuş çoxsaylı tədqiqatlar aparılır və bu tədqiqatların eksperimental yoxlanmasını həyata keçirmək üçün möhtəşəm simulyasiya proqramları yaradılır. Bu tip proqramlara aşağıdakılar aiddir [4]:

CloudSim. Bulud texnologiyaları infrastrukturlarını və tətbiqi proqram xidmətlərini modelləşdirməyə, simulyasiya və sınaq etməyə imkan verən ümumi simulyasiya vasitəsidir. CloudSim Melburn Universitetinin (ing. University of Melbourne) “Kompüter elmləri və proqram mühəndisliyi” şöbəsinin CLOUDS laboratoriyasında yaradılmışdır və GridSim proqram təminatının genişlənmiş formasıdır. CloudSim – verilənlər anbarını, veb-servisləri, virtual maşınlar arasında resursların paylanmasını, federallaşmış buludları və s. modelləşdirməyə imkan verir. CloudSim toolkit 3.0 proqram təminatı 2012-ci ilin yanvar ayında istifadəyə verilmişdir. CloudSim sistemindən HP və başqa təşkilatlar, o cümlədən dünyanın bir sıra universitetləri bulud resurslarının paylanması, verilənlər mərkəzinin resurslarında enerji qənaətinin idarə edilməsi, bulud texnologiyalarının optimallaşdırılması və elmi fəaliyyətlərdə istifadə edirlər. Bu sistemin çatışmazlığı onun qrafiki istifadəçi interfeysinin olmamasıdır.

CloudAnalyst. Bulud texnologiyaları mühitini və tətbiqi proqramları analiz etmək üçün CloudSim əsasında qurulmuş vizual modelləşdirmə vasitəsidir.

GreenCloud. Verilənlər mərkəzinin kompüter serverləri, şəbəkə kommutatorları, rabitə əlaqələri kimi İT qurğularının enerji sərfiyyatını modelləşdirməyə xidmət edir. GreenCloud simulyatoru NS2 (Network Simulator) şəbəkə simulyatorunun genişlənmiş formasıdır, GreenIT və ECOCLOUD layihələri çərçivəsində işlənib hazırlanmışdır.

MDCSim. 2009-cu ildə yaradılmışdır, çoxsəviyyəli verilənlər mərkəzlərini dəriniə analiz etməyə xidmət edən çevik, miqyaslana bilən, kompleks simulyasiya platformasıdır. Təyinatı real verilənlər mərkəzlərini təqlid etməkdir, bunun üçün verilənlər mərkəzlərinin serverlər, müxtəlif vendorların təqdim etdiyi kommutatorlar, kommunikasiya əlaqələri kimi xüsusi aparat təminatı xarakteristikaları ilə təchiz edilmişdir. Simulyator verilənlər mərkəzlərinin ötürücülük qabiliyyətini, cavabvermə vaxtını, enerji sərfiyyatını qiymətləndirə bilir.

iCanCloud. Böyük həcmdə saxlancları olan bulud infrastrukturlarını (minlərlə qovşağa malik) simulyasiya etmək üçün 2010-cu ildə yaradılmış platformadır. Platforma açıq kodludur və istifadəyə yararlılıq, çeviklik, məhsuldarlıq və miqyaslılıq kimi xüsusiyyətlərə malikdir. Təyinatı xüsusi aparat qurğusundakı müəyyən tətbiqi proqramın qiyməti ilə məhsuldarlığı arasında nisbəti proqnozlaşdırıb çəkiləcək xərclər haqqında istifadəçiləri məlumatlandırmaqdır.

NetworkCloudSim. CloudSim simulyatorunun genişlənmiş formasıdır, verilənlər mərkəzlərindən fərqli olaraq verilənlər axınını dəstəkləyir.

VirtualCloud. Bulud texnologiyaları mühiti yaratmağa və yeni yaradılmış həlləri test etməyə imkan verən simulyatordur.

Miqyaslılıq və eksperimentlərin həyata keçirilməsində təkrarlamaların mümkünlüyü baxımından üstünlük təşkil etdiyinə baxmayaraq simulyasiya platformaları həqiqi bulud texnologiyalarını tam əks etdirə bilmir. Bu səbəbdən tədqiqatçılar real bulud test mühitlərindən istifadə olunmasına üstünlük verirlər. *Ticari bulud test mühitləri* və *bulud test mühitləri* real bulud test mühitləridir. Bu test mühitləri haqqında ətraflı məlumat aşağıda verilir.

III. TİCARİ BULUD TEST MÜHİTLƏRİ

Bulud texnologiyaları mühitini eksperimental məqsədlərə sazlamaq üçün zəruri şərt burada aparat təminatı infrastrukturlarına giriş imkanının və onu idarə etmək üçün proqram təminatının olmasıdır. Tədqiqatçılar eksperimental tədqiqatların aparılması üçün həm ümumi məqsədlərə xidmət edən ticari bulud servislərindən və həm də elmi məqsədlər üçün xüsusiləşdirilmiş bulud test mühitlərindən istifadə edir. Ticari bulud servisləri böyük miqdarda investisiya xərcləri tələb etmədən sürətli və kiçikölçülü eksperimentlərin aparılması üçün əlverişlidir, xüsusiləşdirilmiş bulud test mühitləri isə uzun sürən eksperimentlərin aparılması üçün əlverişlidir.

Elmi tədqiqatlarda istifadə olunan *ticari bulud test mühitləri* aşağıdakılardır:

Amazon Elastic Compute Cloud (EC2). Buludda miqyaslanan hesablama resursları təqdim edən veb-servisdir, əksər elmi tədqiqatlarda istifadə olunmuşdur. İstifadəçilərin virtual maşınların sayında dəyişiklik etməsinə imkan yaradır. EC2 platformasının tətbiqini və ya onun məhsuldarlığının qiymətləndirilməsini hədəfə alan məqalələrə [5, 6, 7] ədəbiyyatlarını göstərmək olar.

Amazon S3 saxlanc xidməti (S3 storage services). İstənilən həcmdə verilənləri saxlamağa imkan yaradan, Amazon təşkilatı tərəfindən təqdim edilən bulud xidmətidir, 2006-cı ildə yaradılmışdır. Bu platformanın üzərində eksperimentlərin aparıldığı məqaləyə [8] aid etmək olar. Burada S3 və EC2 platformalarından istifadə edərək verilənlərin ötürülmə anında məhsuldarlıqda baş verən dəyişikliklər tədqiq edilmişdir.

Google App Engine. İstifadəçilərin proqram tətbiqlərini Google-nin infrastrukturunda yaratmasına və icra etməsinə şərait yaradan PaaS xidmətidir. Burada verilənlər saxlancının tutumu istifadəçinin tələbatına və trafikin həcminə görə artırıla bilir. Bu platformanın məhsuldarlığının qiymətləndirilməsinə ədəbiyyat [9]-də baxılmışdır.

Google Apps. Təbiətinə görə SaaS tipli proqram tətbiqidir. İstifadəçilərə ənənəvi ofis proqram təminatlarının (məsələn, *Word* redaktoru) funksional imkanlarına bənzər veb-tətbiqlər təqdim edir, onların bir-biri ilə asanlıqla qarşılıqlı əlaqə qurmasına şərait yaradır. Burada bütün tətbiqlər onlayn saxlandığından və onlara giriş veb-brauzer vasitəsi ilə mümkün olduğundan, istifadəçilər öz hesablarına İnternet şəbəkəsinə qoşulmuş istənilən kompüterdən daxil ola bilir və lokal olaraq əlavə proqram təminatı quraşdırmağa zərurət qalmır. [10]-da Google Apps əsasında e-təlim mühitinin yaradılması üçün yanaşma təklif edilmişdir.

Windows Azure. Platforma xidmətidir, təyinatı Windows proqram tətbiqlərini icra etməkdir. Burada proqram tətbiqləri və virtual maşınlar yaratmaq və onları idarə etmək üçün platforma təqdim edir. [11]-də Windows Azure platformasının məhsuldarlığının yüksəldilməsi üçün yanaşma təklif edilir.

IV. BULUD TEST MÜHİTLƏRİ

OpenCirrus. Bulud texnologiyaları sahəsində tədqiqatların aparılması üçün 2008-ci ildə yaradılmış ən böyük test mühitidir. Mövcud alternativlərdən fərqli olaraq paylanmış verilənlər mərkəzlərinin federallaşmasını təmin edə bilir. Layihə HP, Intel, Yahoo! və bir çox təşkilatların birgə təşəbbüsü nəticəsində yaradılmışdır. OpenCirrus (http://opencirrus.org) tədqiqatçılara fiziki və virtual maşınlardan, o cümlədən vahid giriş, monitorinq, saxlanc kimi qlobal serverlərdən ibarət bulud steki təqdim edir. OpenCirrus dörd məqsədə xidmət edir: bulud texnologiyalarında sistem səviyyəsində tədqiqatların aparılmasına köməklik, proqram tətbiqləri səviyyəsində tədqiqatların aparılmasına köməklik, eksperimental verilənlər çoxluğunun təqdim edilməsi, buludlar üçün açıq kodlu sistemlərin yaradılması. Test mühitinin tərkibinə şimali Amerika, Avropa və Asiya ölkələrindən olan on bölmə daxildir. [12]-də xüsusi buludun OpenCirrus test mühitində məhsuldarlığının qiymətləndirilməsi məsələsinə baxılmışdır.

Açıq Bulud Test mühiti (Open Cloud Testbed, OCT). Açıq Bulud Konsorsiumu (Open Cloud Consortium, OCC) tərəfindən 2009-cu ildə yaradılmışdır və bulud texnologiyalarını dəstəkləyən geniş sahəli test mühitidir [13]. OCT bulud texnologiyaları tədqiqatlarını inkişaf etdirmək, müxtəlif bulud arxitekturlarının cəlb etdiyi eksperimental tədqiqatların aparılmasını mümkün etmək, bulud interoperabeliyinin öyrənilməsi üçün platforma təqdim etmək, bulud texnologiyaları bencmarklarını dəstəkləmək məqsədi ilə yaradılmışdır. Hazırda o ABŞ-ın Baltimor (John Hopkins University), Çikaqo (University of Illinois at Chicago və StarLight) və San Dieqo (California Institute for Telecommunications and Information Technology) şəhərlərindəki dörd verilənlər mərkəzini özündə birləşdirir. OCT təşəbbüsünün ikinci fazasında buraya həmçinin Kembric (MIT Lincoln Lab) və Pitsburq (Pittsburgh Supercomputer Center / Carnegie Mellon University) şəhərlərindəki daha iki verilənlər mərkəzi də əlavə edilmişdir.

Bu test mühiti digər paylanmış infrastrukturlar üçün problem sayılan olduqca böyük verilənlər axınının tələblərini ödəyə bilir. OCT platformasının tərkibində Eucalyptus, Hadoop, CloudStore (KosmosFS), Sector/Sphere və Thrift kimi müxtəlif bulud sistemləri və xidmətləri quraşdırılmışdır və onlar sistemdə tədqiqatların aparılması üçün əlçatar formada mövcuddur. OCT-nin tərkibinə daxil olan bu sistemlər bencmark tədqiqatların aparılması prosesini asanlaşdırır. Burada həmçinin bulud texnologiyaları stekində eksperimental tədqiqatların həyata keçirilməsini mümkün etmək üçün şəbəkə kitabxanaları, monitorinq sistemləri də yaradılmışdır.

Elmi tədqiqatların aparılması üçün əlçatar olan digər bulud test mühitlərinə Science Clouds, Virtual Computing Lab, Future Grid, Grid’5000, Okeanos, Illinois Cloud Computing Testbed (CCT), IEEE Intercloud Testbed infrastrukturlarını göstərmək olar.

V. BÖYÜK VERİLƏNLƏR VƏ BULUD TEST MÜHİTLƏRİ

Hazırda mövcud olan bulud test mühitləri böyük verilənlər axınının emalında öz təsirini itirir. Bu səbəbdən yuxarıda sadalanan bulud simulyatorlarının, test mühitlərininin böyük verilənlərin emalında yararlı olmasını təmin etmək üçün bu bulud vasitələrinin tərkibinə paralelləşmə yarada bilən MapReduce və s. modellərini daxil edirlər. Bu tip yanaşmalara [14, 15] ədəbiyyat mənbələrində baxılmışdır.

NƏTİCƏ

Test mühitləri tədqiqatçıların parallel və bulud texnologiyaları sahəsində innovativ yanaşmalarını reallaşdırmaq və test etmək məqsədi ilə yaradılır. Mövcud bulud test mühitləri tədqiqatçılara paylanmış sistemlərdən istifadə edərək öz eksperimentlərini həyata keçirmək imkanı verməkk üçün onlara infrastruktura təqdim edir. Lakin bulud test mühitləri böyük verilənlərlə bağlı tədqiqatların aparılmasını dəstəkləyə bilmir. Bu test mühitlərinin böyük verilənlərin tələblərini ödəmək məqsədi ilə onların modifikasiyası istiqamətində aktiv şəkildə araşdırmalar aparılmalıdır. Bu tip araşdırmalar həmçinin nəhəng elmi maliyələşdirmə qurumlarının da prioritet məsələləri siyahısına daxil edilməlidir.

MİNNƏTDARLIQ

Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişafı Fondunun maliyyə yardımı ilə yerinə yetirilmişdir – **Qrant № EİF-2013-9(15)-46/16/1**

ƏDƏBİYYAT

[1] G. Sakellari, G. Loukas, “A survey of mathematical models, simulation approaches and testbeds used for research in cloud computing,” Simulation Modelling Practice and Theory, 2013, Vol. 39, pp. 92–103.

[2] NSF Report on Support for Cloud Computing, 2012, 21 p.

[3] Enabling a new future for cloud computing, 2014, http://nsf.gov/news/news_summ.jsp?cntn_id=132377

[4] A.K. Pathan, M.M. Monowar, S. Khan, “Simulation technologies in networking and communications: Selecting the best tool for the test,” 2014, 648 p.

[5] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, D. Epema, “A performance analysis of EC2 cloud computing services for scientific computing, Proc. of the first international conference cloud computing (CloudComp),” 2009, pp. 115-131.

[6] A. Iosup, S. Ostermann, M. Yigitbasi, R. Prodan, T. Fahringer, D. Epema, “Performance analysis of cloud computing services for many-tasks scientific computing, IEEE transactions on parallel and distributed systems,” 2011, Vol. 22, No 6, pp. 931-945.

[7] S. Islam, K. Lee, A. Fekete, A. Liu, “How a consumer can measure elasticity for cloud platforms, Proc. of the 3rd ACM/SPEC international conference on performance engineering,” 2012, pp. 85–96.

[8] J. Wang, P. Varman, C. Xie, “Avoiding performance fluctuation in cloud storage, Proc. of the international conference on high performance computing,” 2010, pp. 1–9.

[9] A. Iosup, N. Yigitbasi, D. Epema, “On the performance variability of production cloud services, Proc. of the 11th IEEE/ACM international symposium on cluster, cloud and grid computing (CCGrid),” 2011, pp. 104-113.

[10] J. Alemany, X. Perramon, L. Panadès, “Improving the quality of the practicum. The use of moodle and google docs in monitoring the practicum process after the EHEA,” 2012.

[11] W. Lu, J. Jackson, J. Ekanayake, R. Barga, N. Araujo, “AzureBlast: A case study of developing science applications on the cloud,” Proc. of the 19th ACM international symposium on high performance distributed computing, 2010, pp. 413-420.

[12] C. Baun, M. Kunze, “Performance measurement of a private cloud in the OpenCirrus testbed,” Proc. of the parallel processing workshops, 2010, pp. 434-443.

[13] R. Grossman, Y. Gu, M. Sabala, C. Bennet, J. Seidman, J. Mambratti, “The open cloud testbed: A wide area testbed for cloud computing utilizing high performance network services,” Proc. of the GridNets, 2009, pp. 1-6.

[14] Jung J., Kim H., “MR-CloudSim: Designing and implementing MapReduce computing model on CloudSim,” Proc. of the international conference on ICT convergence (ICTC), 2012, pp. 504-509.

[15] Laszewski G., Fox G. “The FutureGrid testbed for big data, Cloud computing for data-intensive applications,” 2014, pp. 27-59.

Böyük Həcmli Mətnlərin Yaxınlığının Müəyyənləşdirilməsində Big Data Texnologiyalarından İstifadə

Nadir Ağayev[1], Əminə Ağazadə[2]
[1]Milli Aviasiya Akademiyası, Bakı, Azərbaycan
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*nadir_avia@yahoo.com*, [2]*emina635@hotmail.com*

***Xülasə* — Məqalədə böyük həcmli mətnlərin yaxınlığının müəyyənləşdirilməsində rast gəlinən problemlər araşdırılmış, informasiya həcminin sürətlə artdığı şəraitdə bu problemlərin həlli üçün big data texnologiyalarından istifadə təklif edilmişdir.**

Açar sözlər — mətnlərin yaxınlığının müəyyənləşdirilməsi; antiplagiat sistemləri; Big data texnologiyaları.

I. GİRİŞ

Hazırda qloballaşan dünyamızda material, enerji və maliyyə resurslarından daha çox informasiya resurslarına tələbat hiss edilir. Bunu bilavasitə qloballaşmanın təzahürü kimi müxtəlif sivilizasiyaların inteqrasiyası və iqtisadiyyatın məhdud coğrafi məkandan transmilli məkana inikası nəticəsində idarəetmədə düzgün qərarın operativ, dəqiq və tam informasiyanın olmasından bilavasitə asılılığı ilə izah edilir. Təbii ki, eyni zamanda hamı tərəfindən daha operativ, daha dəqiq və tam informasiyanın əldə edilməsi istəyi informasiya mənbələrindən istifadədə müəyyən problemlər yaradır. Bu problemlər bir tərəfdən informasiyanın həcminin artımı ilə əlaqədar onların saxlanılması və emalında yaranan problemlər, digər tərəfdən informasiyanın əldə edilməsinin qanuniliyinə nəzarətə tələblərin yüksəlməsi ilə əlaqədar rast gəlinən problemlərdir. Hər iki halda problemin mənbəyində böyük həcmli informasiyanın emalı səbəbindən yaranan problemlər durur. Bu problemləri tədqiqatçılar hazırda big data texnologiyalarından istifadə etməklə həll etməyə çalışırlar [1]. Bu məqalədə biz informasiyanın əldə edilməsinin qanuniliyinə nəzarət məsələlərinin həllində big data texnologiyalarından istifadə imkanlarını araşdıracağıq.

II. MƏTNLƏRİN YAXINLIĞININ MÜƏYYƏNLƏŞDİRİLMƏSİ TEXNOLOGİYALARI

Hazırda informasiya texnologiyasının (İT) inkişafı elə səviyyəyə çatmışdır ki, cəmiyyət tərəfindən istifadə edilən istənilən informasiyanı elektron daşıyıcılarda saxlamağa imkan verir. Bu informasiyalar içərisində mətn informasiyalar həcmcə digər növ informasiyalardan daha yüksək çəki əmsalına malikdirlər. Məhz bu səbəbdən hazırda mətnlərin analizi və ilkin emalı məsələlərinin həllinə yönəldilmiş sistemlərin yaradılması əsas problemlərdən biri kimi tədqiqatçılar qarşısında dayanır. Bu problem xüsusilə informasiyadan qanuni istifadəyə nəzarət sistemlərinin (qlobal və lokal şəbəkələrdə yerləşdirilmiş informasiyalar üçün bu sistem antiplagiat sistemləri adlanır) yaradılması zamanı daha sərt şəkildə özünü biruzə verir. Bu sistemlərdə -"Findsame", Eve2, Turnitin, Copy Catch, Word CHECK, "Advego Plagiatus", ANTIPLAGIAT və digər sistemlərdə [2, 3, 4, 5, 6] mətnlərin yaxınlığı dörd mərhələdə yoxlanılır [7]:

I mərhələ: müxtəlif mənbələrdən – İnternetdən, referat və elmi məqalələrin açıq və ya icazəli bazalarından mövcud axtarış sistemləri – Google, Yahoo, Yandex və s. vasitəsi ilə informasiyanın toplanması və sistemləşdirilməsi.

II mərhələ: toplanmış informasiyanın filtrasiyası (reklamlardan, menyulardan, başlıqlardan və s. təmizlənməsi).

III mərhələ: informasiyanın sistemin tələblərinə uyğunlaşdırılaraq məlumat bazasına daxil edilməsi və növbə ilə yoxlanması.

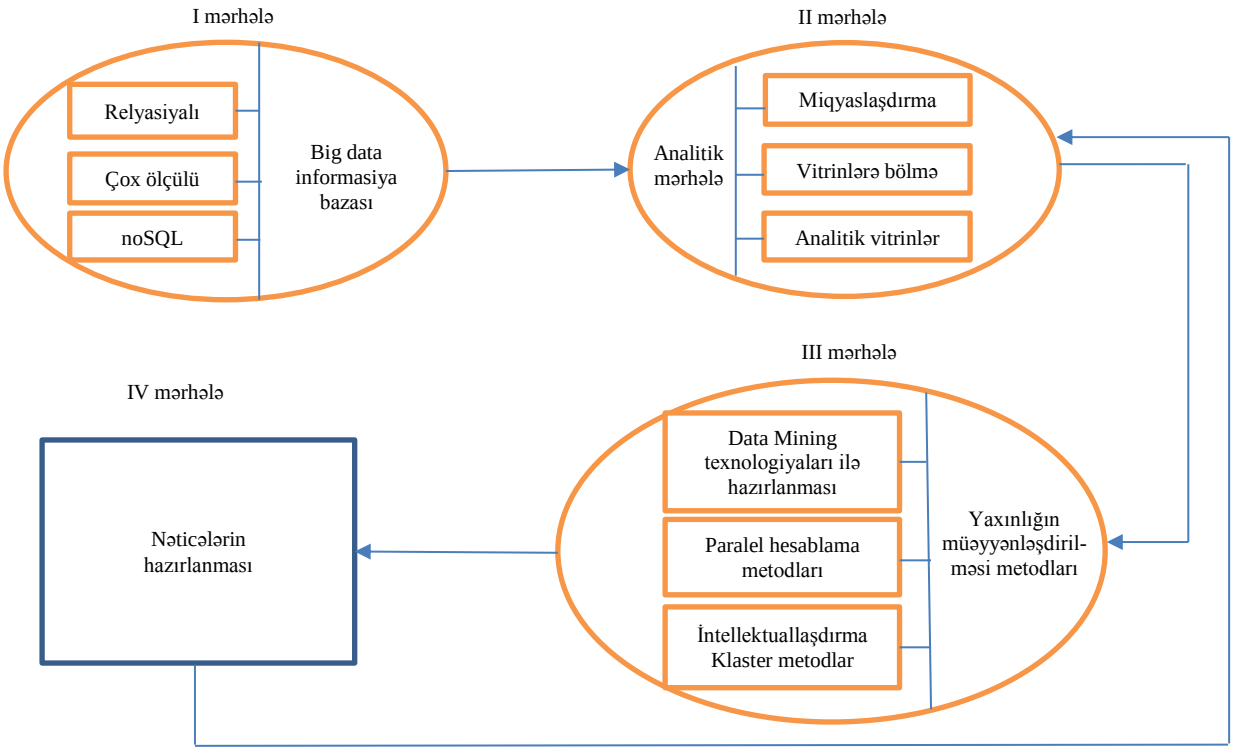
IV mərhələ: normal başa çatmış yoxlamadan sonra nəticələr haqqında hesabatın hazırlanması.

Yuxarıda qeyd edilən mərhələlər yaxınlığın aşkarlanmasının nəzəri, proqram və texniki realizasiyasından asılı olmayaraq istənilən sistemlərdə tətbiq edilir. Mərhələlərdə adi sorğulardan və informasiyanın klassik emalı metodlarından kəmiyyət və keyfiyyətcə fərqli texnologiyaların tətbiq edilməsi zərurəti təkcə böyük həcmli informasiya ilə işləmək tələbi deyil, həm də müəyyən edilmiş informasiyanın elə çevrilməsidir ki, ondan sonrakı mərhələlərdə və qərar qəbul etmədə istifadə edilməsi mümkün olsun. Bir neçə il əvvəl yoxlanılan informasiyaların həcmi ən çoxu terabaytla (Fast Data) ölçülürdü. Bu halda emal prosesini izləmək və əvvəlcədən irəli sürülən hipotezləri statistik yoxlamaq mümkün olduğundan qərar qəbul etmədə klassik metodlara əsaslanaraq yaxınlığı müəyyən etmək olurdu. Bu metodlarda yaxınlıq ölçüləri kimi Jakkard, Days, kosinus və s. [8] klassik riyaziyyatdan məlum olan yaxınlıq ölçülərindən istifadə edilir və əsasən mətnlərdə verilmiş sözlər ardıcıllığının müəyyən qaydalarla hesablanmış xarakteristikalarına əsasən (sintaksis metodlar) onların eyniliyi müəyyənləşdirilir. Məhz bu səbəbdən qeyd edilən prinsiplə qurulmuş metodlar əsasında yaradılmış antiplagiat sistemləri məhdud həcmli informasiyalarla işləyə bilir.

Sonrakı illərdə informasiyanın həcmi artdıqca problemi həll etmək üçün yeni yanaşmalarla verilənlər bazası yaradılsa da (Greenplum, Netezza, Oracle Exadata, Teradata, Vertica tipli verilənlər bazasının idarəetmə sistemi (VBİS)) [1] problem həll edilmədi. Bunun əsas səbəbi axtarılan informasiyanın təkcə həcmcə (informasiyanın həcmi petabaytlarla ölçüldükdə) deyil, həm də keyfiyyətcə dəyişiklikləri idi (Big Data Analytics). Bu halda yaxınlığın yoxlanılması yanaşmalarında artıq intellektuallıq elementlərindən istifadə cəhdləri göstərilirdi. Belə metodlarda müxtəlif yanaşmalarla - lüğətlər, etalon sözlər və terminlər bazasından və s. istifadə etməklə (bu metodlar leksik metodlar adlanır) emal edilən informasiyanın həcmini sintaksis metodlara nisbətən artırmaq mümkün oldu. Lakin bu metodlarda öyrətmə prosesi “müəllimlə öyrətməyə” əsaslandığından intellektuallıq zəif strukturlaşdırılmış və ya strukturlaşdırılmamış informasiyalar üçün gözlənilən nəticəni vermir. Məsələnin həlli üçün son dövrlərdə təklif edilən klasterləşdirmə metodları da eyni problemi yaşayır.

III. BIG DATA TEXNOLOGİYALARINDAN İSTİFADƏ

Hazırda internet və elektron daşıyıcılarda yerləşdirilmiş mətn tipli informasiyalar əksər hallarda strukturlaşdırılmamış, zəif strukturlaşdırılmış və yaxud müxtəlif tipli mənbələrdən alınaraq yaradılmış və özündə fərqli informasiya tiplərini (audio, video, mətn, verilənlər bazası) birləşdirir. Bu informasiyanın çox az hissəsi, əsasən strukturlaşdırılmış məlumatlardan təşkil edilmiş verilənlər bazası metaverilənlərə malik olur, digərlərində az, yaxud tamamilə olmur. Bu hal əsasən İnternet şəbəkələrində xaotik şəkildə, müəyyən subyektiv faydalılığı nəzərdə tutaraq (reklam, şəxsi veb-saytlar, sosial şəbəkələrdə yerləşdirilmiş məlumatlar və s.) yerləşdirilmiş informasiyalara aid olur. Bu növ informasiyadan mövcud analitik emal metodlarını tətbiq etməklə bilavasitə insan-ekspertin iştirakı ilə xülasə məsələlərin həllində istifadə etmək olur. Hazırda bu emal metodları güclü riyazi aparata malikdir və qruplaşdırılaraq müxtəlif sistemlərdə - Data Mining, Business Intelligence və s. sistemlərdə [9] müvəffəqiyyətlə istifadə edilir. Lakin bu metodlar ancaq rəqəmli verilənlərə aiddir və zəif strukturlaşdırılmış informasiyanın analizi hazırda yalnız insan-analitikin iştirakı ilə mümkün olur. Digər tərəfdən internetdə mövcud informasiyanın böyük əksəriyyəti zəif strukturlaşdırılmış informasiya olduğundan onlara klassik analiz metodlarını tətbiq emək mümkün olmur. Hazırda informasiyanın həcminin ekzabayt və zetabaytla ölçüldüyü bir şəraitdə yaxınlığın yoxlanılması metodları tamamilə strukturlaşdırılmamış informasiya ilə işləmək bacarığı ilə yanaşı emalın və qərar qəbuletmənin yüksək sürətini də təmin etməlidir. Metodların tətbiqi üçün böyük həcmli informasiya ilə işləyə bilən xüsusi kompüterlər və onlar üçün proqram-aparat sisteminin yaradılması təklif edilir (Şəkil 1). İnformasiyanın saxlanılması və emalında belə sistemlər - SAP HANA, Oracle Big Data Appliance, Oracle Exadata Database Machine, Oracle Exalytics Business Intelligence Machine, Teradata Extreme Performance Appliance, NetApp E-Series Storage Technology, IBM Netezza Data Appliance, EMC Greenplum, HP Converged Infrastructure bazasında Vertica Analytics Platform [10] yaradılmış və müxtəlif məsələlərin həllində istifadə edilir. Bu sistemlərdən böyük həcmli mətnlərin yaxınlığının müəyyən edilməsi üçün istifadə edilə bilər. Yaxınlığı müəyyən edilmiş məlumatlar sonrakı emalda



Şəkil 1. Big data texnologiyalarından istifadə edərək yaxınlığın müəyyən edilməsi sxemi

istifadə edilməsi üçün yaddaşda saxlanılır. Daxili və xarici mənbələrin mürəkkəb əlaqələri və məlumatların fərqliliyi - strukturlaşdırılmış, strukturlaşdırılmamış və kvazistruktur-laşdırılmış məlumatlar üçün müxtəlif indeksləşdirmə (relyasiyalı, çoxölçülü, noSQL) sxemləri tətbiq etmək olar [11]. Bunun üçün ilkin xammal məlumatlar yuxarıda qeyd edilən sistemlər əsasında qurulmuş analitik maşının girişinə verilir. Məlumatların emal edilməsi üçün aralıq vitrinlər (Independent Data Mart, IDM) [12] yaradılır, sonra isə onlar xüsusi analitik vitrinlərə (Analytical Data Mart, ADM) yığılır və yaradılmış yaxınlığın müəyyən edilməsinin metod və alqoritmləri tətbiq edilir. Nəticələr In-Database Analytics və ya No-Copy Analytics yanaşmalarını tətbiq etməklə bilavasitə bu məqsəd üçün yaradılmış bazalarda saxlanıla bilər [13]. Belə bazalar adətən paralel və ya analitik VBİS kimi qurulmalıdır. Sonrakı yoxlama prosesi bilavasitə tam strukturlaşdırılmış “hazır” məlumatlar bazasından ibarət olduğundan emal prosesi də nisbətən sürətlənir. Yaxınlığın müəyyən edilməsində analitik vitrinlərdə yerləşən məlumatlara müxtəlif paralel hesablama metodlarını tətbiq etmək olar.

NƏTİCƏ

Məqalədə internet şəbəkəsində böyük həcmli informasiyanın mövcudluğunun informasiyadan qanuni istifadədə yaratdığı problemlər araşdırılmışdır. Göstərilmişdir ki, belə informasiyalar içərisində strukturlaşdırılmamış və yaxud zəif strukturlaşdırılmış mətn tipli informasiyalar əhəmiyyətli çəkiyə malikdir. Bu halda mətnlərin yaxınlığının müəyyən edilməsində klassik metodların tətbiqi gözlənilən nəticəni vermir. Bu səbəblə, böyük həcmli mətn informasiyaların yaxınlığının müəyyən edilməsində big data texnologiyaları əsasında yaradılmış proqram-aparat sistemlərindən istifadə edilməsi təklif olunmuşdur. Belə ki, daxili və xarici mənbələrin mürəkkəb əlaqələri və məlumatların fərqliliyi şəraitində müxtəlif indeksləşdirmə (relyasiyalı, çoxölçülü, noSQL) sxemləri tətbiq etməklə Big Data Analytics üsulları ilə yaxınlığın tam avtomatlaşdırılmış şəkildə müəyyən edilməsi təklif edilmişdir. Nəticələrin In-Database Analytics və ya No-Copy Analytics yanaşmalarını tətbiq etməklə bilavasitə bu məqsəd üçün yaradılmış bazalarda saxlanıldığından və emal prosesində müxtəlif paralel hesablama metodları, həmçinin miqyaslaşdırma ilə klaster metodları tətbiq etməklə qərar qəbuletmə prosesini əhəmiyyətli dərəcədə yüksəltmək olar.

ƏDƏBİYYAT

[1] В. М. Шенбергер, К. Кукьер, “Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим”, 221 с., Москва, 2014.

[2] A. H. Osman, N. Salim, Y. J. Kumar, A. Abuobieda, “Fuzzy Semantic Plagiarism Detection” // Journal of Advanced Machine Learning Technologies and Applications Communications in Computer and Information Science, Vol. 322, pp 543-553, 2012.

[3] S. Sandhya, S. Chitrakala, “Plagiarism Detection of Paraphrases in Text Documents with Document Retrieval” // Journal of Advances in Computing and Information Technology Communications in Computer and Information Science, Vol. 198, pp 330-338, 2011.

[4] R. Əliquliyev, N. Ağayev, R. Alıquliyev, “Kompüter proqramları piratçılığı və antiplagiat sistemlərindən istifadə problemləri” // “İnformasiya texnologiyaları problemləri” jurnalı, №1, səh. 3-9, 2015.

[5] http://turnitin.com

[6] http://www.canexus.com

[7] R. Əliquliyev, N. Ağayev, R. Alıquliyev, “Plagiatlıqla mübarizə texnologiyaları”, Bakı, “İnformasiya Texnologiyaları” nəşriyyatı, 2015, 165 səh.

[8] H. Finch, “Comparison of Distance Measures in Cluster Analysis with Dichotomous Data” // Journal of Data Science Vol. 2, No. 3, pp 85-100, 2005.

[9] N. Kerdprasop, K. Kerdprasop “Moving Data Mining Tools toward a Business Intelligence System”, International Journal of Intelligent Technology, Vol. 2 No. 2, pp. 99-104, 2007.

[10] https://www.vertica.com/wp-content/uploads/2014/05/VerticaOverview.pdf

[11] C. Ji, Y. Li, W. Qiu, Y. Jin, Y. Xu, U. Awada, K. Li, “Big data processing: big challenges and opportunities”, Journal of Interconnection Networks, Vol. 13, No. 3 & 4, pp. 1-19, 2012.

[12] R. Chhabra, P. Pahwa, “Data Mart Designing and Integration Approaches”, International Journal of Computer Science and Mobile Computing, Vol.3, No. 4, pp 74-79, April 2014.

[13] http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataanalyticswpoaa-1930891.pdf

Böyük Videoinformasiya Resurslarının Emalı Problemləri Haqqında

Bikəs Ağayev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
depart6@iit.ab.az

***Xülasə*— Məqalədə cəmiyyətin inkişaf səviyyəsi, istehsal edilən informasiyanın həcmi və informasiya daşıyıcılarının tutumu arasındakı səbəb-nəticə əlaqələri araşdırılır. İnformasiyanın klassik emal texnologiyalarından Big Data-nın yaranma zərurəti analiz edilir. Big Data texnologiyalarının videoanalitika istiqamətinin videonəzarət sistemlərinə tətbiqi problemləri, həll nəticələri tədqiq edilir.**

Açar sözlər— informasiya resursları; sıçrayışlı informasiya istehsalı; informasiya inqilabları; informasiya daşıyıcıları; Big Data; audio-video informasiya; videoanalitika; videonəzarət

I. İSTEHSAL EDİLƏN İNFORMASİYA HƏCMİ CƏMİYYƏTİN İNKİŞAF SƏVİYYƏSİNİN MEYARIDIR

İnsanlar öz aktiv əmək fəaliyyətini biliklər və qazanılmış təcrübə əsasında yaratdığı texnika və informasiya vasitələrindən istifadə etməklə reallaşdırır. Bu vasitələrin inkişafı cəmiyyətin yeni inkişaf mərhələlərinə keçməsinə, maddi və mənəvi dəyərlərin köklü dəyişikliklərinə səbəb olur. Cəmiyyətin sosial-mədəni, siyasi-iqtisadi təkamülü həmin dövrlərdə informasiyanın əldə edilməsi və istifadə edilməsi sahəsindəki sıçrayışlı inkişafı zəminində baş verir. Bu gün formalaşmaqda olan İnformasiya Cəmiyyəti də məhz insanların informasiyanın əldə edilməsinə, istifadə edilməsinə olan zəruri tələbatının ödənilməsi forması kimi yaranmaqdadır.

Cəmiyyətin informasiyalaşması fəlsəfəsini, informasiya bolluğu və inkişaf arasındakı əlaqələri tədqiq edən görkəmli alim, akademik Moiseyev N.N. yazır [1]: “Cəmiyyət tarixinin öyrənilməsini tarixin bir informasiya prosesi kimi öyrənilməsi işinə gətirmək olar, çünki hər bir cəmiyyətin iqtisadi, mədəni və s. sahələrdə inkişafına təsir edən əsas amillərdən biri cəmiyyət tarixinin informasiya tərkibidir”. Digər görkəmli alim N.A.Neqodayev sivilizasiyanın inkişaf səviyyəsi ilə cəmiyyətin istehsal etdiyi informasiya tutumu arasında bir əlaqənin mövcudluğunu qeyd edir. Onun fikrincə, informasiya azdırsa, cəmiyyətin inkişaf səviyyəsi də aşağıdır və əksinə, informasiya bolluğu yarandıqca cəmiyyət də inkişaf edir. Cəmiyyətin inkişafı, onun dünəni, bu günü və sabahı texniki vasitələrdən, xüsusən də, informasiyanın mənimsənilməsi və ondan istifadə edilməsi prosesləri ilə üzvi surətdə bağlıdır. Bu nöqteyi-nəzərdən informasiya cəmiyyətin sosial-mədəni, siyasi-iqtisadi inkişafının mühüm hərəkətverici qüvvəsi olmuşdur. Başqa sözlə, cəmiyyətin yaratdığı və istifadə etdiyi informasiyanın miqdarı bu cəmiyyətin inkişafının ölçü meyarı xüsusiyyətinə malikdir.

Alim bu postulatdan çıxış edərək sivilizasiyanı istehsal olunmuş informasiyanın həcminə müvafiq olaraq aşağıdakı səviyyələrə bölür [2]:

- səviyyə 0 – ilk insanın beyninin informasiya tutumu – 1,0 Kb;
- səviyyə 1 – icma, tayfa və ya kənd daxilində dövr edən şifahi nitq informasiyasının tutumu 10,0 Kb;
- səviyyə 2 – yazılı mədəniyyət mövcuddur. İnformasiya cəmiyyətinin ölçüsü kimi 532800 ədəd lülə formalı əlyazmalardan ibarət olan İsgəndəriyyə kitabxanasının informasiya tutumu 1,0 Gb qəbul olunur;
- səviyyə 3 – kitab mədəniyyəti mövcuddur. Bu dövrdə dərc olunan kitabların, jurnalların ümumi informasiya tutumu 0,1 Tb həddində qiymətləndirilir;
- səviyyə 4 – İnformasiyanın elektron emal texnologiyaları meydana gəlmişdir. Bu dövrdə istehsal olunan informasiyanın tutumu təxminən 10,0 Pb-dir.

Əgər informasiya bolluğu cəmiyyətin bütün fəaliyyət sahələrinin inkişafını şərtləndirirsə elm, texnika və texnologiyaların inkişafı da öz növbəsində daha böyük həcmdə informasiya hasil edən və ya istifadə edən vasitələrin (xidmətlərin, münasibətlərin və s.) yaranması ilə nəticələnir və bu proses dövrü olaraq təkrarlanır. Növbəti yaranan vasitələr özündən əvvəlki ilə müqayisədə daha məhsuldar, daha güclü informasiya generatoru olması səbəbindən, yəni daha dinamik inkişafa səbəb olduğu üçün növbəti vasitənin yaranmasını sürətləndirir, mövcudluq dövrünü qısaldır. Bəşəriyyətə məlum olan informasiya inqilablarının xronologiyasını nəzərdən keçirsək bu tezislərin düzlüyünə əmin olarıq. Həqiqətən:

- insanın təşəkkül tapması ilə I informasiya inqilabı (nitqin yaranması) arasındakı müddət – 1,5 mln. il;
- I və II informasiya inqilabı (yazının yaranması) arasındakı müddət – təxminən 35-37 min il;
- II və III informasiya inqilabı (kitab çapının ixtirası – 1497) arasındakı müddət – 4,5 min il;
- III və IV informasiya inqilabı (elektrik cərəyanının kəşfi nəticəsində teleqrafın (1844), telefonun (1876), radionun (1868), televizorun (1932) icad edilməsi) arasındakı müddət – 400 il;

- IV və V informasiya inqilabı (1971-1981 – fərdi kompüterlərin və 1969-1989 – İnternetin kütləvi istifadəsi) arasındakı müddət – 100 il.

Bu gün sivilizasiya elə bir inkişaf səviyyəsinə çatmışdır ki, özünün istehsal etdiyi çox böyük həcmli informasiyanı axın halında onlayn rejimdə, istərsə də bazalarda arxivləşdirilmiş informasiyanı oflayn rejimdə emal etməkdə çətinlik çəkir. Bu tip problemlərin həlli zərurəti informatika elminin yeni bir istiqaməti olan Böyük Verilənlər (ingilis dilli ədəbiyyatda – Big Data) adlandırılmış tədqiqat sahəsinin (texnologiyalarının) yaranmasına səbəb olmuşdur.

Məqalədə audio-video informasiyanın Big Data texnologiyaları ilə səbəb-nəticə əlaqələri aydınlaşdırılır, bu tip informasiyanın hazırkı emalı vəziyyəti və problemləri araşdırılır.

II. İNFORMASİYA ARTIMININ SƏBƏB-NƏTİCƏ ƏLAQƏLƏRİ

Yuxarıda qeyd edildiyi kimi son bir neçə on illikdə müxtəlif mənbələr tərəfindən generasiya edilən strukturlaşmış və strukturlaşdırılmamış informasiyanın artım dinamikası elə bir həddə çatmışdır ki, artıq mövcud klassik qaydalarla bu qədər informasiyanı qaneedici keyfiyyət göstəriciləri çərçivəsində (müddət, dəqiqlik və s.) həll etmək mümkün olmur. Burada “emal” dedikdə təkcə verilənlər üzərində bilavasitə hesablama vasitələri ilə riyazi və məntiqi əməliyyatların aparılması yox, eyni zamanda verilənlərin toplanması, qeydiyyatı, arxivləşdirilməsi və ötürülməsi prosesləri də nəzərdə tutulur [3].

Digər tərəfdən informasiyanın Big Data kateqoriyasına aidiyyatı emal nöqteyi-nəzərindən onun üç xüsusiyyətdən ibarət çoxluqdan istənilən kombinasiyaya malik olması imkanını nəzərdə tutur: həcmin böyüklüyü, yüksək ötürmə sürəti, tip müxtəlifliyi [4]. Deyilənlər xüsusilə müasir verilənlər bazalarının strukturlaşma formatlarına uyğun gəlməyən strukturlaşdırılmamış verilənlər axınının (audio-video trafik, veb mətnlər, e-mail məlumatları, texnoloji sensorların axınları və s.) onlayn emalına aiddir. Formalaşmaqda olan Big Data texnologiyalarının predmeti də qeyd olunan hər üç meyara malik informasiyanın emalı üçün yeni yanaşmaların və üsulların işlənməsidir.

Verilənlərin həcm böyüklüyü Big Data texnologiyaları üçün əsas problem hesab edilir. Bu problemin mahiyyətini ədədi qiymətlərlə xarakterizə edək. Texniki ədəbiyyatda informasiyanın həcmə görə Big Data-ya aidiyyatı məsələsində fikir müxtəlifliyi var. İnformasiya həcminin artım dinamikası/emal üsullarının funksionallığı (imkanları) aspektindən xronoloji qaydada araşdırmaqla aşağıdakı təsnifatın daha məqsədəuyğun olması fikrinə gəlmək olar.

a) Klassik üsullarla emal edilən verilənlər.

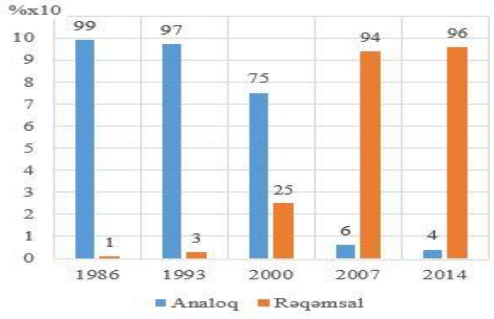
- böyük həcmli verilənlər – yüzlərlə qiqabaytadək;
- ifrat böyük həcmli verilənlər – yüzlərlə terabaytadək.

b) Big Data metodları ilə emal edilən verilənlər – onlarla petabaytadək.

Bir sıra tədqiqatçılar yaxın gələcəkdə daha böyük həcmli (ekzabaytlarla) verilənlərin emalı məqsədilə növbəti texnologiyaların işləniləcəyini proqnozlaşdırır və onu Post Big Data, Extermaly Big Data və s. adlandırırlar [5].

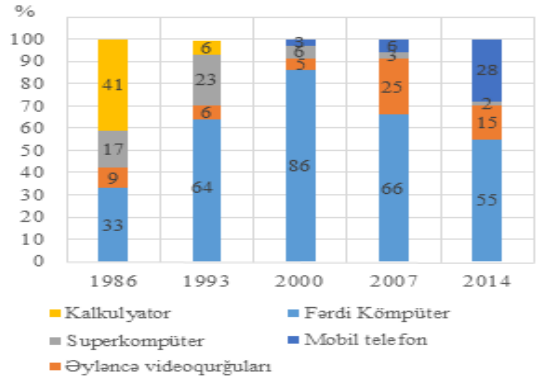
Aparılan araşdırmalar göstərir ki, informasiya həcmi və emal keyfiyyəti arasında korrelyativ əlaqə var və müəyyən zaman kəsiyində stabil qalır: generasiya edilən informasiyanın artan həcm dinamikasına uyğun emal üsulları işlənir. Bu proses az-çox aydın seçilən periodlar (dövrlər) üçün doğrudur, yəni period daxilində korrelyasiya əmsalını stabil qəbul etmək mümkündür. Lakin bir neçə perioddan sonra informasiyanın həcmi xətti artım tendensiyasından çıxır və çox qısa zamanda sıçrayışlı artım yaranır. Bu proses bir qayda olaraq texnika və texnologiyalar sahəsində qlobal əhəmiyyətli yeniliyin yaranması ilə şərtlənir (birinci informasiya inqilabı adlanan nitqin yaranması istisnadır). Nəticədə həcm/emal adekvatlığı pozulur, korrelyasiya əmsalı kəskin azalır. Yaranmış uyğunsuzluğu aradan qaldırmaq üçün yeni metodlar, texnologiyalar axtarışı zərurəti yaranır. Əgər informasiya həcminin son sıçrayışlı artımını IBM markalı fərdi kompüterlərin kütləvi istehsalı/istehlakı (1971-1981), İnternetin korporativlikdən kütləviliyə keçməsi (1969-1989), Web texnologiyaların yaranması (1990) və s. əlaqələndirib V informasiya inqilabı adlandırılırsa, sıçrayışlı informasiya generasiyası prosesini də, uyğun olaraq, VI informasiya inqilabı adlandırmaq olar. Bu nöqteyi-nəzərdən Big Data adlandırılan texnologiya da son illərdə baş verən sıçrayışlı informasiya artımının yaratdığı problemləri həll etmək zərurətindən yaranır. Bu inqilabın yaranması üçün zəmin yaradan səbəbləri (texniki şərtləri) son 30-35 ilin inkişaf dinamikasında, ümumi şəkildə araşdıraq.

İnformasiya sıçrayışına ilk səbəb kimi bu dövr ərzində informasiyanın bütün təsvir formaları (tipləri) üzrə rəqəmləşdirmə prosesinin başa çatması və effektiv kodlama metodlarının yaradılmasını göstərmək olar. Məsələn, müasir texnoloji sensorlar konstruktiv olaraq diskretləmə və kvantlama əməliyyatlarını aparan qovşaqla təchiz edilir (o cümlədən radio vericili sensorlarda), audio-video yazma/oxuma avadanlığı rəqəmləşdirilib. Bu isə asanlıqla daha çox rəqəmsal informasiya hasil etmək deməkdir (şək. 1).



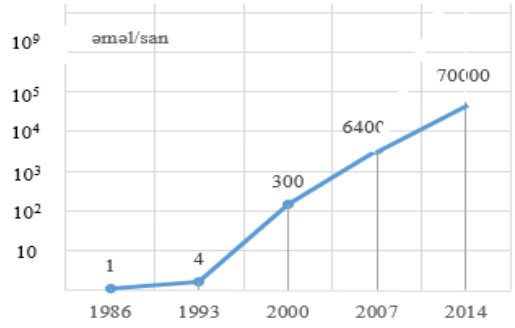
Şəkil 1. Verilənlərin təsvirinin analoq-kod forma transformasiyası

İkinci səbəb kimi verilənlərin emalı vasitələrindən istifadə məqsədlərinin transformasiyasıdır. Həqiqətən, fərdi kompüterlərin, o cümlədən kompüter bazalı mobil telefonların qiymətlərinin əlyetərliyi, kompüterləşmə, şəbəkələşmə istiqamətindəki addımlar və s. bu vasitələrdən istifadə səviyyəsini kəskin yüksəldərək hesablama texnikasını korporativ emal vasitəsindən kütləvi emal vasitəsinə çevirmişdir (şək. 2).



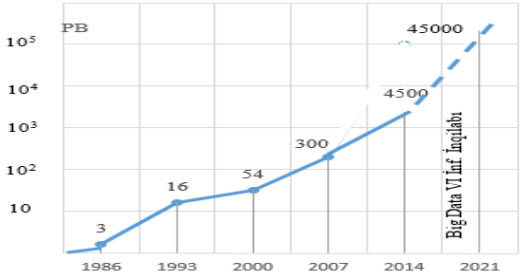
Şəkil 2. Verilənlərin emalı vasitələrinin transformasiyası

Üçüncü səbəb həmin dövr ərzində hesablama vasitələrinin məhsuldarlığının (hesablama sürətinin) artmasıdır. Şəkildən (şək. 3) göründüyü kimi məhsulu informasiya olan bu vasitələrin məcmu istehsal gücü kəskin artmışdır.



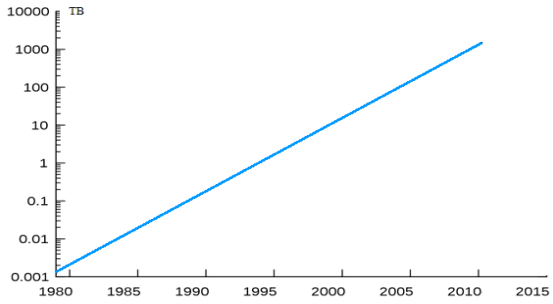
Şəkil 3. Hesablama texnikasının güc (məhsuldarlıq) dinamikası

Nəticədə bir sıra yeni texnika və texnologiyalardan, o cümlədən son nəsil texniki sensorlardan, sosial şəbəkələrdən, İnternet və mobil telefon tətbiqlərindən, böyük axtarış sistemlərindən və s. istifadə hazırda istehsal olunan informasiya həcmini artıraraq EB-lərə çatdırmışdır (şək. 4).



Şəkil 4. Verilənlərin artım dinamikası

Bəzi mənbələrin məlumatlarına görə 2020-ci ildə istehsal olunan informasiyanın həcmi 44 EB-dək artacaq [6]. Hal-hazırda əsas informasiya daşıyıcıları elektron daşıyıcılar olduğu üçün onların da funksionallığı, ilk növbədə tutum (həcm) göstəriciləri informasiya artımına adekvat olaraq yüksəlmişdir. Bu ilk növbədə hesablama vasitələrinin maqnit, maqnit-optik, optik, elektrik prinsipli xarici yaddaş qurğularına, elektron arxiv saxlanclarına aiddir. Məsələn, kütləvi istifadə olunan müasir fərdi kompüterlərin sərt disk qurğusunun (vinçesterin) həcmi 3-5 TB, ölçüsü petabaytlarla olan saxlanclarda isə 8-10 TB-dir. Son 35 ildə sərt diskin həcmi orta hesabla 10 mln. dəfədən çox artmışdır (şək. 5).



Şəkil 5. Xarici yaddaş qurğularının həcminin artım dinamikası

III. VİDEOANALİTİKA BIG DATA TEXNOLOGİYALARININ OBYEKTİ KİMİ

Böyük yaddaş və hesablama resursları tələb edən mənbələrdən biri də audio-video informasiyadır. Böyük həcm bu tip informasiyanın yazılışı və yenidən canlandırılması (uyğun olaraq dinlənilməsi və müşayiət edilməsi) məqsədilə rəqəmləşmə-kodlaşma texnikasının nəticəsi kimi yaranır. Məsələn, “a” simvolunu müasir “genişləndirilmiş ASCII” standartı ilə rəqəmləşdirmək üçün cəmi 8 bit informasiya istifadə edilirsə səslənmə müddəti orta hesabla cəmi 100 msan. olan “a” səsini (Azərbaycan əlifbasının) müasir DVD-audio formatında 24 dərəcəli kvantlama və 192 Khs tezliyi ilə diskretləyib kodlasaq 460 Kbit (24x192 Kbit/san.x0,1) informasiya həcmi sərf edilər. Əksetdirmə qurğularında (monitor, TV, kinoekran və s.) “a” səsini səsləndirən şəxsin təsvirini 100 msan. ərzində müşayiət etmək üçün yüksək keyfiyyətli son rəqəmsal HDMİ 1.3 video standartı ilə (2560x1440 pixel, 50 kadr/san.) 1020 Mbit informasiya yazıb generasiya etmək lazımdır.

Praktikada videoçəkilişlər (videoyazılış) və ya verilişlər zamanı çox böyük informasiya trafikini emal etmək lazım gəlir. Məsələn, AMEA İnformasiya Texnologiyaları İnstitutunun videonəzarət sistemində cəmi 16 ədəd IP-AVTECH markalı rəqəmsal “gecə-gündüz” tipli videokameralar quraşdırılmışdır və əraziyə sutkalıq nəzarət aparılır. Hər kameranın bitreytini orta hesabla 2 Mbit/san. qəbul etsək sistem 1 ay ərzində 43,8 TB informasiya yazılmış olar (real olaraq 3 ədəd 3 TB-lik HDD sərt diski 20 günlük yazılış üçün kifayət edir). Bakıda istismarda olan “Nəqliyyatın idarə edilməsi intellektual sistemi”-nin 1500-dən çox videokameraları 260 TB-dən çox informasiya hasil edir.

Aydındır ki, operatorun bu qədər informasiyanı (kadrları) monitorda onlayn rejimdə fasiləsiz izləməsi çox yorucu və mümkün olmayan işdir. Təcrübələr göstərir ki, operator hətta bir monitoru fasiləsiz və diqqətlə müşayiət edirsə 20-30 dəqiqədən sonra yorulur və 45% informasiyanı “görmür”. Bir saatdan sonra bu “itki” 80-90%-ə çatır və ən mühüm hadisələri ekranda görməmək ehtimalı kəskin artır [7]. Bir operator bir neçə monitora və ya dörddən artıq seqmentlərə bölünmüş bir monitora fasiləsiz nəzarət edirsə onun iş effektivliyi daha da azalır. Digər tərəfdən, böyük həcmli videobazadan lazım olan videofraqmentləri axtarıb tapmaq (oflayn rejimdə) xeyli vaxt tələb edən yorucu işdir. Ona görə də böyük həcmdə audio-video informasiyanın emalı prosesinin avtomatlaşdırılması operativlik, dəqiqlik nöqteyi-nəzərdən əhəmiyyətlidir və operatorların nəzarət funksiyalarının effektivliyini yüksəltmək məqsədi daşıyır. Audio-video verilənlərin analizi, o cümlədən analitik analizi ilə məşğul olan istiqamətlərdən biri də videoanalitikadır. Videoanalitikanın əsas funksiyası müəyyən hərəkət ardıcıllığı ilə baş vermiş yerdəyişmələri (hadisələri) bu hadisələri xarakterizə edən şablonlarla (etalon təsvir formaları ilə) müqayisə etməklə onları videoaxın və ya videoarxiv verilənlərindən mümkün qədər tez aşkarlayıb tapmaqdır.

Bu gün videoanalitikaya Big Data texnologiyalarının bir istiqaməti – videoinformasiyanın analitik emalı vasitəsi kimi baxılır. Videonəzarət sistemləri isə onun tətbiq sahələrindən biridir. Videoanalitika son bir neçə ildə formalaşmaqda olan sahədir və ilkin mərhələdə əsasən aşağıdakı məsələlərin həlli məqsəd kimi qarşıya qoyulmuşdur:

- arxivləşdiriləcək videoinformasiyanın həcmini azaltmaq məqsədilə giriş axınının filtrasiyası, yəni nəzarət məqsədlərinə uyğun mühüm hadisələrin aşkarlanması və seçilib yazılması;
- operativ axtarış meyarları çoxluğunun müəyyənləşdirilməsi və metodlarının işlənməsi, məsələn, arxiv verilənlərinin müxtəlif meyarlara görə strukturlaşdırılması;
- nəzarətə götürülən hadisələr çoxluğunun xarakterik şablonlarının (etalonlarının) müəyyənləşdirilməsi və uyğun verifikasiya metodlarının işlənməsi.

Hal-hazırda videoinformatikada qeyd olunan məsələləri həll etmək üçün fərqli prinsiplərə əsaslanan üç sinif, nisbətən sadə hesab edilən analitik alqoritmlərdən istifadə edilir:

- təsvirlərin piksellərinin müqayisəsi. Bu sinif alqoritmlər eyni formatlı kadrlardakı təsvirlərin bir-biri ilə müqayisəsinə əsaslanır;
- şablonlarla müqayisə. Cari təsvir bazada yerləşdirilmiş xarakterik hərəkətləri (statik vəziyyəti və ya vəziyyətlər ardıcıllığını) əks etdirən etalon nümunələrlə (şablonlarla) müqayisə edilir. Müqayisə obyekti videoaxın və ya baza verilənləri ola bilər;
- obyekt davranışının öyrənilməsi və qiymətləndirilməsi. Lakin inkişaf tendensiyası yaxın illərdə daha mürəkkəb (funksionallıq nöqteyi-nəzərdən) və dəqiq (yüksək verifikasiya dərəcəli) sinif alqoritmlərin işlənəcəyini proqnozlaşdırmağa əsas verir.

Hal-hazırda videoanalitikanın videonəzarət sistemlərinə tətbiqi ilə həll olunan məsələlər sırasına mühafizə edilən əraziyə daxil olma, virtual xətti adlama, yanğınların aşkarlanması, avtomobillərin nişan nömrələrinin və sürücünün bir sıra yol hərəkəti qaydalarını pozması hallarının aşkarlanması, sahibsiz qalan əşyaların aşkarlanması, nəzarətə götürülmüş ərazi daxilində qarət və oğurluq hallarının, kütlə içində axtarılan şəxsin olmasının aşkarlanması və s. daxildir.

Müasir analitik analiz proqramlarının qiyməti kifayət qədər yüksək olduğu üçün kiçik və orta miqyaslı videonəzarət sistemlərində nəzarət və analiz funksiyaları avtonom şəkildə – operatorlar tərəfindən həyata keçirilir. AMEA İnformasiya Texnologiyaları İnstitutunun mühafizə videonəzarət sistemi də bu rejimdə işləyir. Yuxarıda qeyd olunmuş ikinci sistem isə avtonom rejimlə bərabər bir sıra sadə analitik analiz funksiyalarını da yerinə yetirir: nəqliyyat vasitəsinin hərəkət sürətinin müəyyənləşdirilməsi, dövlət nişan nömrəsinin, işıqforun qırmızı işığında hərəkət, qadağan olunmuş xətti adlama hallarının aşkarlanması və s. Təhlükəsizlik kəmərindən istifadə edilib-edilməməsinin, sürücünün maşına sahibliyinin və ya onu əvəz edənin idarəetmə hüququnun aşkarlanması və s. funksiyalarının icrası isə hələlik mümkün deyil.

NƏTİCƏ

Məqalədə cəmiyyətin inkişaf səviyyəsi, istehsal edilən informasiyanın həcmi və informasiya daşıyıcıları arasındakı səbəb-nəticə əlaqələri araşdırılır. Göstərilir ki, sivilizasiyanın, insan cəmiyyətlərinin, ölkələrin, xalqların generasiya etdiyi informasiyanın həcmi nə qədər çoxdursa inkişaf səviyyəsi də bir o qədər yüksəkdir. Tarixən müəyyən bir dövr üçün informasiya həcmi/inkişaf səviyyəsi nisbəti stabil qalsa da inkişaf dinamikasının müəyyən həddində yeni texnika və texnoloji vasitələrin ixtirası, yaradılması bu stabilliyi pozur: informasiya istehsalında sıçrayışlı artım yaranır, yeni və daha yüksək inkişaf dövrü başlayır. Bu dövrülük tarix boyu təkrarlanır. Son 30-40 ildə informasiyanın sıçrayışlı artım halını da informasiya inqilabı kimi xarakterizə edilir. Məqalədə informasiya daşıyıcılarının müsbət dinamikası da artan informasiyanın saxlanması zərurəti ilə əlaqələndirilir və bu dövr üçün zaman/həcm əlaqəsi cədvəli verilir. İnformasiyanın klassik emalı texnologiyalarından Big Data-ya keçid prosesi analiz edilmiş, bu keçidin son illər sıçrayışla artmaqda olan informasiya emalı problemlərinin həlli zərurətindən yaranması məlum olmuşdur. Qeyd edilir ki, bu problemlərin həlli klassik metodlardan fərqli prinsiplərə əsaslanan yeni yanaşma və metodların işlənməsini tələb edir. Big Data texnologiyalarının videoanalitika istiqaməti, onun videonəzarət sistemlərinə tətbiqi problemləri, müasir vəziyyəti və praktiki həll nümunələri verilir.

ƏDƏBİYYAT

[1] Моисеев Н.Н. Информационное общество: возмщжности и реальность // Информационное общество. М.: АСТ, 2003, с. 248-251.

[2] Негодаев И.А. Информационные революции в истории общества // Вестник ДГТУ, 2006, т.6. №2 (29), с. 144-153.

[3] Большие данные (Big Data). http://www.tadviser.ru/index.php/Статьи

[4] D. Leney. 3D data management: Controlling data volume, velocity and variety. Technical report META Group. 2001. http://blogs.Gartner.com

[5] Tomas Coughlin. 2013 Digital Storage for Media and Entertainment Report. http://www.tomcoughlin.com/Techpapers/M&E%20Storage%20Report%20Brochure,%202013,%20072413.pdf

[6] Worldwide Big Data Techology and services 2013-2017 Forecast. http://www.ide.com

[7] IP KVM с потоковым видео HD-качества. «Журнал сетевых решений/LAN». №3, 2014. http://www.osp.ru/lan/2014/03/13040186/

Böyük Verilənlərdən Faydalı İnformasiyanın Aşkarlanmasında Mövcud Yanaşmalar

Rəna Qasımova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
depart1@iit.ab.az

Xülasə — **Yüzlərlə terabayt və ekzabayt həcmində böyük verilənlərin saxlanması, real-vaxt ərzində analitik emalı, onlardan faydalı informasiyanın əldə edilməsi ciddi problemlərin yaranmasına səbəb olmuşdur. Məqalə böyük verilənlərin emalı problemlərinə həsr olunur və böyük verilənlərdən faydalı informasiyanın aşkarlanmsında mövcud yanaşmalar analiz olunur. Eyni zamanda büyük verilənlərin əsas xarakteristikaları və xüsusiyyətləri tədqiq olunur.**

Açar sözlər — verilənlər xəzinəsi, bulud, struktulaşdırılmamış informasiya, big data, NoSQL, MapReduce, Hadoop.

I. GİRİŞ

Son zamanlar böyük verilənlər (Big Data) ideyası aktuallıq kəsb etmiş, bu sahəyə olan maraq daha da artmışdır. Big Data termini böyük həcmli verilənlərlə işləyən bütün sferalarda, iqtisadiyyatda, bank fəaliyyətində, istehsalatda, marketinqdə, telekommunikasiyada, veb analitikada, tibbdə və s. sferalarda istifadə edilir [1-3]. Hazırda IBM, Oracle, Microsoft, SAS, SAP, HP, Teradata, EMC və s. kimi nəhəng İT informasiya texnologiyaları istehsalçıları tərəfindən böyük vüerilənlərin (BV) emalını təmin edən startaplar, mürəkkəb informasiya sistemləri yaradılır [4-8]. Bütün aparıcı sahə analitikləri rəqabət qabiliyyətini təmin etmək üçün BV ilə iş bacarığının nə qədər vacib olmasına diqqəti cəlb etməyə çalışırlar. Bu bir daha böyük həcm, sürət və mürəkkəblik kimi xarakteristikalarla təyin olunan BV sahəsindəki problemlərin nə qədər aktual olmasından xəbər verir. BV-nin saxlanması, idarə edilməsi, analitik emalı, təhlükəsizliyi və s. məsələlərinin həlli yeni texnologiyalar, idarəetmə metodları və daha mükəmməl intellektual analiz vasitələrini tələb edir.

Bununla belə, BV problemi hələ ilkin araşdırmalar səviyyəsindədir, yəni bu sahə hələ də tam olaraq təhlil olunmayıb. Aparılan bütün tədqiqatlar “Big Data” anlayışını, mahiyyətini, onun müxtəlif xarakteristikalarını təsnifləndirməyi, BV-nin mənbələrini, bu texnologiyanın imkanlarını, problemlərini, təhlükəsizlik məsələlərini hərtəfli tədqiq etməyə imkan verir. Eyni zamanda tədqiqatlar göstərir ki, BV-in emalı və analizi üçün mükəmməl analitik texnologiyalar və alətlər tələb olunur [9-11].

II. BİG DATA XÜSUSİYYƏTLƏRİ

Bu gün Big Data İT sahəsindəki nəhəng vendorların və beynəlxalq analitik agentlərinin istifadə etdikləri anlayışdır. Qeyd edək ki, Big Data texnologiyasını və bu texnologiyanı digərlərindən fərqləndirən xüsusiyyətlərini araşdırarkən müxtəlif fikirlərə rast gəlmək olur. Mütəxəssislərin fikrincə Big Data texnologiyalar çoxluğudur və aşağıdakı xüsusiyyətlərə malikdir [12, 15]:

– ***Çox böyük həcm.*** Hazırda BV-nin miqyası terabaytlardan zetabaytlara qədər həcm ilə xarakterizə olunur. Bu gün müəssisə səviyyəsində verilənlərin belə həcmi artıq heç kəsi təəccübləndirmir.

– ***Çoxlu sayda informasiya mənbələri.*** Məlumdur ki, biznes analitika (Business intelligence, BI) emal olunmayan informasiyanın anlamlı, rahat formaya çevirmək üçün lazım olan metod və alətlərdir. Ənənəvi BI-də bir neçə informasiya mənbələri var, Big Data isə onlarla, minlərlə xarici mənbələrə malikdir.

– ***Strukturlaşdırılmamış informasiya.*** Ənənəvi verilənlər xəzinəsi (VX) relyasion baza əsasında qurulmuşdur və strukturlaşdırılmış informasiyanın axtarışı üçün nəzərdə tutulmuş alətdir. Adi relyasion bazalarının emal edə bilmədiyi yüzlərlə terabayt və ekzabayt həcmində müxtəlif tip strukturlaşdırılmamış informasiyanın toplanması və saxlanması, axtarışı, təhlükəsizliyi, analizi kimi məsələlərin həlli üçün yeni alətin yaradılmasına ehtiyac yaranır. Big Data mənbələrinin əsas hissəsini də strukturlaşdırılmamış və qismən strukturlaşdırılmış verilənlər təşkil edir. Bu halda Big Data yanaşması prinsipial olaraq daha mükəmməl həll olan şablonlar əsasında axtarışı təklif edir, bu isə öz növbəsində relyasion verilənlər bazasından fərqli olan informasiyanın saxlanması strukturunu nəzərdə tutur.

– ***Dinamik dəyişən informasiya***. Böyük verilənlər çox böyük informasiya massivlərinin toplanması ilə xarakterizə olunur ki, bu da hazırda istənilən təşkilatın qarşılaşdığı problemdir. Böyük həcm problemi ilk növbədə saxlama problemi yaradır ki, bu da genişmiqyaslı saxlama və paylanmış emal tələb edir. Eyni zamanda bütün verilənləri, hətta filtrlənmiş verilənləri də saxlamaq getdikcə daha baha başa gəlir. Məlumdur ki, saxlama texnologiyaları bahadır və onların uçuzlaşması yeni verilənlər mənbəyinin yaranmasına nisbətən daha yavaş baş verir. Bununla bağlı təşkilat səviyyəsində bu və ya digər verilənlərin hansı vaxta qədər saxlanması dəqiq müəyyən olunmalıdır. Məsələn, bəzi verilənlər təşkilatda uzun illər ərzində tələb oluna bilər, digərləri isə artıq bir neçə saatdan sonra, analitiklər onlardan lazım olanı götürdükdən sonra faydasız ola bilər.

Son zamanlar isə verilənlərin emalı və yadda saxlanılmasına xidmət edən "grid" və "cloud computing" texnologiyalarının tətbiqi saxlama sahəsindəki problemləri demək olar ki, aradan qaldıra bilmişdir. Beləliklə, Big Data BI-nin ənənəvi alətləri ilə həll olunan oxşar məsələləri həcm, mənbə, struktur və paylanma mərhələsində daha geniş kontekstdə həll edir. Nəticədə Big Data və BI texnoloji səviyyədə bir-birindən ciddi fərqlənirlər (cədvəl 1.) [16-19].

CƏDVƏL 1. Big Data və biznes analitika arasındakı fərqli xüsusiyyətlər

Ənənəvi BI	Big Data
Vahid korporativ VX	Xəzinə paylanmış fayl sistemində yerləşə bilər
Verilənlərin formatı emal funksiyalarının tələbinə uyğunlaşdırılır	Emal funksiyaları müxtəlif verilənlərin formatına uyğunlaşdırılır
Verilənlərin formatı strukturlaşdırılmışdır	Həm strukturlaşdırılmış, həm də strukturlaşdırılmamış informasiya ilə işləmək nəzərdə tutulur
Tarixi informasiya	Ən yeni verilənlərlə iş
Verilənlərin əldə edilməsi, ötürülməsi və emalı əsasən ardıcıl, dəqiq müəyyən olunmuş prosedurlara uyğun baş verir	Verilənlərin kütləvi paralel emalı tətbiq olunur (Massively Parallel Processing, MPP)

III. BİG DATA EMALININ TEXNOLOJİ MƏRHƏLƏLƏRİ

Hazırda aparılan elm-tədqiqat işlərinin əksəriyyəti BV-nin texnoloji problemlərinin həllinə yönəlmişdir. Yüzlərlə terabayt və ekzabayt həcmində BV-nin saxlanması, real-vaxt ərzində analitik emalı, onlardan faydalı informasiyanın əldə edilməsi ciddi problemlərin yaranmasına səbəb olmuşdur. Big Data-nın əsas mənbələrini sensor və sosial şəbəkələr, müxtəlif sahələr haqqında məlumat ötürücüləri, bank əməliyyatları, coğrafi informasiya sistemləri (GİS), Qlobal Mövqetəyinetmə Sisteminin siqnalları (GPS), elmi eksperimentlər, elektron poçt, smartfonlar vasitəsilə alınmış rəqəmsal foto və videolar, böyük şirkətlərin bazarları, böyük satış mərkəzləri, domen adları serverləri (Domain Name Server, DNS) və s. təşkil edir. Bu mənbələr tərəfindən yaradılan böyük həcmli verilənlərin məhsuldar istifadəsi üçün uyğun texnologiyalar lazımdır. Bunlar verilənlərin toplanması, saxlanması və analitik emalı sistemlərindən ibarətdir. Onun müxtəlif hissələrində aşağıdakı texnologiya qruplarından istifadə edilə bilər [20]:

- analitik alqoritmlər;
- paralel proqramlaşdırma metodları;
- bulud hesablama resursları;
- fərdi kompüterdən strateji təyinatlı superkompüterlərədək hesablama sistemləri;
- saxlama sistemləri;
- səbəkələr;
- mürəkkəb teleskoplardan və tomoqraflardan sadə radiotezlikli identifikasıya (Radio Frequency Identification, RFID) texnologiyarınadək müxtəlif növ daxiletmə qurğuları və s.

Sadalanan siyahıda səviyyələr yüksəldikcə verilənlərin biliklərə yaxınlığı da artır. Analitika verilənlərin biliyə çevrilməsi prosesini başa çatdırır. Siyahıda aşağı səviyyələr yoxlanmış elmi və mühəndis həllərinə əsaslanır, yuxarı səviyyələr isə yeni olduqlarına görə zəif inkişaf etmişlər və kifayət qədər elmi təminata malik deyillər. Mütəxəssislər bilavasitə İT-dən Big Data üçün texnologiyalar kateqoriyasına verilənlərin kütləvi paralel emalı platformalarında analitik sistemləri, verilənlərin bulud servisini, Hadoop və MapReduce texnologiyalarını və NoSQL tipli paylanmış verilənlər bazasının idarəetmə sistemlərini (VBİS) və s. daxil edirlər. Apache Software Fondation-un layihəsi olan Hadoop daha geniş yayılmış texnologiyadır, paylanmış hesablama mühitində böyük həcmli verilənlərin (petabayt miqyasında) emalı və analizi üçün əsas platformadır, MapReduce modelinin açıq kodlu (open access) sistemidir. Hadoop iki əsas komponentdən təşkil olunmuşdur: Hadoop MapReduce və Hadoop Distributed File System (HDFS). Burada MapReduce paralel hesablamalara, HDFS paylanmış fayl sistemi isə verilənlərin idarə edilməsinə cavab verir [21-25].

Böyük verilənlərin növbəti dalğasını insanın kompüterlə qarşılıqlı əlaqəsini təmin edən yeni qurğular yarada bilər. Onlara elektron kağız, taktil əks əlaqə texnologiyaları (haptics), müxtəlif videoşlemlər (video visor), həmçinin xüsusi yaddaşlı məhsul və əşyalar, Vikipediyaya oxşar açıq kontekstlə işləmək üçün sistemlər və s. aiddir. Məlumdur ki, geniş mənada texnologiya, əmtəə və xidmətlərin istehsalı üçün zəruri bilik kapitalı, tətbiqi mənada isə məhsulun hazırlanması prosesində maddənin, enerjinin, informasiyanın çevrilməsi, materialların emalı və yenidən emalı, hazır məhsulların yığılması, keyfiyyət və idarəetməyə nəzarət üçün lazım olan metodlardır. Texnoloji ardıcıllıqdakı maşın və mexanizmlər ilkin xammalı istehlaka hazır olan məhsula çevirir. Hazırki inqilab amorf, qeyri-müəyyən "informasiya texnologiyaları"nı rədd edərək daha aydın "verilənlər texnologiyaları"nı təsdiq edir. Texnoloji ardıcıllığın girişində xam verilənlər, çıxışda isə insan tərəfindən istifadəyə hazır olan verilənlər olmalıdır. Bu aspektdən müasir texnologiyalar verilənlərdəki informasiyanın yeni biliklərə çevrilməsinə və ya biliklərin əldə edilməsinə imkan verməlidirlər [26, 27].

Verilənlərlə işləyən texnologiyaların əsas fərqləndirici xüsusiyyətlərindən biri ondadır ki, onlar insan üçün işləyirlər. Avtomatik quraşdırılmış sistemlər istisna olmaqla qalan kompüter sistemləri son nəticədə insanın istifadə etdiyi verilənlərin hazırlanmasına xidmət edir, yalnız bu sistemlər verilənləri informasiyaya və sonra biliyə çevirə bilər. Bu cür texnologiyaların paradoksallığı ondadır ki, texnoloji ardıcıllığın girişində verilənlərin həcmi daim artırsa, buradan da böyük verilənlərin problemi yaranır. Giriş və çıxışdakı verilənlərin həcmi arasındakı disproporsiya mahiyyətcə verilənlərlə işləyən texnologiyaların inkişafında əsas istiqaməti müəyyən edir: daxil edilən verilənləri itirməməklə hansısa bir yolla giriş axınını cilovlamaq, sonra bütün verilənlərdən daha əhəmiyyətlilərini seçmək və insanın qavraması üçün asan şəkildə təqdim etmək lazımdır. Hazırkı iqtisadi böhran exafloodla (exaflood, petabaytdan sonra gələn, 1018 bayta bərabər olan verilənlərin ölçü vahidi exabyte və flood, yəni, daşqın sözlərindən yaranmışdır) bağlı sahələrdə inkişafı dayandırmamış, lazımi verilənləri o qədər də mühüm

olmayanlardan ayırmaq üçün filtr rolunu oynayan mürəkkəb hadisələrin emalı texnologiyalarının (Complex Event Processing, CEP) inkişafını sürətləndirmişdir. Bu qrup texnologiyalar BV-lərlə bağlı əsas məsələlərdən birini həll edir, yəni, xam verilənlərdən yeni, gizli biliklərin və faydalı məlumatların əldə olunmasına keçidi asanlaşdırır [28].

Bu gün böyük həcmli verilənlərdən faydalı informasiyanın aşkarlanması üçün ən qabaqcıl metodlar təmin edilir, belə məsələləri effektiv həll etməyə imkan verən texnoloji alətlər yaradılır. Mütəxəssislər tətbiq sahəsindən asılı olmayaraq böyük verilənlər emalının texnoloji ardıcıllığını yeddi əsas mərhələyə bölürlər [29-31]:

– ***Verilənlərin toplanması.*** Xam verilənlər xəzinələrdən, qurğuların vericilərindən, şəbəkə mənbələrindən toplana bilər. Bu mərhələ mühəndis nöqteyi-nəzərindən daha ənənəvidir, lakin burada da verilənlərin tipini ayırmaq lazımdır, məsələn, mətn verilənlərini ədədi verilənlərlə səhv salmaq olmaz.

– ***Verilənlərin sintaksis və qramatik təhlili***. Burada verilənlərin strukturlaşması, onların kateqoriyalar üzrə paylanması həyata keçirilir, təhlil isə bir neçə səviyyədə yerinə yetirilir. Aşağı səviyyəyə fiziki siqnalların emalı, sıxılmış faylların açılması və deşifrələnməsi daxildir. Bit səviyyəsində mətn, media və digər faylların ayrılması, mətn səviyyəsində qramatik və struktur təhlili aparılır.

– ***Filtrasiya.*** Burada verilənlərin analizi metodlarına, məsələn, mürəkkəb hadisələrin emalı texnologiyaları vasitələrinə müraciət etmədən giriş axınını azaldaraq yalnız faydalı verilənləri saxlamaq lazımdır.

– ***Verilənlərin əldə edilməsi.*** Verilənlərin əldə edilməsi təsvir və verilənləri ayırmağa imkan verən statistik və digər metodlardan ibarətdir, verilənləri uyğun riyazi kontekstə həll edir.

– ***Təqdimat.*** Verilənlər üçün daha yaxşı təqdimat formasını müəyyən etmək tələb olunur (diaqramlar, siyahılar, ağaclar və s.). Verilənlərin vizuallaşma üsulu sadə cədvəl və qrafiklərdən tutmuş mürəkkəb iki və üç ölçülü təsvirlərə qədər dəyişir.

– ***Təqdimatın təkmilləşdirilməsi.*** Burada verilənlərin təqdimatı formalarının redaktəsi aparılır.

– ***Qarşılıqlı əlaqə.*** Verilənlərlə fəal işləmək üçün verilənlərlə manipulyasiya fəndləri və daha əyani təqdimatı təmin edən üsullar işlənir.

Verilənlərin toplanması və onların təhlili ənənəvi texnologiyalar vasitəsilə reallaşır, filtrasiya və əldə edilmə verilənlər haqqında yeni elmin predmetidir, verilənlərin təqdimatı və dəqiqləşdirilməsi qrafiki dizayn sahəsinə daxildir.

NƏTİCƏ

Big Data sadə düşüncə deyil, gələn texniki inqilabın simvoludur. BV-lərlə analitik iş zəruriliyi İT-sənayesini tamam dəyişəcək yeni proqram və aparat platformalarının yaranmasını stimullaşdırır. Artıq bu gün böyük həcmli verilənlərin analizi üçün ən qabaqcıl metodlar təmin edilir: süni neyron şəbəkələri – bioloji neyron şəbəkələrinin təşkili və funksiyası prinsipi üzrə qurulmuş modellər, prediktiv analitika, statistika, Natural Language Processing metodları və s. Həmçinin kraudsorsinq, A/B testləşmə, sentiment analiz və ekspertləri cəlbedən metodlardan istifadə edilir.

Beləliklə, dünyada gedən proseslərə və aparıcı ölkələrin təcrübəsinə nəzər saldıqda aydın olur ki, müxtəlif mənbələrdən avtomatik və fasiləsiz olaraq generasiya olunan verilənlərin real-vaxt ərzində analitik emalı və analizində fərqli modellər, konseptual yanaşmalar təklif olunur. Bu mühüm faktı nəzərə alaraq, beynəlxalq təcrübədə BV-nin analitikası istiqamətində aparılan tədqiqatların araşdırılması olduqca zəruridir. Aydındır ki, çox böyük həcmdə rəqəmsal verilənlərin toplanması cəmiyyətin bütün sferalarının qarşılaşdığı problem olduğu üçün, onunla bağlı elmi-tədqiqat işləri də bir çox elm sahəsi üçün aktualdır. Çünki, bu xam, srtukturlaşdırılmamaış verilənlər cəmiyyətin bütün sahələrini kökündən dəyişə biləcək təsirə malik bilik mənbəyidir. Bu baxımdan, Big Data texnologiyasının bütün aspektlərinin elmi-nəzəri əsaslarının işlənməsi böyük əhəmiyyət kəsb edir. Baxılan problemin araşdırılması, elmi-nəzəri tədqiqatların təhlili əsasında alınan nəticə göstərir ki, BV-dən yeni biliklərin və faydalı məlumatların əldə olunması verilənlərin idarə edilməsində, onlardan dəyər yaradılmasında, mərkəzləşdirmə, interaktivlik, onlayn rejim və digər spesifik xüsusiyyətləri sayəsində ictimai sferanın yeni, unikal modeli formalaşır.

ƏDƏBİYYAT

[1] Clifford L. Big data: How do your data grow?, Nature, 2008, vol. 455, p. 28-29.

[2] The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe

[3] Wei Fan, Albert Bifet. Mining big data: current status, and forecast to the future, ACM SIGKDD Explorations Newsletter, 2012, vol. 14, No. 2, p. 1-5.

[4] InfoSphere Platform: Big Data Analytics, 2013, http://www-01.ibm.com/software/

[5] Oracle and Big Data: Big Data for the Enterprise, 2013, http://www.oracle.com/

[6] Big Data, 2013, http://www.microsoft.com/.

[7] Big Data - What Is It? 2013, http://www.sas.com/big-data/

[8] SAP HANA integrates predictive analytics, text and big data in a single package, 2013, http://www54.sap.com/

[9] Big Data Solutions, 2013, http://www8.hp.com/

[10] Madden S. From Databases to Big Data, IEEE Internet Computing, 2012, vol.16, issue 3, p. 4-6.

[11] What is big data? - Bringing big data to the enterprise, 2013. http://www-01.ibm.com

[12] Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001. http://blogs.gartner.com

[13] Birke R., Björkqvist M., Chen L. Y., Smirni E., Engbersen T. (Big)data in a virtualized world: volume, velocity, and variety in cloud datacenters, Proceedings of the 12th USENIX conference on File and Storage Technologies (FAST'14), USENIX Association Berkeley, CA, USA, 2014, p. 177-189.

[14] Richard Price. Volume, velocity and variety: key challenges for mining large volumes of multimedia information, Proceedings of the 7th Australasian Data Mining Conference (AusDM '08), Australia, 2008, vol. 87, p. 17-17.

[15] Chiang R.H.L., Goes P., Stohr E.A. Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline, Journal ACM Transactions on

Management Information Systems, 2012, vol. 3, No. 3, Article 12 (p. 12:1-12:13).

[16] Chen H., Chiang R.H. L., Storey V.C. Business intelligence and analytics: from big data to big impact, Journal MIS Quarterly, 2012, vol. 36, No. 4, p. 1165-1188.

[17] Omar El-Gayar, Prem Timsina. Opportunities for Business Intelligence and Big Data Analytics in Evidence Based Medicine, HICSS '14 Proceedings of the 2014 47th Hawaii International Conference on System Sciences(HICSS '14), USA, 2014, p. 749-757.

[18] Statchuk C., Iles M., Thomas F. Big data and analytics, Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research (CASCON '13), USA, 2013, p. 341-343.

[19] Foster Y., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations, Intern. J. of High Performance Computing Applications, 2001, vol. 15, No. 3, p. 200-222,

[20] Черняк Леонид. Большие Данные – новая теория и практика, Открытые системы, 2011, № 10, с. 18-25.

[21] Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters, Communications of the ACM, 2008, vol. 5, No.1, p. 107-113.

[22] Lee K-H., Lee Y-J., Choi H., Chung Y.D., Moon B. Parallel data processing with MapReduce: a survey, ACM SIGMOD Record, 2011, vol. 40, No. 4, p.11-20.

[23] Brunozzi Simone. Big Data and NoSQL with Amazon DynamoDB, Proceedings of the 2012 workshop on Management of big data systems (MBDS '12), 2012, USA, p. 41-42.

[24] Weiyi Shang, Zhen Ming Jiang, Hadi Hemmati, Bram Adams, Ahmed E. Hassan, Patrick Martin. Assisting developers of big data analytics applications when deploying on hadoop clouds, Proceedings of the 2013 International Conference on Software Engineering (ICSE '13), NJ, USA, 2013, p. 402-411.

[25] Чак Лэм. Hadoop в действии, Издательство: ДМК Пресс 2012, 424 с.

[26] Черняк Леонид. Вычисления с акцентом на данные, Открытые системы, 2008, № 8, с. 36-39.

[27] Wu X., Zhu X., Wu G., Ding W. Data Mining with Big Data, Journal IEEE Transactions on Knowledge and Data Engineering, USA, 2014, vol. 26, No.1, p. 97-107.

[28] Wang Y.H., Cao K., Zhang X.M. Complex event processing over distributed probabilistic event streams, Computers & Mathematics with Applications, 2013, vol. 66, No. 10, p. 1808-182131.

[29] Черняк Леонид. Смутное время СУБД, Открытые системы, 2012, № 2, с. 16-21.

[30] Черняк Леонид. Что делать с хаосом данных?, Открытые системы, 2013, № 9, с. 16-20.

[31] Дубова Наталья. Большие Данные крупным планом, Открытые системы, 2011, № 10, с. 30-33.

Bibliometric Analysis of Big Data Research

Ramiz Aliguliyev[1], Nigar Ismayilova[2]
[1,2] ANAS Institute of Information Technology, Baku, Azerbaijan
[1]*r.aliguliyev@gmail.com*, [2]*nigar@iit.ab.az*

***Abstract* — The paper analyzes research performance on an area of big data during 2001-2015 years, using data from Scopus. Also was demonstrated geographical distribution of big data research, and predicted future of this area by means of received information.**

Keywords — big data; bibliometric analysis; research performance

I. INTRODUCTION

Big Data is related to technologies for collecting, processing, analyzing and extracting useful knowledge from very large volumes of structured and unstructured data generated by different sources at high speed [1]. The term Big Data is used almost anywhere in modern times; from news articles to professional journals, from tweets to YouTube videos and blog discussions. The term coined by Roger Magoulas from O'Reilly media in 2005 [2], refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity [3]. This paper is aimed to provide a summary of research activity on big data and characterize its most important aspects.

II. RESULTS OF BIBLIOMETRIC ANALYSIS

Scopus is the largest bibliometric database covering peer-reviewed literature: scientific journals, books and conference proceedings [4]. Scopus was used as a main source in our analysis, the results of which are presented below.

A. Publication output

Big data research has begun in 2001with one published document in Scopus, and 2014 was the most productive year with 3472 documents (45 % of all publications from 2001 to 2015) (figure 1).

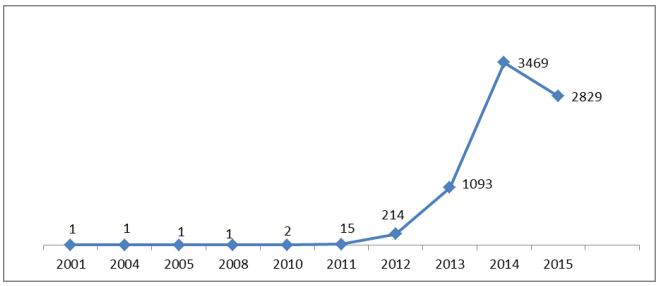


Figure 1. Number of publications in big data.

Exponential and polynomial dependence of publications' number on publication year are C=0.0567exp(1.1011Y) and C= 83.76Y2 -582.95Y+744.02, where Y – is the year and C- is the cumulative number of published documents. According to these equations we can predict that, the number of publications in big data research will be approximately 4467 in 2016.

B. Type of document

The queries result consists of 7633 documents; most of them are conference papers - 4887 (64 %) and journal articles - 2190 documents (28,7 %) . Reviews (204 publications – 2.7%), notes (15 publications – 0.2%), editorial materials (32 publications – 0,4 %) and book chapters (33 documents - 0,5 %) showed much-lesser significance than articles and conference materials. Totally, 7685 articles were published in 98 sources listed in the Scopus. Most of the articles published in journals edited by IEEE, Springer, Elsevier and reported in International Conferences as IEEE Big Data, Data Engineering, Advanced Cloud and Big Data, Symposium on Computational Intelligence in Big Data, Big data Congress, etc. The most productive authors in big data are from Australia and China (Table 1).

TABLE I. TOP TEN AUTHORS ON BIG DATA

N	*Authors*	*Affiliations*	*h-index*	*Number of publications in big data*
1	Ranjan Rajiv	Commonwealth Scientific and Industrial Research Organization, Melbourne, Australia	14	22
2	Wang Lizhe	Chinese Academy of Sciences, Institute of Remote Sensing and Digital Earth, Beijing, China	19	15
3	Cuzzocrea Alfredo	Universita della Calabria, Cosenza, Italy	18	13
4	Liu Chang	Tongji University, Department of Chemistry, Shanghai, China	51	14
5	Chen Jinjun	University of Technology Sydney, Faculty of Engineering and IT, Sydney, Australia	20	14
6	Wang Wei	Xi'an Jiaotong University, School of Electronic and Information Engineering, Xi'an, China	85	15
7	Zhang Xuyun	NICTA, Machine Learning Research Group, Canberra, Australia	6	15
8	Rong Chunming	Universitetet i Stavanger, Centre of Innovation Technology, Stavanger, Norway	10	13

N	*Authors*	*Affiliations*	*h-index*	*Number of publications in big data*
9	Puliafito Antonio	Universita degli Studi di Messina, Department of Civil Engineering, Messina, Italy	21	12
10	Xhafa Fatos	Universitat Politecnica de Catalunya, Department of Languages and Informatics Systems, Barcelona, Spain	21	12

C. Distribution in subject areas

All the documents were published in 26 scientific areas. Most of them (76, 7 %) belongs to Computer Sciences, 23.1 % Engineering and 14.8 % Mathematics. The publications was in Social sciences, Business, Medicine, Physics and other disciplines showed lesser significance (Figure 2).

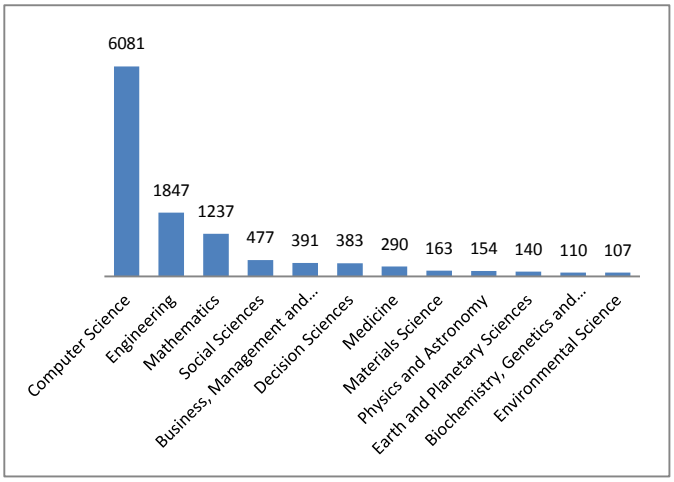


Figure 2. Distribution of publications by subject areas

D. Geographical distribution of publications

The analyzed documents were from 100 countries; researchers from USA have published 2250 documents (29 %), researchers from China 1881 documents (24,4 %). Asian countries as India and South Korea are in top ten of the most productive countries in big data research. The publications are in 13 languages, the documents published in English is 99,7 %, Chinese is 3,3, % and 1% of all publications are written in German, Spanish, French, Japanese, Turkish, Russian, Korean, Poland, Hungarian, Italian and Serbian. The notion of big data was first used by German researcher in 2001, and afterwards by the US and Chinese authors in 2004 and 2005 accordingly. In 2010 were fixed 2 documents from UAE and South Korea. During 2011-2012 years USA was the leader in 'big data' research, in 2013-2014 China have gone after the USA with the insignificant difference and in 2015 have passed it. They are followed by Italy in 2011, UK in 2012 and 2013, Germany in 2014 and India in 2015. In 2015 the Asian countries become more active than the developed European countries (Figure 3, 4, 5, 6).

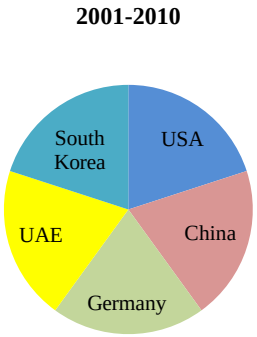


Figure 3. Distribution of publications by countries in 2001-2010 years

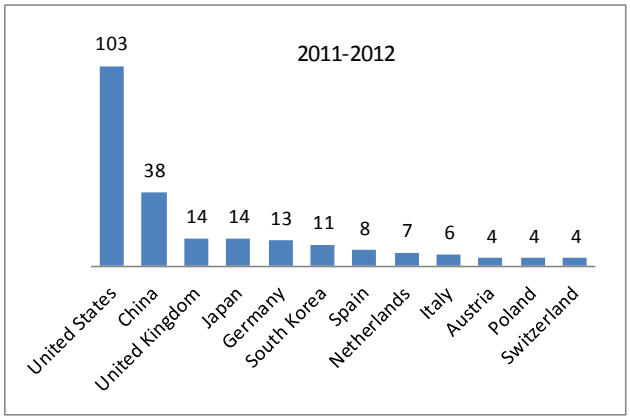


Figure 4. Distribution of publications by countries in 2011- 2012

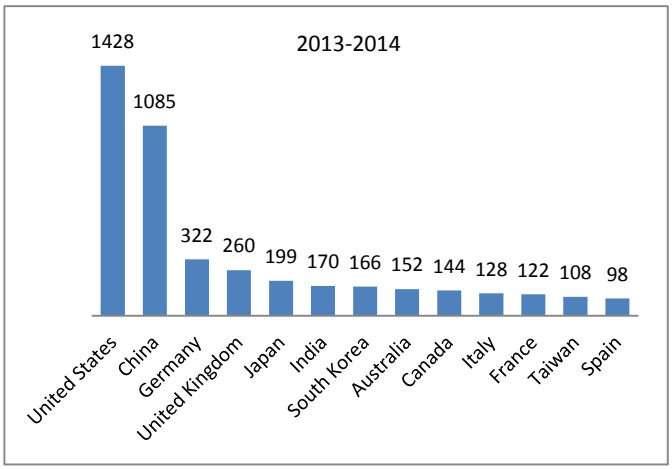


Figure 5. Distribution of publications by countries in 2013-2014

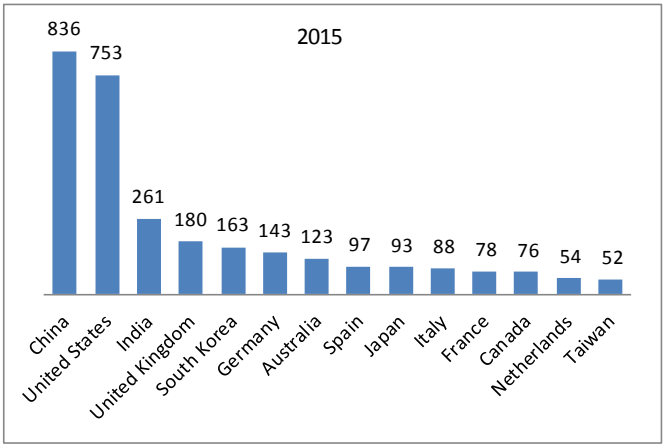


Figure 6. Distribution of publications by countries in 2015

We can conclude that, big data research gets its beginning from Germany, has grown in developed countries as the USA and UK and is prospering in Asian countries as China and India.

E. Topic longevity and the most cited article

Using the citation data we can calculate topic longevity of big data [5]:

$$a = i\sqrt{\frac{k}{k+l}}, \quad (1)$$

where n depends on three parameters, i - the number of years (in this case 6), k - the number of cited papers in topic t published i or more years before the current year, and l - the number of cited papers in topic t published less than i years previously. These parameters determine the aging rate a. The estimated half-life is then

$$\text{Longevity(t)} = -\frac{\log 2}{\log a}. \quad (2)$$

The cumulative number of cited documents during 2008-2014 and 2010 – 2014 years is equal to 1089, according to equation 2 topic longevity is approximately 0.28 [6]. Small longevity score of big data topic indicates that, publications in this topic are relatively recent. The most cited paper by Chen H. et al. was published in 2012 in the journal MIS Quarterly: Management Information Systems named as "Business intelligence and analytics: From big data to big impact" was cited 256 times during 2012-2015 years (Figure 7).

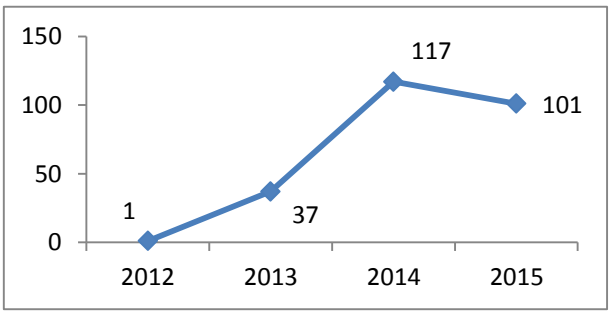


Figure 7. The history of the most cited article

CONCLUSION

Bibliometric analysis carried out in this paper, gives reason for following discourses: 'big data' topic now is progressing and it is directed from Europe and the USA toward Asian countries. Asian scientists have begun actively research and provide their contributions to big data research.

REFERENCES

[1] R. Alguliyev, Y. Imamverdiyev "Big Data: Big Promises For Information Security 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)", 2014, pp. 1-4.

[2] http://radar.oreilly.com/2010/01/roger-magoulas-on-big-data.html

[3] G. Halevi, H. Moed "The Evolution of Big Data as a Research and Scientific Topic", Research Trends Issue, 2012. pp. 03-04

[4] https://www.elsevier.com/solutions/scopus

[5] R. Əliquliyev, N. İsmayılova "Bibliometriya: müasir vəziyyəti, problemləri və inkişaf perspektivləri," Bakı, 2015, 78 səh.

[6] S. Mann, D. Mimno, A. McCallum " Bibliometric Impact Measures Leveraging Topic Analysis", Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 2006, pp. 1-10.

Plan Məlumatları Əsasında Virtual Hava Məkanında Proqnozlaşdırılan Münaqişə Vəziyyətinin İmitasiya Modeli

Nadir Ağayev[1,2], Ağası Məlikov[3], Hüseyn Babayev[4], Nəsrəddin İsgəndərov[5]
[1,3]Milli Aviasiya Akademiyası, Bakı, Azərbaycan
[2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[4,5]Azərbaycan Hava Yolları Qapalı Səhmdar Cəmiyyəti, Bakı, Azərbaycan
[5]*nesiazeri@yahoo.com*

Xülasə **– Məqalədə plan məlumatları əsasında hava gəmiləri arasında münaqişələrin aşkarlanması zamanı hava məkanındakı vəziyyətin əks etdirilməsi üçün üsul təklif olunur.**

Açar sözlər – hava gəmisi, münaqişə vəziyyəti, uçuş planı.

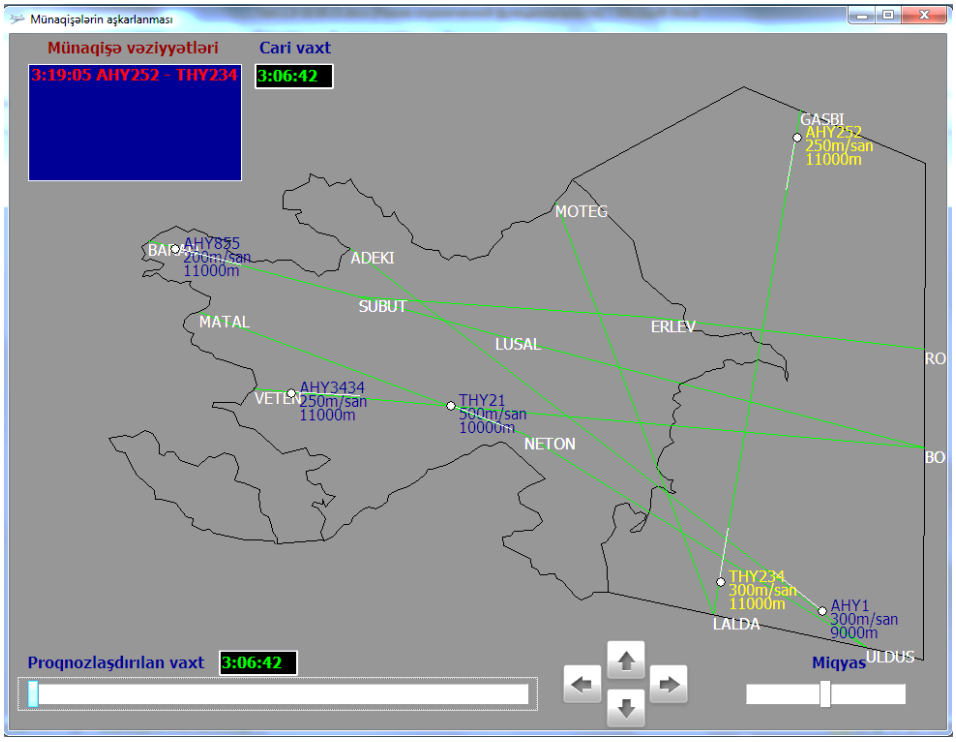
I. GİRİŞ

Havadakı hərəkətin intensivliyinin artması şəraitində hava məkanının buraxma qabiliyyətini artırmaq və HG-nin (hava gəmisi) hərəkətinin təhlükəsizliyini maksimum təmin etmək üçün havadakı hərəkətin müasir avtomatlaşdırılmış idarəetmə sistemləri (HHAİS) radar məlumatını emal edir, HG-nin koordinatlarını göstərir, uçuş planlarını emal və idarə edir, uçuşlar zamanı münaqişələri xəbər verir və dispetçer sorğularını yerinə yetirir [1,2]. HHAİS-in tərkibində olan münaqişələrin qabaqcadan xəbərdarlıq funksiyası əvvəlki sistemlərdə yalnız radar məlumatından istifadə edərək qısa müddət üçün (STCA) [3,4] nəzərdə tutulsa da, hazırki sistemlərdə həm də plan məlumatları əsasında HG arasında orta müddətli münaqişələri (MTCD) [5,6] aşkarlama funksiyasından istifadə olunur. Lakin araşdırmalar [7,8] göstərir ki, bu funksiyanın həm dəqiqlik baxımından, həm də münaqişənin baş verə biləcəyi zaman anında hava məkanındakı vəziyyəti əks etdirmə baxımından çatışmazlıqları mövcuddur. Buna görə də havadakı hərəkətin idarə olunmasının effektivliyini təmin etmək məqsədilə hərəkətin idarə edilməsi sistemlərinin təkmilləşdirilməsi üçün SESAR [9] proqramında bir çox yeniliklərin tətbiqi və bəzi funksiyaların yenidən işlənməsi planlaşdırılır. HG koordinatlarının hesablanması dəqiqliyinin artırılması üçün [10]-da Vincenty üsulunun hesablamalara tətbiqi, [11]-də isə bu üsuldan istifadə edərək uçuş planları əsasında HG arasında münaqişələrin aşkarlanması alqoritmi verilmişdir.

II. TƏKLİF OLUNAN MODEL

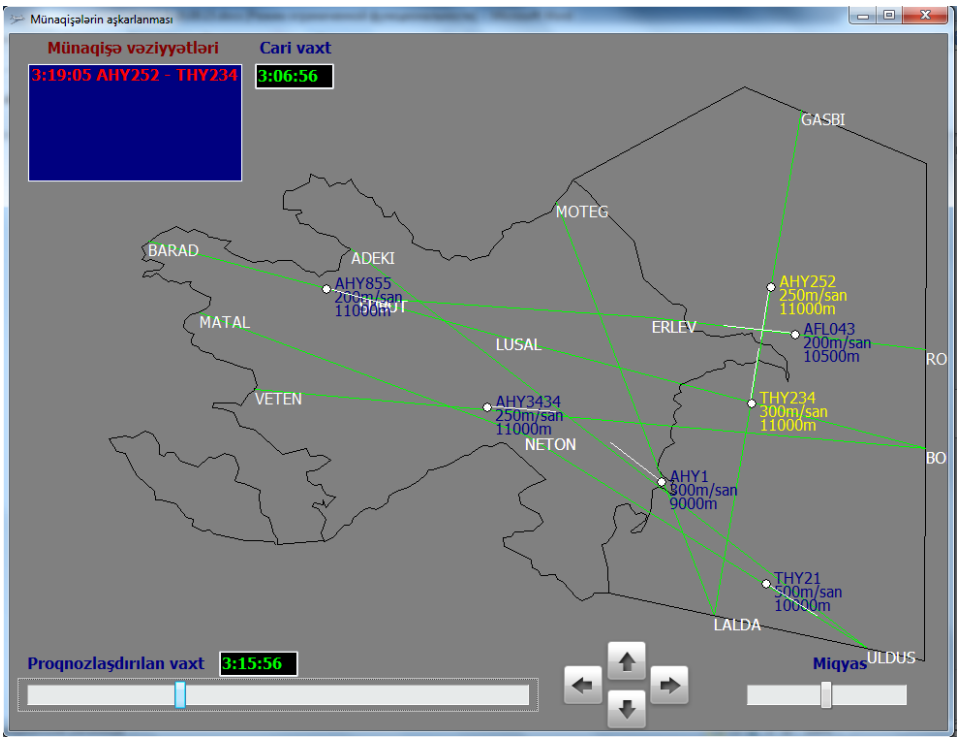
Hava məkanında plan məlumatları əsasında münaqişələrin aşkarlanması zamanı vəziyyətin əks olunması üçün təklif etdiyimiz funksiyanın plan məlumatlarını idarə edən dispetçerin monitorunda ayrıca bir pəncərədə verilməsi nəzərdə tutulur. Şəkil 1-də verilmiş proqram pəncərəsi Delphi [12] proqramlaşdırma dilinin imkanlarından istifadə olunaraq hazırlanmışdır. HG-nin hərəkəti və HG arasında münaqişələrin aşkarlanması [11]-də təklif etdiyimiz alqoritm əsasında həyata keçirilir. Proqram eyni zamanda verilənlər bazası faylının tərkibinə daxil olan “xəritə”, “nöqtələr”, “hava yolları” və “uçuş planları” cədvəllərindən istifadə edir. “Nöqtələr” cədvəlində nöqtələrin coğrafi koordinatları verilir. “Hava yolları” cədvəlində hansı nöqtələr arasında hava xətlərinin olması göstərilir. “uçuş planları” cədvəlində isə uçuş planlarının idarə olunması sistemindən qəbul olunmuş uçuş planları saxlanılır. Əgər cari vaxta uyğun uçuş planı varsa həmin HG-nin hərəkəti ekranda əks olunur.

Proqnozlaşdırılan vaxt ilkin halda cari vaxta bərabər götürülür. Biz proqnozlaşdırılan vaxtı 30 dəq-dək irəli çəkərək HG-nin harada olacağı haqda vizual məlumat ala bilərik. Eyni zamanda proqrama miqyas, xəritənin sağ, sol, yuxarı, aşağı hərəkət etdirilməsi funksiyaları əlavə edilmişdir. Münaqişə vəziyyətləri pəncərəsində 30 dəqiqə ərzində münaqişə vəziyyətində ola biləcək HG-nin siyahısı əks olunur. Burada HG-nin adları, münaqişə baş verəcək zaman anı göstərilir və həmin sətri seçdikdə proqnozlaşdırılan vaxt həmin zamana dəyişəcək və həmin zaman anında hava məkanındakı vəziyyət əks olunacaq. Şəkil 1-də AHY252 və THY234 HG arasında 03:19-da münaqişə vəziyyətinin yaranacağı görünür. Dispetçer bu HG-dən hər hansı birinin hündürlük və ya istiqamətinin dəyişməsi haqda qərar qəbul etməlidir. Lakin dispetçer əvvəlcə proqnozlaşdırılan vaxtı münaqişənin baş verə biləcəyi zamana yaxınlaşdıraraq hava məkanındakı gözlənilən vəziyyəti müşahidə edir (şəkil 2). Bu ona qərar qəbul etməkdə kömək göstərəcək. Şəkil 2-dən göründüyü kimi münaqişə vəziyyətində olan, 11000m hündürlükdə uçan HG-dən hər hansı birinin hündürlüyünün dəyişdirilməsinə qərar verərkən həmin hava gəmiləri ilə yolları kəsişən və yaxınlıqda olan, 10500m hündürlükdə uçan AFL043 HG-nin hündürlüyü nəzərə alınacaq və bu HG ilə də münaqişənin yaranmaması üçün mütləq digər hündürlüklərin seçilməsinə qərar veriləcək.



Şəkil 1. Cari vaxtda hava məkanındakı vəziyyətin əks olunması.

.



Şəkil 2. Proqnozlaşdırılan vaxta uyğun hava məkanındakı vəziyyətin əks olunması.

ƏDƏBİYYAT

[1] Р.М. Ахмедов, А.А. Бибутов, А.В. Васильев и др. Автоматизированные системы управления воздушным движением: Новые информационные технологии в авиации: Учеб. пособие – СПб. Политехника, 2004, 446с.

[2] Air traffic control automation system. www.indracompany.com

[3] EUROCONTROL Guidance Material for Short Term Conflict Alert. Appendix A: Reference STCA System” Edition number:2.0, Document identifier: EUROCONTROL-GUID-123, 2009.

[4] N.B.Ağayev, N.İ.İsgəndərov “Havadakı hərəkətin avtomatlaşdırılmış idarə edilməsi zamanı münaqişələrin aşkarlanması və qərar qəbul edilməsi üsullarının sistem struktur analizi” Elmi məcmuələr. Cild 15, № 2, “Azərbaycan Hava Yolları” Qapalı Səhmdar Cəmiyyəti Milli Aviasiya Akademiyası. 2013, səh. 21-27.

[5] EUROCONTROL Specification for Medium-Term Conflict Detection. Edition 1.0, EUROCONTROL-SPEC-139, 2010.

[6] Jean-Marc Alliot, Nicolas Durand. A mathematical analysis of the influence of wind uncertainty on MTCD efficiency. Dans : The Controller, IFATCA - International Federation of Air Traffic Controllers' Associations, Montreal - Quebec, Numéro spécial *Meteorology and ATC*, Vol. 1 N. Spring 2011, p. 17-19, mars 2011. http://www.alliot.fr/papers/ifatca.pdf

[7] C. Beers and D. Dehn, “Medium Term Conflict Detection (MTCD) local validation plan, For shadow mode trials at Rome Area Control Centre (ACC),” NLR-CR-2002-545, NLR, Tech. Rep., 2003.

[8] Alam S., Abbass H.A., Lokan C.J., Ellejmi M., and Kirby S. Computational red teaming to investigate failure patterns in medium-term conflict detection. Defence & Security Applications Research Centre, University of New South Wales; Australian Defence Force Academy, Canberra, Australia; EUROCONTROL Coorperative Network Design Cooperative Network Division (CND), France 01/2009.

http://www.eurocontrol.int/eec/public/standard_page/DOC_Conf_2009_009.html

[9] European ATM Master Plan, Edition 2, October 2012. www.atmmasterplan.eu

[10] N.B.Ağayev, N.İ.İsgəndərov “Vincenti üsulu ilə hava gəmisi koordinatlarının təyin edilməsi”, İnformasiya texnologiyaları problemləri, 2014, №2, 92-100.

[11] Nadir Agayev, Nasraddin Isgandarov “Flight plan based aircraft conflict detection method”, International Journal of Latest Research in Science and Technology. Volume 3, Issue 5: Pages 63-66. September-October 2014.

http://www.mnkjournals.com/ijlrst.htm.

[12] Осипов Д. Л. Базы данных и Delphi. Теория и практика. – СПб.: БХВ-Петербург, 2011.

Наиболее Распространенные Задачи Data Mining

Сайяр Абдуллаев[1], Судаба Абасова[2], Сабина Фоменко[3]
[1,2,3]Институт Информационных Технологий НАНА, Баку, Азербайджан
[1,2,3]depart5@iit.ab.az

Аннотация **– в статье рассматривается понятие Data Mining. Описываются возникновение Data Mining, основная суть задач, а в некоторых из них рассмотрены процесс и методы решения, а также применение. Сравниваются Big Data и Data Mining.**

Ключевые слова **–** ***Big Data, Data Mining, классификация, кластеризация, прогнозирование, визуализация.***

I. ВВЕДЕНИЕ

Для получения конструктивной и нужной информации прежние методы, которые применяли статики и математики, требовали большого количества времени. Это привело к формированию Data Mining.

Data Mining как термин возник в 1978 году. До середины 1990-х годов данные обрабатывались и анализировались в рамках прикладной статистики. При этом чаще всего решались задачи обработки маленьких баз данных. В этот промежуток времени понятие Data Mining начало приобретать популярность в современной интерпретации. Одним из основателей направления Data Mining является Григорий Пиатецкий-Шапиро [1].

Понятие Data Mining означает добыча (mining) данных (data). Синонимами Data Mining можно считать понятия "обнаружение знаний в базах данных" (Knowledge Discovery in Databases, KDD) и "интеллектуальный анализ данных".

Data Mining применяется везде, где имеются какие-либо данные. Но на сегодняшний день методы Data Mining в первую очередь заинтриговали коммерческие предприятия, которые развертывают проекты на основе информационных хранилищ данных.

Главной характерной чертой Data Mining является сочетание большого математического инструментария (от традиционного статистического исследования до самых новых кибернетических способов) с новейшими достижениями в области информационных технологий.

Data Mining – это мультидисциплинарная сфера, возникшая и формируемая на основе таких наук, как распознавание образов, теория баз данных, прикладная статистика, искусственный интеллект, и др. (рис. 1).

Каждый год Data Mining посвящается множество научных и практических конференций. Одной из них является Международная конференция по Knowledge Discovery Data Mining (International Conferences on Knowledge Discovery and Data Mining).

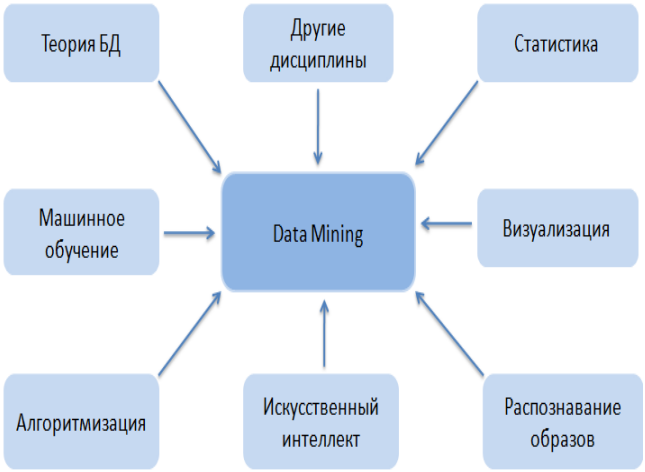


Рис. 1. Data Mining в роли мультидисциплинарной области

II. ЗАДАЧИ DATA MINING

Концепция шаблонов составляет базу технологии Data Mining. Концепция отражает в данных неожиданные, неочевидные регулярности. Эти регулярности составляют скрытые знания. Концепция представляет собой закономерности. Им свойственны выборки данных, выраженные в понятной для человека форме. Задачи Data Mining решаются результатом обнаружения незаметных закономерностей. Иногда задачи Data Mining называют закономерностями или техниками [2].

Нет единого мнения о том, какие задачи можно причислять к Data Mining. Во многих изданиях перечислены нижеприведенные задачи Data Mining:

1) классификация;
2) кластеризация;
3) прогнозирование;
4) ассоциация;
5) визуализация;
6) анализ и обнаружение отклонений;
7) оценивание;
8) анализ связей;
9) подведение итогов.

Задача классификации. Самой часто решаемой, а также самой простой задачей Data Mining является классификация.

Классификация – это упорядоченная в соответствии с определенным принципом масса объектов, имеющих схожие классификационные признаки, с помощью которых между этими объектами определяются сходства либо различия.

В задаче классификации должны соблюдаться следующие правила:

- в каждой части деления должно применяться только одно основание;
- общий объем категориальных понятий должен равняться объему родового понятия, которое делится. Деление обязательно соразмерно;
- объемы членов деления не могут пересекаться. Нужно, чтобы они взаимно исключали друг друга;
- деление обязательно должно быть последовательно.

Можно различить искусственную и естественную классификации. Искусственной называется классификация, осуществляемая по внешнему признаку. Вспомогательная классификация нужна для придания большинству процессов и явлений заданного порядка. Естественной называется классификация, которая осуществляется по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Естественная классификация есть результат и важное средство научного исследования.

Задачу классификации можно решить, используя следующие способы:

- ближайшего соседа;
- индукцию деревьев решений;
- байесовские сети;
- нейронные сети.

Задача кластеризации. В кластеризации классы исследуемого набора данных предварительно неопределенны.

Кластеризация используется с целью разбиения совокупности объектов в схожие кластеры либо классы. Представив данные выборки в виде точек в признаковом пространстве, задачу кластеризации можно свести к понятию "сгущений точек".

С помощью кластеризации можно провести исследовательский анализ с целью изучения "структуры данных".

Кластеризация применяется во многих отраслях. Первоначально она использовалась в биологии, антропологии и психологии.

Кластеры бывают непересекающимися и пересекающимися.

Иногда применение всевозможных способов кластеризации приводит к различным результатам. Это зависит от особенностей работы какого-либо алгоритма. Существует более сотни разнообразных алгоритмов кластеризации.

Примером способа решения задачи кластеризации является обучение "без учителя".

Задача прогнозирования. В самых различных сферах человеческой деятельности можно столкнуться с задачей прогнозирования. Развитие информационных технологий и методов прогнозирования напрямую связано друг с другом. Так как оно требует детального исследования исходного набора подходящих для анализа данных и методов, его можно считать одной из самых сложных задач Data Mining. Цель прогнозирования – это предсказание будущих событий.

Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе анализа его состояния в прошлом и настоящем. Следовательно, решение задачи прогнозирования требует некоторой обучающей выборки данных.

При решении задачи прогнозирования устанавливается функциональная зависимость между независимыми и зависимыми переменными. Прогнозирование является распространенной и востребованной задачей во многих областях человеческой деятельности. В результате прогнозирования уменьшается риск принятия неверных, необоснованных или субъективных решений.

Задача ассоциации. При решении задачи ассоциации отыскиваются закономерности среди связанных событий в наборе данных.

Ассоциации отличаются от классификации и кластеризации тем, что в них поиск закономерностей происходит между несколькими событиями, протекающими одновременно.

Задача поиска ассоциативных правил решается с помощью многих алгоритмов.

Задача визуализации. Визуализация позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом. Результатом использования визуализации является графический образ данных. Визуализация представляется в виде гистограмм, графиков, диаграмм, схем и т.д. Визуализацию используют в задачах классификации и кластеризации.

В визуализации практически полностью отсутствует необходимость в специальной подготовке пользователей.

Задачами визуализации как компонента Data Mining, являются: поддержка интерактивного и согласованного исследования; помощь в предоставлении результатов; использование зрения для создания зрительных образов и осмысление их.

Анализ связей. Это задача, с помощью которой находятся зависимости в наборе данных.

Подведение итогов. Целью этой задачи является описание конкретных групп объектов, выбранных из анализируемого набора данных [3].

III. DATA MINING И BIG DATA

Big Data и Data Mining – это два разных понятия. Каждое из понятий применяется в работе по обработке больших объемов данных. Кроме того, эти два термина используются для двух различных направлений такого рода работы.

Big Data – это термин, означающий набор больших данных. Big Data являются те данные, которые уже перерастают простые базы данных, используемые в ранние периоды, стоили они дороже и имели меньше возможностей. Примером этого могут служить наборы данных, которые слишком большие для обработки в разных системах.

Понятие Big Data в последнее время, можно сказать, стало мейнстримом. С Big Data тесно связаны такие термины, как анализ данных, наука о данных, аналитика данных, сбор данных и машинное обучение.

В отличие от Big Data, Data Mining используется для поиска нужной информации среди большого набора данных. Синонимом этого вида деятельности можно считать высказывание "искать иголку в стоге сена". То есть автоматически собираются большие наборы данных, которые могут быть однородными, похожими. Data Mining может включать в себя использование различных программ [5].

Data Mining – это процесс очистки больших данных и подготовки их последующему анализу или использованию в алгоритмах машинного обучения. Data Mining должны обладать исключительными распознавательными качествами, чудесной интуицией и техническими умениями для объединения и трансформирования огромного количества данных [6].

ЗАКЛЮЧЕНИЕ

Таким образом, технология Data Mining с каждым днем становится все более востребованной. С помощью Data Mining можно решить множество задач и обработать большое количество данных.

На сегодняшний день существует несколько точек зрения на Data Mining. Некоторые считают, что это мираж, который отвлекает внимание от классического анализа данных. Другие же считают, что Data Mining – это альтернатива традиционному подходу к анализу. А также, как и во всех случаях, существует середина, которая рассматривает возможность совместного использования современных достижений в области Data Mining и при классическом статистическом анализе данных.

Для того чтобы пользоваться Data Mining, нужна специальная квалификация. Не имея ее, очень трудно работать с Data Mining. Квалификация должна постоянно повышаться.

ЛИТЕРАТУРА

[1] И.А.Чубукова, Data Minig. Курс лекций интернет-университета INTUIT, 2008, с. 7–16.
[2] В.Дюк, А.Самойленко. Data Mining. Издательство "Питер", 2001, с. 14–20.
[3] http://bug.kpi.ua/stud/work/RGR/DATAMINING/tasksofdm.html
[4] www.kdnuggets.com
[5] https://www.techopedia.com/7/29678/technology-trends/what-is-the-difference-between-big-data-and-data-mining
[6] http://habrahabr.ru/company/io/blog/264125/

О Методах Сбора и Хранения Сетевого Трафика Компьютерных Сетей

Рамиз Шыхалиев
Институт Информационных Технологий НАНА, Баку, Азербайджан
ramiz@science.az

Аннотация **− процесс сбора и хранения сетевого трафика компьютерных сетей (КС) является одним из основных этапов процесса мониторинга. Однако в современных КС процесс сбора и хранения полного сетевого трафика представляет собой очень сложную проблему. Так как с ростом скорости и масштаба КС растут и объемы сетевого трафика, который в день может потребовать петабайты объемов памяти для хранения. Существуют различные методы сбора и хранения сетевых данных, которые при правильном выборе могут существенно снизить объем собранных данных, следовательно, и объем требуемых для хранения данных. Исходя из этого, в статье рассматриваются подходы к решению вопросов сбора и хранения данных, а также проблемы, имеющиеся в этой области.**

Ключевые слова − компьютерные сети, мониторинг, сетевой трафик, сбор сетевого трафика, хранение сетевого трафика.

I. ВВЕДЕНИЕ

Сегодня компьютерные сети, особенно Интернет, стали глобальной инфраструктурой, обеспечивающей повсеместно доступные, интерактивные и безопасные сервисы. Для того чтобы обеспечить высокий уровень QoS (Quality of Service) этих сервисов, необходима эффективная инфраструктура мониторинга. При этом самым подходящим методом для инфраструктуры мониторинга КС является пассивный мониторинг [1], который позволит определить общее состояние КС и безопасность, а также уровень QoS предоставляемых сервисов и т.д. Для этого необходимо постоянно осуществлять сбор, хранение и анализ большого объема данных, то есть постоянно осуществлять сбор, хранение и анализ сетевого трафика. Однако это представляет из себя очень сложную задачу, в особенности при мониторинге больших КС. Так как с повышением скорости и масштабов КС растет и объем сетевого трафика, который в день может измеряться петабайтами. Вместе с тем для того, чтобы провести всеобъемлющий анализ состояния и безопасности КС, необходимо произвести сбор и хранение всей информации о трафике. Например, при мониторинге безопасности КС сбор всех сетевых пакетов необходим для обнаружения вредоносных активностей или определения вирусных атак (например сетевыми червями) и т.д.

Известно, что при пассивном мониторинге КС собираются большие объемы данных мониторинга, что приводит к возникновению проблем, связанных с их хранением и уменьшением эффективности анализа. Поэтому, наряду с необходимостью разработки новых методов анализа больших объемов сетевых трафиков, актуальной проблемой является разработка новых подходов для сбора и хранения больших сетевых трафиков для мониторинга КС. При этом очень важным является сокращение размерности признакового пространства сетевого трафика, который используется для мониторинга КС [2]. Исходя из этого, в статье рассматриваются подходы к решению вопросов сбора и хранения, а также проблемы, имеющиеся в этой области.

II. МЕТОДЫ СБОРА СЕТЕВОГО ТРАФИКА

Существующие сегодня средства мониторинга осуществляют сбор различных типов и объемов информации. При этом имеются три основных метода сбора трафика, которые имеют различные требования к объему памяти: сбор всех пакетов; сбор сетевого потока и так называемый сбор расширенного потока.

Целью сбора всех пакетов является сбор всего сетевого трафика, который генерируется компьютерами и устройствами КС, при котором осуществляются сбор и хранение данных заголовка каждого пакета и передаваемой в пакетах информации. Другими словами, производятся сбор, обработка и хранение копии каждого пакета трафика для последующего анализа. Эти собранные данные обеспечивают аналитиков полной информацией о трафике: полной информацией заголовков пакетов и передаваемой в пакетах информацией. Следовательно, такой метод сбора данных мониторинга может быть наиболее универсальным, так как большой объем информации может интенсивно храниться и обрабатываться [3].

Сетевой поток определяется как множество IP-пакетов, проходящих через точку наблюдения в сети в течение определенного интервала времени. Все пакеты, принадлежащие к определенному потоку, имеют набор общих свойств. Требования к потокам IP-пакетов определены в RFC 3917 [4]. Согласно приведенному определению, сетевой поток представляет из себя потоки сетевых пакетов, для которых выполняются следующие условия:

- происходят в течение одного и того же периода времени;
- имеют один и тот же адрес источника и номер порта;

- имеют один и тот же адрес назначения и номер порта;
- используют один и тот же протокол.

При этом, если не рассматривать передаваемую в пакетах информацию и информацию некоторых полей заголовков пакетов, а также объединить некоторые пакеты, то уменьшится объем данных, что приводит к уменьшению требуемой памяти для хранения потоков. Однако это вызывает понижение качества анализа сетевого трафика [5].

Сбор расширенного потока включает в себя сбор всех пакетов и сетевого потока. При этом к информации потока добавляется информация, взятая непосредственно из заголовков пакетов или из передаваемой в пакетах информации. Вместе с тем расширенный поток также может содержать дополнительную информацию о каком-то внешнем источнике, например, о географическом расположении IP-адресов источника и назначения. Поэтому некоторые решения сбора расширенного потока рассматривают эту информацию как метаданные [6].

Большая часть современных исследований в области сбора сетевого трафика посвящена вопросам сбора пакетов в скоростных сетях с минимальной потерей данных и сжатием данных после сбора, то есть снижению объемов. Например, в работах [7, 8] авторы соответственно обсуждают вопросы преобразования данных для их эффективного хранения и обработки и мониторинга в облаке (Cloud). В работах [9, 10] авторы предлагают подход к разработке приложений по сбору данных в скоростных сетях, основанный на стандартных аппаратных средствах. А в работе [11] для полного анализа сетевого трафика авторы предлагают метод агрегации потоков.

III. МЕТОДЫ ХРАНЕНИЯ СЕТЕВОГО ТРАФИКА

Другой проблемой эффективного анализа сетевого трафика является хранение собранных данных, которые должны сохраняться достаточно долго и надежно, чтобы при необходимости аналитики могли иметь доступ к ним. При этом, в зависимости от места и способа хранения данных, может существенно изменяться требуемый объем памяти для хранения, а также возникать проблемы, связанные с администрированием и обслуживанием и т.д. Вместе с тем данные могут быть сохранены локально в организации, облаке или другом внешнем хранилище и могут быть использованы различные способы хранения данных, такие, как файлы (например, лог-файлы); базы данных и их комбинации. Каждый из этих способов имеет собственные аспекты.

Обычно в большинстве организаций КС производят сбор сетевого трафика в нескольких точках. Поэтому очень важен выбор места физического расположения собранных данных. Например, централизованное хранение всех собранных данных в одном месте может упростить управление и анализ данных, однако требует передачи данных в центр, что приводит к неэффективному использованию полосы пропускания каналов передачи сети. Также такой способ хранения данных нецелесообразен с точки зрения безопасности, так как при компрометации хранилища может произойти несанкционированное изъятие данных. Альтернативой централизованного хранения данных является распределенное хранение данных, однако при таком подходе хранения усложняется процесс анализа данных, а также администрирования и обслуживания. Одним из видов распределенного хранения данных можно считать облачное хранение данных [12, 13], который может также осуществлять и сбор данных.

Большинство средств сбора сетевого трафика записывают полученные данные в файлы (лог-файлы) и обычно имеют собственные форматы файлов. При этом очень важно знать формат хранимого файла, чтобы без затруднения организовать передачу данных между приложениями сбора и анализа данных, так как большинство из них поддерживают определенные форматы файлов. Однако имеются некоторые общие форматы (например, pcap), которые поддерживаются большинством приложений сбора и анализа данных. Вместе с тем форматы создаваемых файлов могут определить необходимые объемы для хранения файлов. Несмотря на то, что в малых объемах данных различия между форматами не существенны, в больших объемах данных выбор того или иного формата является существенным. Уменьшение объема памяти, необходимого для хранения данных, также может быть достигнуто сжатием данных, которое может сделать хранение и анализ данных более эффективными. Эффективный алгоритм сжатия может не только уменьшить дисковое пространство, необходимое для хранения данных, но также уменьшить время, требуемое для извлечения данных из этого диска. Например, алгоритм сжатия данных lzo1x может уменьшить размер записей примерно на 50% [14].

Некоторые средства сбора данных для хранения данных могут использовать базы данных. Приложения, которые поддерживают только файлы, могут хранить их в базе сами или с помощью приложения анализа данных. Аналитик это может сделать вручную. Вместе с тем при использовании базы данных для хранения сетевого трафика необходимо учитывать ожидаемый размер и количество записей и свойства базы данных, ограничивающих общий размер базы данных, размер записи и т.д. Учитывая то, что реляционные базы данных не так масштабируемы, как при хранении данных в виде файлов. Для решения проблемы масштабируемости могут быть использованы NoSQL базы данных, такие, как Hadoop [15].

ЗАКЛЮЧЕНИЕ

Сегодня мониторинг сетевого трафика КС является одним из основных средств обеспечения их нормальной работы и безопасности. При этом сбор и хранение являются основой мониторинга. Для получения полной информации о деятельности КС необходимы постоянный и полный сбор и хранение сетевого трафика, что позволяет своевременно и эффективно реагировать на отказы в работе и инциденты безопасности. Однако это требует постоянного сбора и хранения большого объема данных

мониторинга, что может требовать огромных объемов памяти, а также снизить эффективность анализа собранных данных. Другой проблемой сбора и хранения большого объема сетевого трафика, на наш взгляд, является неправильный выбор соответствующих методов сбора и хранения данных в зависимости от задач мониторинга.

В статье были проанализированы существующие методы сбора и хранения сетевого трафика, а также проблемы, имеющиеся в этой области.

В результате анализа можно сделать вывод, что при сборе сетевого потока или расширенного потока требуется гораздо меньше памяти для хранения. Проведенный в статье анализ методов сбора и хранения сетевого трафика поможет администраторам КС в соответствии с задачей мониторинга выбрать необходимый метод.

ЛИТЕРАТУРА

[1] Şıxəliyev R.H. Kompüer şəbəkələrinin monitorinqi üsulları və vasitələri haqqında // İnformasiya cəmiyyəti problemləri, №2(4), səh. 61-70, 2011.

[2] Шыхалиев Р.Г. Об одном методе сокращения размерности анализируемых признаков сетевых трафиков, используемых для мониторинга компьютерных сетей // Телекоммуникации. - N: 06. - с. 44-48, 2011.

[3] Bejtlich Richard., Why Collect Full Content Data? http://taosecurity.blogspot.com', 2012

[4] Quittek J., Zseby T., Claise B., Zander S., RFC 3917: Requirements for IP Flow Information Export (IPFIX). Internet Engineering Task Force, 2004. http://tools.ietf.org/html/rfc3917

[5] RFC 7011, Specification of the IP Flow Information Export (IPFIX) Protocol, a standardized network flow format, provides a more technical definition of flow. http://tools.ietf.org/search/rfc7011

[6] National Information Standards Organization (NISO). Understanding Metadata. NISO, 2004.

[7] Aceto G., Botta A., Pescape A., Westphal C. Efficient Storage and Processing of High-Volume Network Monitoring Data, IEEE Transactions on Network and Service Management, vol. 10, issue 2, pp. 162-175, 2013.

[8] Aceto G., Botta A., de Donato W., Pescape A., Cloud Monitoring: A Survey, Computer Networks vol.57, issue 9, pp. 2093-2115, 2013.

[9] Deri L., Cardigliano A., Fusco F., 10 Gbit Line Rate Packet-to-Disk Using n2disk, Proceedings IEEE INFOCOM. Turin, Italy, Apr. IEEE, pp. 3399-3404, 2013.

[10] Banks D., Custom Full Packet Capture System, SANS, 2013.

[11] Francois J., State R., Engel T., Aggregated Representations and Metrics for Scalable Flow Analysis, IEEE Conference on Communications and Network Security (CNS). Washington, D.C., pp. 478-482, 2013.

[12] Sivashakthi T., Prabakaran N., A Survey on Storage Techniques in Cloud Computing, International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue 12, pp. 125-128, 2013.

[13] Spoorthy V., Mamatha M., Santhosh Kumar B., A Survey on Data Storage and Security in Cloud Computing, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.6, 2014, pp. 306-313.

[14] Software Engineering Institute, Carnegie Mellon University. SiLK FAQ https://tools.netsa.cert.org/silk/faq.html (2014).

[15] Hadoop, http://hadoop.apache.org/

Veb İstifadəçilərin Profilinin Qurulması Metodu

Babək Nəbiyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
babek@iit.ab.az

Xülasə – **Kompüter şəbəkələrinin təhlükəsizliyinin təmin olunması və prosesin optimallaşdırılması üçün bir çox vasitələr mövcuddur. Məlumdur ki, təhlükələrin yaranmasının əsas səbəblərindən biri şəbəkə trafikində anomal və qeyri profil trafikin generasiya olunmasıdır. Bunları nəzərə alaraq şəbəkə trafikində istifadəçilərin profilinin müəyyən olunması üçün vasitə işlənib hazırlanmışdır. İstifadəçilərin profilinin müəyyən olunması üçün K-orta klasterizasiya metodu tətbiq olunmuşdur.**

Açar sözlər – ***şəbəkə trafiki, klasterizasiya, istifadəçi profili, anomal trafik***

I. GİRİŞ

İnternet vasitəsilə qloballaşan dünyada hər bir resurs və ya informasiyanı sürətlə əldə etmək çox asan olmuşdur. Bu informasiya cəmiyyəti nöqteyi nəzərindən çox müsbət haldır. Amma bildiyimiz kimi generasiya olunan informasiyaların heç də hamısı məqsədəuyğun olmur. Bu isə öz növbəsində kompüter şəbəkələrində lazımsız yük yaradaraq əlaqə kanallarının əlyetənlik qabiliyyətini aşağı salır. Bu hadisə şəbəkədən istifadə prosesini davranış profilinə uyğunlaşdırmayan korporativ şəbəkələrin gec-tez qarşılaşa biləcəyi hadisələrdəndir.

Symantec şirkətinin 2014-ci ildə verdiyi hesabata əsasən [1], bir gün ərzində veb-resurslarda qarşısı alınan hücumların sayı 586700 – ə bərabərdir. Bunları nəzərə alaraq, şəbəkə istifadəçilərinin korporativ resurslardan düzgün istifadə edərək təhdidlərlə qarşılaşmaması, məhdud olan informasiya kanalını yersiz yükləməməsi və faydalı iş qabiliyyətinin yüksəldilməsi üçün şəbəkə trafikinin klasterizasiya metodu əsasında şəbəkə trafikində istifadəçilərin davranış profilinin (bundan sonra istifadəçi profili) müəyyən olunması təklif olunur. Şəbəkə trafikinin monitorinqi vasitəsilə alınan verilənlərin klasterizasiya dəyərləri əsasında analizi aparılaraq müyyən istifadəçinin davranış klasterləri və bu prosesin reallaşdırılması K-Orta klasterizasiya alqoritmi vasitəsilə həyata keçirilir.

II. ƏLAQƏDAR TƏDQİQATLARIN ANALİZİ

Şəbəkə trafikinin identifikasiyası və kateqoriyalaşdırılması şəbəkə idarəolunmasının əsas elemetlərindən biridir, buna axının prioritetləşdirilməsi, trafikin formalaşdırılması və siyasəti, diaqnostik monitorinqi misal gətirmək olar.

Bütün dünyada IP şəbəkələr vasitəsilə böyük həcmli informasiyalar ötürülür və qəbul olunur. Mütəxəssislər bütün bu prosesi nəzarətdə saxlayır və bunun vasitəsilə təhdidləri müəyyən edərək aradan qaldırırlar. IP paketlər istifadəçilər haqqında bir çox informasiyanı əldə etməyə imkan verir. IP paketləri analizin edərək şəbəkənin idarəolunması və optimimallaşdırılması, təhdidlərin aradan qaldırılması zamanı informasiya mənbəyi kimi istifadə etmək mümkündür. [2]-də IP paketlərin başlıqlarından istifadə edərək şəbəkənin axın prosesini və istifadəçilərin davranış profilini geniş formada izah edən çoxsəviyyəli klasterizasiya metodu təklif olunur. Əlavə olaraq demək lazımdır ki, IP paketlərin başlıqlarından istifadə edərək aparılan analiz prosesi istifadəçilərin şəxsi məlumat toxunulmazlığını təmin edir.

Şəbəkə trafiki və ya ümumiyyətlə şəbəkə haqqında toplanmış loq-fayllar vasitəsilə anomaliyaların və təhdilərin aşkarlanmasını həyata keçirmək olar. Bu proses üçün müxtəlif metodlardan, vasitələrdən istifadə olunur. Məsələn, [3]-də K-Orta klasterləşmə alqoritmindən istifadə edərək trafik axınında anomaliyaları aşkarlamaq təklif olunur. Şəbəkə trafikinin işarələnməmiş verilənləri iki klasterə bölünür, bunlar normal və anomal trafiklərdir. Yeni monitorinq verilənlərində anomaliyaların aşkarlanması əsasında effektiv məsafənin seçilməsi üçün müəyyən olunmuş klasterlərdə şablon olaraq ağırlıq mərkəzi istifadə olunur.

Mərkəzi idarəetmə olmadan özü təşkilatlanan və nəzarət prosesi olmayan klasterləşmə metodu ən yeni yaxınlaşmalardan biridir. Bunun üçün [4]-də qarşılıqlı əlaqəyə əsaslanan qarışqa davranışı metodu istifadə olunur. Bu metodun üstünlüyü ondan ibarətdir ki, ilkin verilənlərə və ya klasterlərin sayının əvvəlcədən müəyyən olunmasına ehtiyac yoxdur. Virtual qarışqalar hər biri ayrı-ayrılıqda şəbəkəni tədqiq edərək klasterləmə prosesini yerinə yetirirlər. Amma bu yeni metod olduğuna görə aparılan prosesin dəqiqlik əmsalı şübhə doğurur.

“Machine learning” yanaşması şəbəkə trafikində anomal axınların unikal statistik xarakteristikalara əsaslanaraq müəyyən olunması üçün geniş istifadə olunur. Qeyri-səlis klasterləşdirmə ənənəvi klasterləşdirməyə nəzərən daha çevikdir, müdaxilələrin aşkarlanması və verilənlərin təbii emalı üçün daha məqsədəuyğundur [5].

Bir çox klasterləşdirmə metodları müdaxilələrin aşkarlanması üçün normal və anomal trafikin ayrılmasını nəzərdə tutur. Klasterləşdirmə metodları trafik sessiyalarının fərqlərini və oxşarlıqlarını tapmaq, onların hər birini müvafiq qruplara bölərək təsnif etmək üçün tətbiq edilir [6]. Bu qruplar onlara verilmiş nişanlar ilə təmsil olunur. Daha sonra bu nişanlar daxil olan şəbəkə trafikinin növünü proqnozlaşdırmaq üçün istifadə olunur.

Şəbəkə trafikinin tez və dəqiq indentifikasiyası QoS-un idarə edilməsi, şəbəkə təhlükəsizliyinin monitorinqi və s. funksiyalar üçün ən vacib məsələlərdən biridir. Lakin son zamanlar P2P-dən istifadə edən qovşaqlar çoxalıb və onlar müxtəlif portlardan istifadə edərək özünü hər hansı qurğu, lazımlı məlumat axını və ya sifrlənmiş məlumat axını altında gizlədərək lazımsız informasiya axınını generasiya edirlər. Bu halda klassik yanaşmalar sayılan “port mapping” və ya “payload analysis” yanaşmalarının istifadəsi effektiv deyil. Alternativ yanaşma şəbəkədə TCP trafiki ilə əlaqədar ilk bir neçə paket daxilində davranışı tədqiq edərək klassifikasiya etməkdir. Bu gələcəkdə bütün informasiyanı klasterləşdirərək identifikasiya prosesini asanlaşdırmaq üçün istifadə etməyə imkan verərdi [7].

III. VEB LOQ FAYLLAR

Verilənlər 5000-dən çox IP ünvandan ibarət olan AzScienceNet şəbəkə mühitində toplanıb və bu şəbəkə də özlüyündə bir neçə xırda şəbəkəyə bölünür. İstifadəçilərin məxfiliyinin pozulmaması məqsədilə ilə AzScienceNet şəbəkəsinin istifadəçi siyasətinə əsaslanmış və əlavə olaraq istifadəçilər kimliyi haqqında verilənlər adsızlaşdırılmışdır. Bu verilənlər 10 dəyişəndən ibarətdir [7], bunlar cədvəl 1-də göstərilmişdir.

Cədvəl 1-də göstərilən 7 dəyişəni aşağıdakı kimi izah etmək olar:

1. Unix zaman damğası – bu Squid tərəfindən əməliyyatın qeyd olunduğu vaxtdır və bir qayda olaraq əməliyyatın həyat dövrünün bitdiyi vaxtı yəni, müraciət olunan informasiya təmamilə əldə olunduğu vaxtı qeyd edir.
2. Əməliyyat zamanı - əməliyyatda keşdə nə qədər zaman keçirdiyini qeyd edir. Yəni, HTTP paketlərin əməliyyat başlayandan son bayt ötürülənə qədər olan zaman kəsiyi. Unix zaman damğasının qeyd etdiyi vaxtdan əməliyyat zamanını çıxdığınız halda əməliyyatın başlama anını əldə etmək mümkündür.
3. Lokal IP ünvan – informasiya üçün müraciət edən IP ünvanlar qeyd olunur.
4. Nəticə kodları – özündə müraciətlərin cavablandırılması, imtina olunması və s. haqqında məlumat toplayır.
5. Həcm (bayt) - Göndərilən və qəbul olunan bütün paketlərin kontent həcmi qeydə alınır. Bu ümumi trafikin həcmini müəyyən etmək üçün vacibdir.
6. Sorğu metodu – adətən böyük hərflərlə yazılmış kiçik ingilis sözlərindən ibarət olur GET, HEAD və s. Bu metodlar əsasında veb resurs istifadəçinin nə üçün sorğu göndərdiyini müəyyən edir və ona uyğun cavab verir.
7. URL (Uniform Resource Locator) – şəbəkə istifadəçiləri tərəfindən müraciət olunan URL liklər qeyd olunur.
8. Ierarxiya kodu – müraciətlərin emal formaları haqqında məlumat verir. Məsələn müraciət bir başa cavablandırıldı və ya partnyor serverə göndərildi və s.
9. Təyinat IP-si – müraciətə cavab verən IP ünvan.
10. Məzmun – HTTP cavabın başlığında yerləşir və obyektin məzmun növünü göstərir.

İndeks	Dəyişənlərin təsviri
1	Unix zaman damğası
2	Əməliyyat zamanı (ms)
3	Local IP
4	Nəticə kodları
5	Həcm (bayt)
6	Sorğu metodu
7	URL
8	İerarxiya kodu
9	Təyinatlı IP
10	Məzmun

Cədvəl 1. Squid loq dəyişənlərinin təsviri

Bütün bu məlumatlar Squid proxy server vasitəsilə toplanır. Squid proxy server [8] - şəbəkə trafikinin loq fayllarının toplanması və idarə olunması prosesini reallaşdırmaq üçün istifadə edilir. Squid proxy server açıq kodlu proqram təminatıdır və bir gün ərzində internetlə işləyən istifadəçilərin sayı 2000-dən çox olan böyük şəbəkələrdə istifadə olunması əlverişlidir. Squid proxy serverin əsas üstünlüyü keşlənən proxy server olmasıdır, bu halda müraciət olunan resurslar keşdə toplanır və onlara yenidən müraciət olunduğu halda emal prosesi nisbətən daha sürətlə sona çatır. Bu da öz növbəsində şəbəkənin əlyetənliyinə müsbət təsir göstərir. Squid proxy server vasitəsilə toplanan loq fayllar xüsusi verilənlər bazasında toplanaraq analiz prosesində istifadə olunur.

IV. LOQ MƏLUMATLARIN TƏMİZLƏNMƏSİ

Squid proxy server vasitəsilə toplanan loq fayllar müxtəlif interpretasiyalar üçün geniş imkanlar yaradır. Bu da öz növbəsində bir növ loq fayldan müxtəlif məqsədlər üçün istifadə edilməsinə şərait yaradır. Cədvəl 2-də squid loq serverin topladığı verilənlər göstərilmişdir. Amma konkret olaraq bu məqalə çərçivəsində squid proxy serverin təqdim etdiyi 10 dəyişənin hamısına ehtiyac yoxdur. İnformasiya təhlükəsizliyi yanaşmasında istifadəçi profilinin identifikasiyası üçün müraciətin məzmunu, təyinat ip-si, http ierarxik kodu, sorğu metodu və nəticə kodları kimi məlumatlar gərəkli deyil. Ona görədə loq-faylların analizi zamanı emal prosesinin asanlaşdırılması və sürətləndirilməsi üçün bu dəyişənlər nəzərə alınmır.

V. İSTIFADƏÇİ PROFİLİNİN İDENTİFİKASİYASI

İstifadəçinin profili dedikdə - müraciət olunan veb resurslar əsasında tərtib olunan maraqlar vektoru və tematik seçimlər nəzərdə tutulur istifadəçilərin tematik profillər toplusu matris yaradır. Bu matrisdə hər bir sətir istifadəçi, hər bir sütun isə əlamətləri göstərir.

UNIX zaman damğası	Əməliyyat zamanı (ms)	Local IP	Nəticə kodları	Həcm (bayt)	Sorğu metodu	URL	İyerarxiya kodu	Təyinat İP-si	Məzmun
1444780867.298	39	10.100.80.51	TCP_MISS/200	10946	GET	http://pagead2.googlesyndication.com	HIER_DIRECT	216.58.208.98	application/x-shockwave-flash
1444795608.042	3598	10.100.80.23	TCP_MISS/301	567	POST	http://v.icecentury.com/	HIER_DIRECT	54.169.165.185	text/html
1444795738.177	222	10.100.80.14	TCP_MISS/304	318	GET	http://code.createjs.com	HIER_DIRECT	23.77.228.124	application/x-javascript
1444799392.183	38	10.100.80.61	TCP_MISS/200	345	HEAD	http://ds.download.windowsupdate.com	HIER_DIREC	188.43.72.35	application/octet-stream

Cədvəl 2. Squid loq serverin topladığı verilənlər

İstifadəçilərin davranış kateqoriyalarına daxil olan resurslara müraciətlərin tezliyindən və daxil olan trafikin həcmindən asılı olaraq əlamətlərin qiyməti hesablanır. Modelin keyfiyyətinin yüksəldilməsi üçün xüsusiyyətlərin normallaşdırılması prosesi aparılaraq [0;1] aralığına gətirilir.

Əlamətlərin lahiyələndirmə prosesi bitdikdən sonra, modelin qurulması üçün daha informativ və dolğun əlamətlər seçilir. Bu emal olunan məlumatların həcmini azaldır, yenidən təlim prosesindən keçməməsinə şərait yaradır və ümumiyyətlə, modelin keyfiyyətini yüksəldir. Baxılan halda resurlar tematik kateqoriyalara uyğun qruplaşdırılır. Məlumdur ki, bir tematik kateqoriyaya aid ola bilən resurslar müxtəlif mənbələrdə yerləşə bilər.

Data mining məsələsinin ilkin mərhələsi əlamətlərin layihələndirilməsidir (feature engineering). Bu ən məsuliyyətli və əziyyətli mərhələ olması ilə yanaşı, prosesin nəticəsinə birbaşa təsir edir. Baxılan halda obyeklər kimi şəbəkə istifadəçiləri, əlamətlər kimi isə onların müraciət etdiyi veb-resurslar nəzərdə tutulur. Bu əlamətlərin təsviri nəticəsində istifadəçilərin tematik profili formalaşır. Nəticədə informativ əlamətlərdən ibarət olan istifadəçilər/kateqoriyalar matrisini əldə edirik. Alınmış matris böyük ölçülərə malikdir (cəd. 3),

	Kat1/ həcm	Kat2/ həcm	Kat3/ həcm	Kat1/ zaman	Kat2/ zaman	Kat3/ zaman	Kat1/ müra.	Kat2/ müra.	Kat3/ müra.
İst 1	12Gb	6Gb	800Mb	126dəq	98dəq	22dəq	5355	4742	1586
İst 2	14Gb	4,8Gb	350Mb	148dəq	71dəq	18dəq	10163	3102	1475
İst 3	3,1Gb	2,7Gb	787Mb	78dəq	38dəq	28dəq	608	1554	3217

Cədvəl 3. İnformativ əlamətlərdən ibarət olan istifadəçilər/kateqoriyalar matrisini

bununla belə seyrək matris formasındadır (sparse matrix).

VI. K-ORTA ALQORİTMİ

İstifadəçi profilinin klasterizasiyası üçün biz K-Orta alqoritmini istifadə edəcəyik. Buna səbəb K-Orta alqoritmin klasterizasiya məsələsini həll etmək üçün çox sürətli və sadə olmasıdır. Əgər, $X = \{x_1, \dots, x_n\}$ isə verilənlər toplusu n trafik sessiyalarından ibarətdir. Hər bir trafik-sessiyası d-ölçülü Evklid mühitində bir veriləndir. $x_i = (f_1, \dots, f_d)$, i trafik-sessiyası üçün $f_1, \dots, f_d$ olduğu halda d xüsusiyyətlər dəyəridir. Əsas məqsəd trafik-sessiyalarını ilə klasterlərə bölməkdir. Bu proses zamanı, *n* verilənlərlə müvafiq K klaster "sentroid"-ləri arasındakı məsafə minimum olsun. Hər bir klasterin sentroid kimi tanınan mərkəzi μ_k var və bu qrupun təmsilçisi hesab edilə bilər.

Belə ki, K-orta alqoritminin girişi "nxd" verilənlər matrisi n – dən ibarətdir, K klasterlərin sayı və ilkin verilənlər isə sentroidlərdir. Burada alqoritmin aşağıdakı mərhələlərdən ibarətdir:

1. İlk olaraq "sentroid" qruplarını təmsil edəcək K nöqtələr müəyyən olunmalıdır;
2. Hər bir verilən ilə ən yaxın "sentroid" arasında Evklid-məsafəsini hesablamaq üçün (1) tənliyindən istifadə olunur:

$$məsafə(x, y) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}} \quad (1)$$

3. Bütün nöqtələr müəyyən olunduqdan sonra, K sentroidlərin mövqeləri yenidən hesablanır və bu o

deməkdir ki, bütün nöqtələrin ortası yenidən təyin olunur.

4. 2-ci və 3-cü bənd sentroidlər mövqeyini dəyişməyənədək təkrarlanmalıdır.

VII. KLASTERLƏRİN SAYININ SEÇİLMƏSİ

Bu bolmədə K-orta alqoritmini tətbiq etməzdən öncə klasterlərin sayını necə seçdiyimiz izah olunacaq. Birinci, nöqtə və cetroid arasındakı məsafəni müəyyən edən klaster-daxili məsafə ölçülür. Bundan sonra bütün bu məsələrin orta müəyyən olunur (2):

$$daxili = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}\|x - x_n\|^2 \qquad (2)$$

buradan, N sessiyaların (nöqtələrin) sayı, K klasterlərin sayı, z_i isə C_i klasterin sentroididir. Sonra isə klasterlər-arası məsafə ölçülməlidir və onlar bir birlərindən nə qədər uzaq olsa o qədər yaxşıdır. Bunun üçün düstur (3) istifadə olunur:

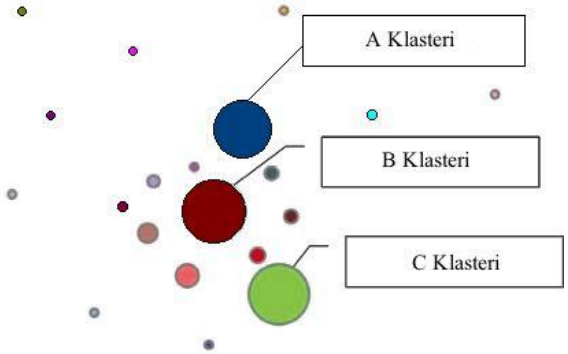
$$arası = min\left(\left\|z_i - z_j\right\|^2\right), i = 1,2,\dots,K-1; \quad j = i+1,\dots,K \qquad (3)$$

Düstur (4) ilə verilmiş etibarlılıq-ölçüsünun minimum dəyəri, K klasterlərin sayını müəyyən etmək üçün istifadə olunur:

$$\text{ölçü} = \frac{daxili}{arası} \qquad (4)$$

VIII. TƏTBİQ

Klasterizasiya modelinin tətbiqi nəticəsində, müəyyən klasterlər formalaşmışdır. Klasterləri əsasən formalaşdıran sosial şəbəkələr, video resurslar və elmi-praktiki resurslardır (şək. 1). Şəkil 1 – də göstərilən nəticə 20 klaster üçün bigml.com resursu [9] vasitəsilə əldə edilmişdir. Ən çox müaraciət olunan A klasteri elmi-praktiki resurslardan ibarətdir. 2-ci ən çox müraciət olunan B klasteri isə sosial şəbəkələrdir. C klasteridir isə, video resurslara olan müraciətlərdən ibarətdir. Digər klasterlərə isə müraciətlər daha azdır. Çünki, istifadəçilərin çoxsu lazım olan informasiyanın çox hissəsini sosial şəbəkələrdən və video resurslardan alırlar.



NƏTİCƏ

Bu məqalədə, AzScienceNet istifadəçilərinin, klasterizasiyası əsasında profilinin müəyyən olunması vasitəsi işlənib hazırlanmışdır. Bunun üçün klasterizasiya metodlarının içində ən sürətli və sadə model olan K-Orta seçilmişdir. Bu prosesin əsas məqsədi şəbəkə resurslarının məqsədə uyğun paylanması, şəbəkə trafikinin optimallaşdırılması, anomal aktivliyin mənbəyinin müəyyən olunması və təhdidlərin vaxtında aradan qaldırılmasını təmin etməkdir.

ƏDƏBİYYAT

[1] http://www.itu.int/en/ITUD/Cybersecurity/Documents/Symantec_annual_internet_threat_report_ITU2014.pdf

[2] Kumpulainen P., Hätönen K., Knuuti O., Alapaholuoma T., **Internet traffic clustering using packet header information** / Joint International IMEKO TC1+ TC7+ TC13 Symposium, Jena, Germany, 2011, pp. 13-20

[3] Gerhard M., Sa L., Georg C., **Traffic Anomaly Detection Using K-Means Clustering** / In Proceedings of performance, reliability and dependability evaluation of communication networks and distributed systems, 4GI/ITG-Workshop MMBnet, Hamburg, Germany, 2007, pp. 25-33

[4] Ekola T., Laurikkala M., Lehto T., Koivisto H., **Network traffic analysis using clustering ants** / Proceedings. World Automation Congress, v. 17, Seville, Spain 2004, pp. 275-280

[5] Duo Liu, Chung-Horng Lung, Lambadaṅs I., Seddigh N. **Network traffic anomaly detection using clustering techniques and performance comparison** / Proceedings the 26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Canada, 2013, pp.1-4

[6] Shokri, R., Oroumchian F., Yazdani N., CluSID: a clustering scheme for intrusion detection improved by information theory / Proceedings of the 7th IEEE Malaysia International Conference on Communications and IEEE International Conference in Networks, Kuala Lumpur, Malasia, 2005, pp.553-55

[7] http://wiki.squid-cache.org/SquidFaq/SquidLogs

[8] http://www.squid-cache.org/Intro/why.html

[9] http://www.bigml.com

Проблемы Визуализации Многомерных Данных

Айтадж Агабейли
Институт Информационных Технологий НАН Азербайджана, Баку, Азербайджан
agabeyli96@gmail.com

Аннотация **– в этой статье описывается термин Big Data в аспектах представления и визуализации данных. Существуют определенные специфические проблемы визуализации больших данных, мы постарались определить эти проблемы и совокупность подходов, помогающих их избежать. Кроме того, сделан обзор существующих методов визуализации применительно к многомерным данным и с учетом описанных проблем. А также была представлена классификация методов визуализации для работы с большими данными.**

Ключевые слова – большие данные, 2D, 3D, визуализация, график, диаграмма, представление.

I. ВВЕДЕНИЕ

Обработка больших данных не тривиальная задача и требует специальных методов и подходов. Графическое мышление очень простой и естественный способ обработки данных для человека, изображение является эффективным методом представления, который позволяет объективно оценить результаты и помогает в принятии правильного решения задач. Но в случае с большими данными множество из классических методов представления становится менее эффективным или даже неприменимым для конкретных задач. Необходимо классифицировать существующие методы визуализации по критерию их применимости к тому или иному классу больших данных.

Для принятия решения и сортировки описанных классов многомерных данных необходимо проанализировать следующие характеристики: применимость для большого объема данных, возможность визуализации данных, представленных в разных форматах, скорость и производительность представления данных.

Визуализация информации – не такая новая область. Первым «визуализатором», по мнению автора [1], стал математик и астроном Урбен Леверье, наиболее известный своим «открытием Нептуна на кончике пера».

Визуальная передача информации известна человеку с тех пор, как мы начали рассказывать друг другу истории. Визуализацией является любая техника создания изображений, диаграмм, карт, таблиц или анимаций. Визуальные образы были эффективным способом общения с древнейших времен. Примеры из истории включают наскальные рисунки, египетские иероглифы, греческую геометрию, революционные методы технического рисования Леонардо да Винчи для инженерных и научных целей.

Автор книги Handbook of Data Visualization Michael Friendly отметил основные опорные точки истории визуализации. Это ранние карты и диаграммы, измерения и теории, новые графические формы, начало современной графики, золотой век статистической графики, смутные годы, возрождение визуализации информации, интерактивная и динамическая визуализации [1].

II. ТИПЫ ВИЗУАЛИЗАЦИИ БОЛЬШИХ ДАННЫХ

В литературе [2−6] были представлены следующие типы визуализации:

A. Линейные графики (line charts)

Линейный график, или line chart, показывает отношение одной переменной к другой. Такие графики наиболее часто используются для отслеживания изменений, происходящих в течение долгого времени. Линейные графики также помогают при сравнении нескольких объектов за один и тот же промежуток времени. Уложенные линии используются для сравнения значений тенденции изменения индивидуальных значений или нескольких переменных. Линейные графики используются, когда необходимо визуально изобразить изменение одной или нескольких переменных, и при изображении тренда или скорости изменения информации о значениях этих переменных [2].

Важно также отметить, что не обязательно выбирать линейный график только потому, что имеется определенное количество точек данных. Для этого можно использовать более простое отображение, например, просто перечислить их в определенном порядке с использованием таблиц. Линейная диаграмма используется тогда, когда необходимо не просто изобразить, а показать взаимосвязь между точками данных.

B. Бар графики (bar charts)

Гистограммы обычно используются для сравнения количественных показателей различных категорий или группы данных. Значения категории представлены в виде баров, они могут быть сконфигурированы либо вертикально, либо горизонтально, а высота или длина полос соответствует определенным значениям данных.

Если эти значения достаточно разнятся и отличия в барах могут быть обнаружены человеческим глазом, можно использовать простую диаграмму. Когда значения баров очень близки друг к другу, становится трудно сравнивать стержни между собой. Чтобы помочь обеспечить визуальный контраст, бары могут иметь различные цвета.

Другой формой гистограммы является прогрессивная диаграмма или waterfall (водопад) график. Водопад-график показывает, как начальное значение данных увеличивается или уменьшается в течение ряда операций или действий. Первый бар начинается в начальном значении, и каждый последующий начинается там, где заканчивается предыдущий бар. Длина и направление бара указывают величину и тип (положительный или отрицательный) действий. Каскад, образующийся в результате диаграммы, покажет, как данное действие или операция приведет к конечному результату [2].

C. Диаграмма рассеяния (scatter plot)

Диаграмма рассеяния (Scatter plot) - это тип математической диаграммы, использующий декартовые координаты для изображения набора данных обычно для двух переменных. Можно увеличить количество отображаемых переменных до трех в случаи, если точки имеют цветовую маркировку. С помощью этих диаграмм можно определить направление и линейность связи между переменными, представленными на диаграмме. С увеличением точек на диаграмме рассеяния увеличивается и степень корреляции.

Как отмечено в публикации [3], диаграммы рассеяния (scatter plots) обычно используются для визуализации многомерных данных. Однако 2D предоставления данных при большом количестве взаимодействий между ними имеют трудности в восприятии. Введенные диаграммы рассеяния, 3D расширений в визуализацию интерактивных данных называются регрессивными кубами (RC). Они значительно увеличивают 3D-диаграммы с тремя аспектами, на которых показаны корреляции между двумя переменными с помощью чувствительных линий и чувствительных потоковых линий (stream lines). Авторы статьи продемонстрировали свою систему двумя примерами и оценкой пользователей, а также показали, как регрессивные кубы дают возможность интерактивного визуального исследования многомерных наборов данных через различные классификации и задачи информационного поиска.

D. Пузырьковая диаграмма (bubble plots)

Пузырьковая диаграмма — это разновидность точечной диаграммы, в которых точки изображены в виде пузырьков, и их размер показывает дополнительные данные. Пузырьковая диаграмма так же, как и точечная, имеют оси значения, которыми являются горизонтальная и вертикальная оси. Значение Z является дополнением к значениям X и Y [2].

Часто пузырьковые диаграммы используют для визуализации финансовых данных. Отличие в размерах пузырьков помогает визуально выделить конкретные значения.

E. Круговые диаграммы (pie charts)

Этот тип диаграмм позволяет графически представить данные как сегменты круга или процентные доли от целого значения. С помощью круговых диаграмм возможно отображение только одного ряда данных, и они показывают отношение размера элементов, образующих ряд, к их сумме. Такие диаграммы используются для изображения составных частей одного целого в виде сегментов круга.

Существует много споров вокруг ценности круговых диаграмм. Могут возникнуть трудности, связанные с интерпретацией оценки результатов, т.к. человеческому глазу сложно сравнивать ломтики круга, близкие по размеру и расположенные не рядом друг с другом. Определенную сложность в оценке и сравнении создает различие углов обзора отдельных сегментов.

В статье [4] показано, что использование круговой диаграммы целесообразно в указанных ниже случаях:

- Нужно отобразить только один набор данных.
- Ни одно из значений, которое нужно отобразить, не является отрицательным.
- Ни одно из значений, которое нужно отобразить, не является нулевым (0).
- Число категорий небольшое.
- Категории представляют части целого круга.

F. Плоское дерево (tree map)

Этот метод показывает весь набор данных в виде элементов, которые являются составляющими иерархического дерева. Они изображаются в виде набора прямоугольников – ветвей дерева, внутри которых находятся дочерние ветви. Прямоугольники отличаются по размеру и цвету соответственно заданным параметрам. Tree map четко показывает отношения данных, но представленных только в определенный момент времени. Например, детальная структура бюджета компании, в котором цветом показан процент изменения каждого пункта по сравнению с предыдущим годом.

G. Секторные диаграммы (sunburst)

Секторные диаграммы (также известные как treepie) являются относительно новым способом представления древовидных структур, часто позиционируемым в качестве альтернативы «плоскому дереву» treemap [5].

Основным отличием между этими методами является то, что переменные параметры отмечаются не как ширина и высота, а как радиус и длина дуги. И это позволяет не

менять всю схему при изменении данных, а только один сектор, содержащий новые данные путем изменения его радиуса. И поэтому этот метод может быть использован для отображения динамики изменения данных, используя анимацию [6].

H. Карты (map)

Часто используемые виды карт это: географическая карта, фотографическая карта, дорожная карта и тематическая карта. Географические карты используются для схематичного изображения географических объектов. Фотографические карты являются фотографиями географических объектов со спутника. На дорожных картах показываются схемы магистралей, трасс, железных и других дорог. На тематических картах объекты изображаются в виде различных маркеров.

III. ПРОБЛЕМЫ ВИЗУАЛИЗАЦИИ БОЛЬШИХ ДАННЫХ

Учитывая свойства больших данных, в [7] были определены следующие проблемы их визуализации:

1. Visual noise / визуальный шум;
2. Large image perception / восприятие большого изображения;
3. Information loss / потеря информации;
4. High performance requirements / высокие требования производительности;
5. High rate of image change / высокая скорость изменения изображения.

1) Визуальный шум

Простая презентация целого ряда данных может создать полный беспорядок на экране, и мы увидим только одно большое пятно, состоящее из точек, представляющих каждую строку данных. Эта проблема состоит в том, что большинство объектов в наборе данных слишком связано друг с другом и на экране наблюдатель не может разделить их в виде отдельных объектов. Так, иногда, анализируя, сложно получить даже немного полезной информации от всей визуализации данных без какой-либо дополнительной обработки. Следует отметить, что в понятие «визуальный шум» не входит любое повреждение или искажение данных, его следует рассматривать как явление потери видимости.

2) Восприятие большого изображения

Следующей проблемой визуализации больших данных является ограничение восприятия слишком крупного изображения. Существует определенный уровень восприятия человеческим мозгом различных визуальных данных. Несмотря на то, что этот уровень для графической визуализации данных значительно выше, по сравнению с визуализацией данных таблицы он имеет свои ограничения. И после перехода этого уровня восприятия человек просто теряет способность приобретать любую дополнительную информацию из перегруженных визуально данных. Все методы визуализации ограничены разрешением технического устройства, которое отвечает за вывод этих данных. Конечно, мы можем заменить устройства на более современные или на группу устройств для частичной визуализации данных, что позволит нам представить более подробное изображение с большим количеством точек данных, но даже если бы мы могли повторить этот процесс бесконечное число раз, мы встретились бы с ограничением восприятия человека. С ростом объема данных, показанных одномоментно, человек сталкивается с трудностями в понимании и анализе этих данных. Таким образом, можно сказать, что методы визуализации данных ограничены не только соотношением и разрешением устройств, но и физическими пределами восприятия.

3) Потеря информации

В связи с вышеизложенным применяются подходы, которые в конечном счете приводят к уменьшению использованных видимых наборов данных. Но, несмотря на решение предыдущего препятствия, эти подходы приводят к появлению другой проблемы, которой является потеря определенного количества информации. Все методы уменьшения визуальной информации производят агрегацию и фильтрацию данных на основе родства объектов в конкретном наборе данных по одному или нескольким критериям. Использование этих подходов может ввести в заблуждение аналитика, который может не заметить некоторые интересные скрытые объекты, а сложный процесс агрегации может потребовать большего количества времени и ресурсов для того, чтобы получить точную и необходимую информацию.

4) Высокие требования производительности

Графический анализ не ограничивается только статической визуализацией изображения, а использует и динамическую визуализацию. Здесь может появиться еще одна проблема, не заметная при статической визуализации. При наличии определенной скорости визуализации появляются требования и к производительности процесса. Процесс анализа определенных данных может занимать много времени при непрерывном увеличении вычислительных ресурсов для фильтрации все большего и большего количества данных.

5) Высокая скорость изменения изображения

Последняя проблема связана с высокой скоростью изменения изображения. Она становится наиболее значимой в процессе мониторинга, когда человек, наблюдающий данные, просто не может реагировать на скорость изменения данных или их интенсивности на дисплее. Снижение скорости меняющихся данных не может обеспечить желаемую эффективность процесса, но скорость реакции человека накладывает определенные ограничения на этот процесс.

В [8] с помощью многих исследований были продемонстрированы преимущества использования прогрессивных иллюминационных моделей в объемных визуализациях. Интерактивная объемная визуализация, делающая этот процесс более эффективным, в качестве инструмента для исследования 3D-данных, включая прогрессивное освещение, была достигнута с ускорением GPU для регулярной решетки объемных данных. В статье [8] представлена интерактивная стратегия иллюминации, которая специально разработана и оптимизирована для объемной визуализации данных неструктурированной решетки. Основой конструкции является дифференциальное уравнение, основывающееся на модели освещения, для того чтобы имитировать распространение, поглощение и рассеяние света в объемной среде. В частности, двухуровневая схема вводится для решения проблем, возникающих в неструктурированных решетках. Результаты испытаний показывают, что добавленные иллюминационные эффекты, такие как глобальная слежка и многократное рассеивание, могут не только привести к визуально более привлекательному представлению, но и существенно повысить восприятие глубины информации и сложных пространственных отношений для людей. усовершенствование объемной визуализации вводится в то время, когда неструктурированные решетки используются в различных научно симулированных приложениях.

ЗАКЛЮЧЕНИЕ

В этой статье мы описали основные проблемы визуализации многомерных данных и проанализировали причины их возникновения. Будущие работы в этой области могут быть проведены по следующим направлениям: исследование методов визуализации и области их применения; принятие решений и рекомендации по выбору методов визуализации для конкретных классов больших данных; формализация требований и ограничений на методы визуализации, применяющиеся к одному или более классам больших данных.

ЛИТЕРАТУРА

[1] M. Friendly, A Brief History of Data Visualization, March 2006, York University.

[2] SAS Institute, Data Visualization Techniques,
http://smartest-it.com/sites/default/files/Data%20Visualization_SAS.pdf.

[3] Y.-H.Chan, C.D.Correa and K.-L.Ma, Regression Cube: A Technique for Multidimensional Visual Exploration and Interactive Pattern Finding, ACM Transactions on Interactive Intelligent Systems (TiiS), April 2014, vol. 4, no. 7.

[4] Cleveland W.S. and R.McGill, Theory, experimentation, and application to the development of graphical methods, Journal of the American Statistical Association, 1984, vol. 79, no. 387.

[5] M.Tennekes and de E.Jonge, Top-down data analysis with treemaps, Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP '11), March 2011, pp. 236–241.

[6] Treemap Visualizations for Analyzing Multi-Dimensional, Hierarchical Data Sets, Panopticon Software,
http://panopticon.com/images/stories/white_papers/wp_treemap_data_visualizations_for_multi-dimensional_data.pdf.

[7] E.Gorodov and V.Gubarev, Analytical Review of Data Visualization Methods in Application to Big Data, Journal of Electrical and Computer Engineering, 2013, pp.7
http://www.hindawi.com/journals/jece/2013/969458/

[8] M.Shih, Y.Zhang and K.-L.Ma, Advanced lighting for unstructured-grid data visualization, IEEE Pacific Visualization Symposium (PacificVis), April 2015, pp. 239-246.

[9] C.Johnson, Top Scientific Visualization Research Problems, IEEE Computer Graphics and Applications, July 2004, vol.24,
http://dl.acm.org/citation.cfm?id=1018051.

[10] N.Cawthon and A.V.Moere, The Effect of Aesthetic on the Usability of Data Visualization,
http://web.arch.usyd.edu.au/~andrew/publications/iv07b.pdf.

[11] J.Heer and B.Shneiderman, Interactive dynamics for visual analysis, Communications of the ACM, 2012, vol. 55, no. 4.

[12] J.Stasko, Visualization for Information Exploration and Analysis, Proceedings of the 4th ACM symposium on Software visualization, 2008, pp. 7-8.

[13] D.Selassie, B.Heller and J.Heer, Divided edge bundling for directional network data, IEEE Transactions on Visualization and Computer Graphics, 2011, vol. 17, no. 12, pp. 2354–2363.

[14] J.Tedesco, A.Sharma and R.Dudko, Theius: a streaming visualization suite for hadoop clusters, Proceedings of the IEEE International Conference on Cloud Engineering, 2013.

[15] T.-Y.Lee, C.Jones, B.-Y.Chen, and K.-L. Ma, Visualizing data trend and relation for exploring knowledge, Proceedings of the IEEE Pacific Visualization Poster, 2010.

[16] P.Zikopoulos, C.Eaton, D. deRoos, T.Deutsch and G.Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2012.

[17] C.Ahlberg and B.Shneiderman, Visual information seeking: tight coupling of dynamic query filters with starfield displays, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '94), April 1994, pp. 313–317.

E-Dövlətin Big Data Mənbələri

Ramiz Alıquliyev[1], Günay Niftəliyeva[2]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*r.aliguliyev@gmail.com*, [2]*gunayniftali@gmail.com*

***Xülasə* — Big data e-dövlət xidmətlərinə transformasiyada, dövlət, vətəndaşlar və biznes sektoru arasında qarşılıqlı əlaqələrdə mühüm əhəmiyyət kəsb edir. Big data e-dövlətdə bir sıra sahələrdə, məsələn: kənd təsərrüfatı, təhsil, səhiyyə və s. problemlərin həllinə köməklik göstərərək qərarların qəbul olunmasında yeni mərhələ açır. Bu məqalədə e-dövlətin big data mənbələri araşdırılmış və big data analitikasının verdiyi imkanlar nəzərdən keçirilmişdir.**

Açar sözlər — big data; e-dövlət; dövlət sektoru.

I. GİRİŞ

Big data ənənəvi relyasion verilənlər bazası üsulları vasitəsilə analizi kifayət qədər mürəkkəb olan strukturlaşdırılmış/struktursuz, olduqca böyük, müxtəlif mənbələrdən əldə olunan rəqəmsal məlumatların böyük həcminə verilən addır. Hər gün internetdə əlavə edilən məlumatların 90%-ni struktursuz məlumatlar təşkil edir. Big data bu məlumatların analizində mühüm vasitələrdən biridir. Big data istənilən təşkilatda qərarların dəstəklənməsinə kömək göstərərək biliyin aşkarlanması, dəyərin yaradılması və zəngin biznes analitikası üçün yeni imkanlar təklif edir.

Big data rəqəmsal dövrdə dövlətin qarşısında duran yeni və mühüm məsələlərdən biridir. Biznes sektoru big data tətbiqlərinin inkişafında öndə olsa da, dövlət sektorunda da buna çox böyük ehtiyac yaranmışdır. İnsanların elektron savadlılığının artması ilə əlaqədar olaraq e-dövlət xidmətlərindən istifadəyə ehtiyac artmışdır. E-dövlətdə veb, bioloji və sənaye sensorları, video, e-poçt və sosial media daxil olmaqla bir çox mənbələrdən əldə olunan, sürətlə böyüməkdə olan informasiyadan real vaxtda qərarların qəbulunun dəstəklənməsində istifadə oluna bilər [1]. Bu məlumat axını e-dövlətdə big data mənbələrinin yaranmasına gətirib çıxarır. Bir sıra tədqiqatlarda dövlətin vətəndaşlara xidmət göstərməsi və milli problemlərin (səhiyyə xərclərinin artması, iş yerlərinin yaradılması, təbii fəlakətlər və terrorizm) aradan qaldırılmasında big data-dan istifadə yolları təklif olunmuşdur [2, 3].

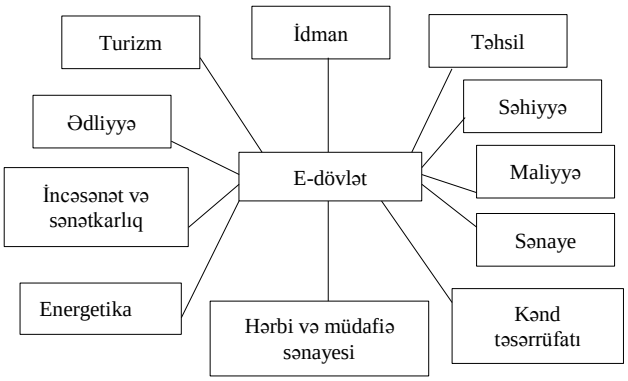
Big data öz həcminə, sürətinə, müxtəlifliyinə və təbiətinə görə ənənəvi data-dan (verilənlər, məlumat) fərqlənir. Dövlətə gələn məlumatların artma sürəti göz qabağındadır. Məsələn, e-hökumət saytlarında vətəndaşlardan real vaxt rejimində bir çox məlumatlar əldə olunur. Təhlükəsizlik kameralarındakı video görüntülər, kosmik kəşfiyyatdan alınan təsvirlər, telekommunikasiya log verilənləri, veb əməliyyat verilənləri və sosial mediada qarşılıqlı əlaqə məlumatları müxtəliflik yaradır.

Məlumatların bu böyük həcmindən dəyərin (biliyin) yaradılması, şəffaflığın artırılması, qərarların dəstəklənməsi və siyasətin formalaşmasında istifadə olunur. Böyük həcmli və açıq məlumatların “ağıllı” kombinasiyası sosial innovasiyalar və kommersiya xidmətlərinin inkişafına imkan verir. Bu hökumət və digər tərəflərin mövqelərinin dəyişməsinə gətirib çıxarır və cəmiyyətdə problemlərin aradan qaldırılması üçün çoxlu həllər irəli sürür. Big datanın artması hökumət üçün həm imkanlar, həm də bir sıra problemlər yaradır. Big datanın yaratdığı imkanlara bunlar daxildir: onlayn məlumatda əhəmiyyətli inkişaf üçün biznes analitikasının yaradılması, hökumət tərəfindən xidmətlərin göstərilməsi, adaptiv və xüsusi e-dövlət təcrübələri üçün real vaxt informasiyasından istifadə, dinamik və birgə dövlət siyasəti qərarlarının qəbul edilməsi üçün dövlət fəaliyyətinin izlənilməsi və vizuallaşdırılması, biznes proseslərinin yenidən təşkili və dövlətin transformasiyası üçün bilik istehsalı.

Big data öz növbəsində informasiyanın keyfiyyəti, siyasi təsirlər, maraqlı tərəflər arasında dəyişən münasibətlər ilə əlaqədar sosial və texniki problemlər də gətirir. Dövlət üçün big data problemləri institusional və texniki xarakter daşıyır. İnstitusional problemlərə səmərəli şəkildə bəzi əsas məsələlərin həlli üçün idarəetmə strukturunun, həmçinin informasiya əlaqələri üçün ümumi verilənlər standartlarının, məlumat paylaşan, məlumat mübadiləsi edən vətəndaşların etibarını qazanmaq üçün məxfilik zəmanətlərinin, fərdiləşdirmə üçün təşkilatlar arasında əlaqə sazişlərinin hazırlanması daxildir. Texniki problemlərə müvafiq qabiliyyətin çatışmazlığı, müvafiq proqram alətlərinin yetərincə inkişaf etməməsi, bir çox məlumat mənbələrinin və formatlarının inteqrasiyası, məlumatların saxlanılması və giriş daxildir [2].

II. E-DÖVLƏT TƏTBİQLƏRİ VƏ BİG DATA

E-dövlət dövlət xidmətlərinin daha yaxşı çatdırılması, biznes və sənaye müəssisələri arasında qarşılıqlı əlaqələrin inkişafı, informasiyaya çıxış vasitəsilə vətəndaş səlahiyyətlərinin genişləndirilməsi və ya daha səmərəli dövlət idarəetməsini nəzərdə tutur. Bu daxili əməliyyatları sadələşdirir və maksimum rahatlıq, az xərclə vətəndaşların bütün təbəqələrinin dövlət xidmətlərindən faydalanmasına kömək edərək dövlət idarələrinin fəaliyyətini yaxşılaşdırır. E-dövlətin əsas məqsədi vətəndaşların demokratiyada iştirakını təmin etmək və dövlət xidmətlərinin bütün növlərinin səmərəliliyinin və effektivliyinin artırılmasından ibarətdir. E-dövlətin əsas tətbiq sahələri şəkil 1-də təsvir olunmuşdur.



Şəkil 1. E-dövlət tətbiqləri

E-dövlət mühitində bir çox big data mənbələri mövcuddur. Big data e-dövlətə səhiyyə, təhsil, milli təhlükəsizlik, hüquq-mühafizə orqanlarında və digər sahələrdə öz öhdəliklərini yerinə yetirməyə kömək edir. Big data texnologiyaları vasitəsilə e-dövlət mühitində verilənlərin toplanması və analizi ilə maliyyə göstəricilərini yaxşılaşdırmaq, gəlirləri və xərcləri optimal idarə etmək və şəffaflığı yüksəltmək mümkündür. Dövlət təşkilatları daxilində mühüm qərarların qəbul edilməsində, ictimai asayişin qorunmasında, sosial təminat, milli təhlükəsizlik, səhiyyə məsələlərində, istehlakçıların alıcılıq qabiliyyətini öyrənməklə marketinq işlərinin yaxşılaşdırılmasında, insanların gizli davranışlarını üzə çıxarmaqda big data texnologiyalarından istifadə oluna bilər [4, 5].

Big data və səhiyyə

Big data səhiyyədə çox böyük rol oynayır və onun sayəsində milyonlarla vəsaitə qənaət oluna bilər. Son bir neçə il ərzində bütün tibbi qeydiyyat məlumatlarına giriş imkanının əldə olunması bu sahədə dövlətə düzgün qərarların qəbuluna kömək edə bilər. İstifadəçi məlumatları düzgün qaydada emal və təhlil olunarsa, ölkənin hansı hissəsində hansı xəstəliklərin yayılması və səhiyyə ilə bağlı hansı problemlərin olduğunu müəyyənləşdirmək olar. Bu dövlətə yeni xidmətlərin göstərilməsində kömək edə bilər və həmçinin xəstələrə xidmət üçün yeni xəstəxanaların açılmasına gətirib çıxara bilər. Bundan əlavə lazımi dərman vasitələrinin tədqiqinə daha çox maliyyə ayırmaqla səhiyyəyə köməklik göstərə bilər [6].

Big data və təhsil

Texnologiya əsaslı təhsil alətləri və platformaları tələbə və müəllimlərə əhəmiyyətli yeni imkanlar təklif edir. Təhsil texnologiyalarının miqyası tələbələr və daha geniş auditoriya arasında artmaqdadır. Beləki, hal–hazırda tələbələr onlayn şəkildə sinif materiallarını əldə edə, təlim videolarına baxa, sinif fəaliyyəti ilə bağlı şərh yaza, bir-birilə əməkdaşlıq edə, ev tapşırıqlarını tamamlaya və testlər edə bilərlər. Bu isə tələbələrlə bağlı böyük həcmli məlumatların, yəni big datanın yaranmasına gətirib çıxarır. Təhsil sektorunda big data mühüm rol oynayır. Tələbələr haqqında müxtəlif məlumatlar onların maraq dairələrini, zəif və güclü tərəflərini müəyyən etməyə kömək edə bilər. Big data şagirdlərin fəaliyyəti və təlimlə bağlı informasiyaları emal etməyə kömək edir. Big data vasitəsilə müəllimlər şagirdlərin biliyini analiz edə və hər bir şagird üçün daha səmərəli metoddan istifadə edə bilərlər. Məlumatların analitikasına diqqət yetirərək müəllimlər daha fərqli yollarla (nüanslarla) təlimi həyata keçirə bilərlər. Onlayn vasitələr tələbələrin fəaliyyətinin daha geniş qiymətləndirilməsinə imkan yaradır, məsələn: elektron resurslar vasitəsilə tələbələrin oxumağa və əsas anlayışları mənimsəməyə nə qədər vaxt sərf etdiyini müəyyən etmək olar. Big data vasitəsilə məlumatların analizi müəlliflərə kitabların təkmilləşməsinə, təhsil işçilərinə bu statistika əsasında düzgün qərarların qəbuluna kömək edə bilər [6, 7].

Big data və hüquq-mühafizə

Hüquq-mühafizə və təhlükəsizlik orqanlarında müntəzəm olaraq cinayətkarlığın aşkarlanması və kəşfiyyat zamanı böyük həcmli məlumatlar toplanır. Elektron əlaqə vasitələrinin inkişafı insanlar və onların fəaliyyəti haqqında çox geniş həcmdə və müxtəlif tipdə məlumatların toplanmasına səbəb olmuşdur. Bu böyük və mürəkkəb həcmli məlumatlar big data yaradır. Big data hüquq-mühafizə orqanlarında güclü vasitə ola bilər. Big data texnologiyaları vətəndaş azadlıqlarının və vətəndaşların qanuni şəxsi maraqlarını həyata keçirmək və ictimai asayişin qorunması üçün istifadə oluna bilər [8].

Big data və maliyyə xidmətləri

Bank və maliyyə xidmətləri sahəsində bir çox məlumatlar toplanır. Belə ki, müştərilərin bank ilə qarşılıqlı əlaqəsi zamanı potensial biznes dəyərli məlumatlar istehsal olunur. İKT-in maliyyə xidmətlərində tətbiqləri böyük məlumatların, yəni big data-nın yaranmasına şərait yaradır. Mobil və rəqəmsal ödəniş zamanı istifadəçilər və müştərilər haqqında ətraflı məlumatları əldə etmək mümkündür. Belə məlumatlar müştərilər üçün xüsusi maliyyə xidmətlərinin yaradılmasında istifadə oluna bilər. Vergi şöbələrində isə böyük həcmli məlumatların yalnız kiçik hissəsindən istifadə etmək lazım gəlir. Həddən artıq çox məlumatları analiz etmək mümkün olmaya bilər. Böyük həcmli məlumatlardan lazımi biliyin çıxarılmasında big data texnologiyasından istifadə oluna bilər [9].

Big data və kənd təsərrüfatı

Kənd təsərrüfatı sahəsində də bir sıra big data mənbələri mövcuddur. Bunlara torpaq temperaturu xəritələri, yağış xəritələri, elektrik keçiriciliyi, rütubətin saxlanması və hava keçiricilik xəritələri, biogen maddələrin tərkibi, son becərmə uçotu (qeydləri), siğorta və gəlirlə bağlı məlumatlar, kənd təsərrüfatı jurnallarında məqalələr və həmçinin sosial media şərhləri daxildir. Big data texnologiyaları vasitəsilə müvafiq lazımi amillərin diaqnostistikası ilə aşağı xərclərlə məhsuldarlığı artırmaq, kənd təsərrüfatı sistemini təkmilləşdirmək üçün tendensiyaları anlamaq, əlaqələri aşkarlamaq mümkündür [10].

NƏTİCƏ

Bir çox dövlət qurumunda müxtəlif mənbələrdən müxtəlif təbiətli məlumatlar toplanır. Bu məlumatların həcminin çox böyük sürətlə artması müvafiq idarəetmə strukturunun həyata keçirilməsində big datanın idarə olunması, müxtəlif məlumat mənbələrinin inteqrasiyası, elektron gizlilik və təhlükəsizlik risklərinin idarə olunması kimi problemlər yaradır. Big data texnologiyaları bu problemlərin aradan qaldırılmasında əhəmiyyətli rol oynaya bilər.

ƏDƏBİYYAT

[1] G.-H. Kim, S. Trimi and J.-H. Chung, “Big-Data Applications in the Government Sector”, Journal for Communications of the ACM, Vol.57, No.3, pp. 78-85, 2014.

[2] Y.-Ch. Chen, T.-Ch. Hsieh, “Big Data for Digital Government: Opportunities, Challenges, and Strategies”, International Journal of Public Administration in the Digital Age, Vol. 1, No.1, pp.1-14, 2014.

[3] J. Šuh, V. Vujin, D. Barać, Z. Bogdanović, B. Radenković, “Designing Cloud İnfrastructure for Big Data in E-Government”, Journal of Universal Excellence, Vol. 4, No. 1, pp. 26-38, 2014.

[4] V. Morabito, “Big Data and Analytics for Government Innovation”, Big Data and Analytics, pp. 23-45, 2015.

[5] R.M. Əliquliyev , M. Ş. Hacırəhimova, “Big Data Fenomeni: Problemlər və İmkanlar”, İnformasiya texnologiyaları problemləri, №2, səh. 3-16, 2014.

[6] S. Kalbande, S. Deshpande and M. Popat, “Review Paper on Use of Big Data in E-Governance of India”, İnternational Journal for Research in Emerging Science and Technology, Vol.2, No. 1, pp.395-403, 2015.

[7] V. Kellen, C. Consortium, A. Recktenwald and S. Burr, “Applying Big Data in Higher Education: A Case Study”, Data Insight & Social BI Executive Report, Vol. 13, No. 8, pp.1-38, 2013.

[8] http://images.politico.com/global/2014/05/01/big_data_privacy_report_may_1_2014.html

[9] P J Godsiff, C.E.A. Mulligan, Z. Gurguc, “ICT & The Future of Financial Services”, Industry Transformation – Horizon Scan, pp.17-18, 2014.

[10] R. Yadav, J. Rathod, V. Nair, “Big Data Meets Small Sensors in Precision Agriculture”, Proceedings of the National Conference Potential Research Avenues and Future Opportunities in Electrical and Instrumentation Engineering, pp. 1-4, 2015.

E-dövlət Sisteminin Analizində Data Mining Texnologiyalarının Tətbiq İmkanları

Ramiz Alıquliyev[1], Günay Niftəliyeva[2]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*r.aliguliyev@gmail.com;* [2]*gunayniftali@gmail.com*

***Xülasə* — Big data dövlətin qarşısında duran yeni və mühüm məsələlərdən biridir. Big data analitikləri e-dövlətin effektiv və səmərəli idarə olunmasında text və data mining metodlarının mühüm rol oynadığını qeyd edirlər. Bunu nəzərə alaraq, məqalədə bu sahədə müxtəlif tədqiqatçılar tərəfindən görülən işlər nəzərdən keçirilmiş, text və data mining texnologiyalarının rolu müəyyən olunmuşdur.**

Açar sözlər — e-dövlət; big data; data mining; text mining.

I. GİRİŞ

Elektron dövlət dörd dəyişənin funksiyası kimi müəyyən edilə bilər: idarəetmə, informasiya və kommunikasiya texnologiyaları (İKT), biznes proseslərinin yenidən təşkili və e-vətəndaş. Eləcədə buraya, rəqəmsal media vasitəsilə vətəndaşlara, biznes və ya digər dövlət orqanlarına milli və dövlət informasiya və xidmətlərinin göstərilməsi daxildir [1]. E-dövlət quruculuğu dövlətlə vətəndaş arasında informasiya mübadiləsi və öhdəliklərin qarşılıqlı olaraq "rəqəmsal mühitdə" davamlı, etibarlı, şəffaf bir şəkildə reallaşdırılmasını ifadə edir. İKT-nin sürətli inkişafı e-dövlətin inkişafına kömək edir. Vətəndaşların gündəlik həyatında İKT-dən istifadənin artması bu sahədə e-dövlət xidmətlərinin genişlənməsinə səbəb olmuşdur. Ədəbiyyatlarda "E-dövlət sistemləri dövlətə inam, sosial inteqrasiya, ictimai rifah kimi siyasi və sosial vəzifələri yerinə yetirmək üçün effektivlik, səmərəlilik və iqtisadiyyat həddini aşan strateji məqsədləri əhatə edir" kimi diqqətə çatdırılır [2].

Big data, bulud hesablamalar e-dövlətin son inkişaf meylləri kimi göstərilir. Big data müəyyən zaman aralığında müxtəlif mənbələrdən toplanan faydalı və istifadə edilə bilən elektron məlumatların böyük həcmidir. Mənbələrdə qeyd olunur ki, "dünya məlumatlarının 90%-i son iki il ərzində toplanmışdır və hər gün 2.5 kvintilyon bayt məlumat əlavə edilir" [2]. Bu məlumatların böyük həcmi öz strukturuna görə relyasion verilənlər bazasında istifadə edilə bilməz. Verilənlər yeni dəyər kəsb edən qiymətli mənbələrdir. Dövlətin vətəndaşlara xidmət göstərməsi, səhiyyə xərcləri, işsizlik, təbii fəlakətlər və terrorizmin artması kimi milli problemlərin aradan qaldırılmasında big data-dan istifadə oluna bilər. Bu yaxınlarda, B. Obama "rəqəmsal məlumatların böyük, mürəkkəb kolleksiyalarından bilik, məlumat əldə etmək imkanlarının yaxşılaşdırılması; elm və texnikanın kəşf tempini sürətləndirmək üçün bu texnologiyalardan istifadə edilməsi; milli təhlükəsizliyin möhkəmləndirilməsi; tədris və təlimi transformasiya etmək üçün" Big Data Tədqiqat və İnkişaf Təşəbbüsünə (Big Data Research and Development Initiative) başlamışdır [2]. Big data bir sıra digər texnologiyaların tətbiqi ilə yanaşı, e-dövlətdən istifadəni genişləndirməyə kömək edə bilər. Verilənlər eksponensial sürətlə artdığından ənənəvi məlumat (data) emalı vasitələri və texnologiyaları bu məlumatların analizində faydalı ola bilmir. Big data analitikləri bunun text və data mining texnologiyaları vasitəsilə mümkün olduğunu qeyd edirlər [1, 3].

II. ƏLAQƏLİ İŞLƏRİN İCMALI

E-dövlətdə data mining - dövlət veb saytlarında verilənlərin (məlumatların) qərar qəbuletmədə istifadə oluna biləcək faydalı biliyə çevrilməsi prosesidir. E-dövlətdə text və data mining texnologiyalarının tətbiqi ilə bağlı bir sıra tədqiqatlar aparılmışdır.

[4]-də mətn məlumat mənbələrindən bilik və əlaqələrin tapılmasında text mining texnologiyasından istifadə olunmasının faydaları müzakirə olunmuşdur. Bu məqalədə dövlət siyasəti və vətəndaş rəyləri arasında əlaqəni aşkar etmək üçün veb forum və bloqlarda text mining metodlarından istifadə olunması və vətəndaş xidmətlərinin bununla təkmilləşdirilə biləcəyi təklif olunur. Bundan başqa, çoxdilli text mining, maşın tərcümə (machine translation), ikidilli text mining-in texniki arxitekturası və s. kimi mühüm məqamlar müzakirə olunur.

[5]-də dövlət qərarlarının qəbul olunmasında daha yaxşı nəticələrin əldə olunması üçün data mining texnologiyasının rolu müzakirə olunur. Bu məqalədə e-dövlətin qurulmasında data mining metodu tətbiq olunur. İlk növbədə ictimai rəy araşdırılır, daha sonra SPSS və korrelyasiya analizi vasitəsilə toplanan məlumatlar emal olunur, insanların real ehtiyaclarına xas qayda tapılır və beləliklə dövlət qərarı üçün daha yaxşı dəstək təmin olunur.

[6]-da e-dövlətdə açıq məlumatların (verilənlərin) analizi ilə bağlı yanaşma təklif olunub. Bu yanaşmaya 4 məsələ daxildir: məlumatların toplanması və emalı, modelin aşkar olunması və modelin təhlili.

[7]-də elektron səsvermə üçün konseptual model və e-səsvermə vasitəsilə toplanmış məlumatlar üzərində data mining texnologiyasının müxtəlif üstünlükləri müzakirə olunmuşdur. Bu məqalədə data mining texnologiyasının e-səsvermə verilənlər bazasından biliyin aşkarlanmasında çox səmərəli olduğu göstərilmişdir.

E-dövlət üçün qeyri-səlis data mining alqoritmi [8]-də təklif olunur. Burada e-dövlət məsələləri sistemində faydalı məlumatı aşkarlamaq üçün qeyri-səlis data mining-dən istifadənin faydaları müzakirə olunur. Sonda bu qənaətə gəliblər ki, e-

dövlətdə qeyri-səlis data mining-dən istifadə siyasətçilərə böyük kömək ola bilər.

[9]-da e-dövlətdə data mining texnologiyasının tətbiqi üçün yanaşma təklif olunur. Bu məqalədə e-dövlətdə data mining texnologiyasının tətbiqi 2 kateqoriyaya bölünür. Birinci kateqoriya ümumi tətbiqdir və bura vətəndaşların identifikasiyası, insan resurslarının və layihələrin idarə olunması daxildir. İkinci kateqoriyaya kənd təsərrüfatı, səhiyyə, təhsil və s. daxildir. Bu tədqiqatlarda həmçinin dövlət təşkilatları daxilində qərarların qəbulu üçün data mining-dən istifadə olunaraq müxtəlif məsələlər, problemlər araşdırılmışdır. Sonda bu qənaətə gəlmişlər ki, data mining texnologiyası təkcə saxtakarlıq (dələduzluq) və təhlükəsizlik sahəsində təhdidləri aşkar etməyə kömək etmir, həmçinin dövlət-dövlət, dövlət-vətəndaş, dövlət-biznes kimi e-dövlət xidmətlərinin təkmilləşməsinə təsir edərək vətəndaşların davranış və istəkləri kimi təsiredici faktların ölçülməsində istifadə oluna bilər.

[10]-da vətəndaşların rifahı üçün e-dövlətin tətbiq və xidmət imkanlarının genişləndirilməsi üçün text və data mining texnologiyalarının tədqiqi təklif olunmuşdur. Bu məqalədə text və data mining texnologiyalarından ictimai rəyin modelləşdirilməsi və hökumət qərarlarının qiymətləndirilməsində istifadə olunur. Daha sonra ictimai rəyin istiqamətini əldə etmək və digər tərəfdən əlaqədar məsələlər üzrə yeni hökumət qərarlarının hazırlanması və çıxarılan rəylər arasında əsas korrelyasını müəyyən etmək üçün vətəndaş rəyləri analiz olunur.

E-dövlət xidmətlərinin təkmilləşdirilməsi üçün [11]-də hibrid yanaşma təklif olunur. Burada, e-dövlət sənədlərinin çoxsəviyyəli klasterləşməsi (multilayered clustering) məqsədilə müxtəlif mətn yaxınlıq ölçüləri və qeyri-səlis konsepsiyanın tətbiqi təklif olunur.

[12]-də veb forumlarda vətəndaşların mesaj və rəylərinin analizi vasitəsilə səsvermə tendensiyalarını müəyyən etmək üçün text mining texnologiyasına əsaslanan yanaşma təklif olunmuşdur.

Hökumət qərarlarına ictimaiyyətin mövqeyini müəyyən etmək üçün [13]-də yanaşma irəli sürülmüşdür. Bu yanaşmada text mining-dən istifadə edərək əvvəlcə dövlət qərarları haqqında vətəndaş rəyləri toplanır və daha sonra analiz olunur. Text mining metodları mətn verilənlərindən xüsusiyyətlərin çıxarılması nəticəsində yaradılan strukturlaşdırılmış verilənlər bazasında istifadə olunur. Burada mətn verilənləri K-Means alqoritmi vasitəsilə biliyin aşkarlanması üçün klasterləşdirilir. Strukturlaşdırılmış mətn verilənlərindən əldə olunan bilik e-dövlət üzrə strateji qərarların dəstəklənməsində istifadə oluna bilər.

III. E-DÖVLƏT SİSTEMİNDƏ TEXT VƏ DATA MİNİNG TEXNOLOGİYALARININ İSTİFADƏ İMKANLARI

E-dövlətdə İnformasiya Texnologiyalarının (İT) rolu elmə yaxşı məlumdur. Verilənlər bazasından biliyin praqmatik istifadəsi, biliyin idarə olunması texnologiyaları e-dövlət mühitində qərar qəbuletmə sistemlərinə böyük töhfə verə bilər. E-dövlətdə müxtəlif təbiətli (mətn, video, audio) məlumatlar toplanır. Aydındır ki, e-dövlətin effektiv idarə olunmasında bu məlumatların analizi mühüm əhəmiyyət kəsb edir. Bu məlumatların həcmi çox böyük sürətlə artır və bu e-dövlət mühitində böyük həcmli verilənlərin analizi probleminə gətirib çıxarır. Hal-hazırda verilənlərin effektiv analizində data mining texnologiyaları çox geniş tətbiq olunur. Data mining verilənlər bazasında biliyin aşkar olunmasında ən mühüm texnologiyalardan biridir və biliklərin idarə edilməsində əhəmiyyətli sahə hesab olunur. Data mining verilənlər bazasından gizli, amma faydalı bilyin aşkarlanmasını həyata keçirir. Verilənlər bazasından biliyin aşkarlanmasının formal tərifi aşağıdakı kimi verilir: Data mining verilənlər haqqında dolayısı ilə əvvəllər məlum olmayan və potensial faydalı informasiyanın çıxarışıdır. Aşkar edilmiş bilik e-dövlət sistemində xidmətin keyfiyyətinin yaxşılaşdırılmasında istifadə oluna bilər. Data mining metodları və bilik idarəetmə texnologiyaları birlikdə biliklə zəngin mühit yaratmaq üçün tətbiq edilə bilər. Verilənlər bazasında biliyin aşkarlanması böyük həcmli verilənlərlə işlədikdə strateji qərarları inkişaf etdirmək və modellərin müəyyən olunmasında effektiv ola bilər. [14, 15].

Qeyd edək ki, data mining metodları marketinq, maliyyə və səhiyyə sahəsində də uğurla tətbiq edilir. Dövlət cinayət və terror fəaliyyətinin, dələduzluq hallarının aşkar olunması, göstərilən xidmətlərin yaxşılaşdırılması, elmi məlumatların təhlili, insan resurslarının idarə olunmasında data mining-dən istifadə edə bilər. Təşkilatlar əhalinin ehtiyaclarını ödəmək üçün nəzərdə tutulumuş siyasətin planlaşdırılması, həyata keçirilməsi və qiymətləndirilməsi üçün data mining texnologiyasını tətbiq edir. Bütün bu sahələrdə planlaşdırma və qərar qəbuletmənin keyfiyyəti dəqiq obyektiv məlumatların toplanması, təhlili və interpretasiyasından asılıdır [16, 17].

E-dövlət mühitində məlumatların çox böyük sürətlə artdığını və bu məlumatların əksəriyyətinin mətn şəklində saxlanıldığını nəzərə alsaq, text mining metodlarının da data mining-lə yanaşı yüksək potensiala malik olduğunu qeyd edə bilərik. Text mining statistika, maşın təlim, məlumat-axtarış, data mining, linqvistika və təbii dil emalı anlayışlarını bir araya gətirən nisbətən yeni fənlərarası sahədir. Bu avtomatik olaraq müxtəlif yazılı mənbələrdən informasiyanın çıxarılması ilə yeni, əvvəllər məlum olmayan informasiyanın kompüter tərəfindən aşkar olunması prosesidir. Text mining bir sıra xüsusiyyətlərinə görə data mining-ə oxşardır. Data mining əsasən rəqəmli strukturlaşdırılmış məlumatlarla, text mining isə struktursuz məlumatlarla işlədikdə faydalıdır [4].

E-dövlətdə istifadə olunan text mining metodları bunlardır: informasiyanın axtarışı, informasiyanın çıxarılması, mövzunun izlənilməsi, referatlaşdırma, təsnifat, konsepsiya (anlayış) əlaqələri, informasiyanın vizuallaşdırılması və sualın cavablandırılması.

İnformasiyanın çıxarılması: İnformasiyanın çıxarılması alqoritmləri mətn daxilində əsas ifadələri və əlaqələri müəyyən edir. Bu "model uyğunluğu (şablon)" adlandırılan prosesdən istifadə edərək mətndə əvvəlcədən müəyyən edilmiş ardıcıllıqların axtarılması ilə həyata keçirilir. Alqoritmlər istifadəçini mənalı fikir ilə təmin etməkdən ötrü bütün müəyyən edilmiş ardıcıllıqlar arasında əlaqələri aşkarlayır. Bu metod böyük həcmli mətn ilə işlədikdə çox faydalıdır.

Təsnifat: Təsnifat sənədin verilmiş mövzular çoxluğuna yerləşdirilməsi ilə onun əsas mövzusunun müəyyən olunmasını nəzərdə tutur. Burada informasiyanın çıxarılması metodunda olduğu kimi aktual informasiyanın emal olunmasına cəhd göstərilmir. Təsnifat yalnız mətndə olan sözləri sayır və bu say sənədin əhatə olunduğu əsas mövzuları müəyyən edir. Təsnifat tez-tez mövzuların əvvəlcədən müəyyən olunduğu lüğətə (thesaurus) əsaslanır və əlaqələr geniş terminlər, dar terminlər, sinonimlər və müvafiq terminlərə görə müəyyən olunur.

Klasterləşmə: Klasterləşmə oxşar sənədləri qruplaşdırmaq üçün istifadə olunan metoddur. Burada təsnifatdan fərqli olaraq əvvəlcədən təyin olunmuş mövzulardan istifadə etmək əvəzinə sənədlərin bir-birinə oxşarlığı əsasında qruplaşdırma aparılır. Əsas klasterləşmə alqoritmi hər bir sənəd üçün mövzuların vektorunu yaradır və sənədin hər bir klasterə nə dərəcədə uyğun gəldiyini ölçür.

Mövzunun izlənilməsi: Mövzunun izlənilmə sistemi istifadəçi profillərini istifadə etməklə işləyir, istifadəçi tərəfindən baxılan sənədlərə əsaslanır və istifadəçidə maraq doğuran digər sənədləri proqnozlaşdırır. Text mining vasitələrindən bəziləri istifadəçilərə maraq doğuran xüsusi kateqoriyaları seçməyə imkan verir və istifadəçinin marağını onun oxuma tarixi və klik məlumatlar əsasında avtomatik olaraq müəyyən edir.

Referatlaşdırma: Mətnin əsas müddəalarını və ümumi mənasını saxlamaqla həcminin azaldılmasından ibarətdir.

Sualın cavablandırılması: Text mining-in digər sahələrindən biri sualın cavablandırılmasıdır. Burada o verilmiş suala ən yaxşı cavabın necə tapılması ilə məşğul olur. Sualın cavablandırılmasında bir çox text mining metodlarından istifadə oluna bilər.

Assosiasiyanın aşkarlanması: Assosiasiya qaydaları mövzular arasında və ya əlaqəli mətn çoxluğunu xarakterizə etməkdə istifadə olunan təsviri anlayışlar arasında olan əlaqələrin və təsirlərin öyrənilməsində istifadə olunur.

Qeyd etdiyimiz bu metodlar text miningin həll etdiyi əsas məsələlərdir. Big data analitikləri bu metodların e-dövlətdə böyük həcmli məlumatların analizində mühüm rol oynayacağını qeyd edirlər.

NƏTİCƏ

E-dövlət İKT-dən istifadə etməklə dövlət və vətəndaşlar arasında münasibətlərə təsir edərək dövlət sektorunda xidmətlərin yaxşılaşdırılması üçün dövlət idarəçiliyində yeni mərhələ hesab olunur. E-dövlətdə toplanan məlumatların çox böyük sürətlə artması bu sahədə bir sıra problemlər yaradır. Bu problemlərə elektron mühitdə eksponensial surətdə yaradılan informasiya axınını analiz etmək üçün səmərəli üsulların inkişaf etdirilməsi tələbatı daxildir. Bu məlumatlar yalnız ölçülərinə görə böyük deyil, həm də dinamik və heterogen xarakterlidir. Yuxarıda göstərilən problemləri nəzərə alaraq tədqiqatçılar tərəfindən bu verilənlərdən biliyin aşkarlanması üçün text və data mining texnologiyalarının istifadə olunmasına baxmayaraq e-dövlətin daha yaxşı həyata keçirilməsi üçün daha səmərəli və etibarlı üsulların inkişaf etdirilməsinə ehtiyac var. Text və data mining texnologiyalarının e-dövlət mühitində həyata keçirilməsində əsas problemlər aşağıdakılardır:

- ✓ Verilənlər bazasının dinamik təbiəti
- ✓ Eksponensial sürətdə yaradılan verilənlər bazasının böyük həcmi
- ✓ Verilənlər bazasının heterogen təbiəti
- ✓ Gizli verilənlər bazaları
- ✓ Atributların seçilməsi

Elektron məlumatların saxlanması problemləri böyük çətinliklər yaradır. Big data analitikanın inkişafı bu problemin öhdəsindən gələ bilər və e-dövlət layihələrinin səmərəli həyata keçirilməsinə kömək edə bilər.

ƏDƏBİYYAT

[1] K.K.Verma, N. Shrivastava, “Exploring Role and Associated Challenges of Data Mining Technique and it's Implementation in E-Governance”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 4, No. 8, pp. 3473-3478, 2015.

[2] Joseph M. Woodside, Sh. Amiri, and B. Boldrin, “The Impact of ICT and Big Data on e-Government”, Proceedings of the Int'l Conf. on Advances in Big Data Analytics, pp. 117-118, 2015.

[3] Y.-Ch. Chen, T.-Ch. Hsieh, “Big Data for Digital Government: Opportunities, Challenges, and Strategies”, International Journal of Public Administration in the Digital Age, Vol. 1, No.1, pp.1-14, 2014.

[4] G. Koteswara Rao, Shubhamoy Dey, “Decision Support for E-Governance: A Text Mining Approach”, International Journal of Managing Information Technology (IJMIT) Vol.3, No.3, pp.73-91, 2011.

[5] G. Guan, L. Zhou, & P. Tang, "Research of Data Mining in Government Transparent Decision-making", The 2009 International Symposium on Web Information Systems and Applications (WISA 2009), pp. 335-338, 2009.

[6] P. Milić, N.Veljković & L. Stoimenov, "Framework for open data mining in e-government", Proceedings of the Fifth Balkan Conference in Informatics ACM, pp. 255-258, September 2012.

[7] S. Kumar, H. Aggarwal, U. Kaur & Sumati, “E-Governance and Data Mining: a methodology for e-voting”, International Journal of Information Technology and Knowledge Management, Vol 2, No. 1, pp. 139-144, 2009.

[8] A. Mujawar and V. Patil , “Fuzzy based data mining system for E-government”, Elixir Comp. Sci. & Engg. 51A, pp. 11066-11068, 2012.

[9] V. R. Rao, " A Framework for e-Government Data Mining Applications (eGDMA) for Effective Citizen Services - An Indian Perspective", International Journal of Computer Science and Information Technology Research, Vol. 2, No 4, pp. 209-225, 2014.

[10] G. Stylios, D. Christodoulakis, J. Besharat, M.-A. Vonitsanou, I. Kotrotsos, A. Koumpouri, S. Stamou, “Data Mining Applications in the Service of E-Government”, Digital Democracy, Vol. 1, pp. 544-558, 2012.

[11] G. Šimić, Z. Jeremić, E. Kajan, D. Randjelović, A. Presnall, “A Framework for Delivering e-Government Support”, Acta Polytechnica Hungarica, Vol. 11, No. 1, pp. 79-96, 2014.

[12] J. Cardeñosa, C Gallardo , J.M. Moreno-Jiménez, “Text mining techniques to support e-democracy systems”, Proceedings of the İnternational Conference on E-Learning, E-Business, Enterprise İnformation System and E-Government, pp. 401-405, 2009.

[13] A. B. Adeyemo and A. K. Ojo, “Classification of Social Blogs Comments Using Text Mining”, International Journal of Computer Science, Vol. 11, No 1, pp. 54-58, 2014.

[14] G. Sangeetha, Dr. L.M. Rao, “A Review on Contribution of Data mining in e-Governance Framework”, International Journal of Engineering Research and General Science, Vol. 3, No. 2, pp. 68-75, 2015.

[15] G. Sangeetha , R. L. Manjunatha, “Efficient Framework of e-Government for Mining Knowledge from Massive Grievance Redressal

Data”, International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 4, No. 5, pp. 671-677, 2015.

[16] P. Desai, “A Survey of Data Mining Research and Implementation for E-Governance Systems”, International Journal of Computer Science, Vol. 4, No.7, pp. 103-105, 2014.

[17] Dr. R. R. Velamala, “A Framework for e-Government Data Mining Applications (eGDMA) for Effective Citizen Services” -An Indian Perspective”, International Journal of Computer Science and Information Technology Research, Vol. 2, No. 4, pp. 209-225, 2014.

Seksiya # 2
Big Data və İnformasiya Təhlükəsizliyi

E-dövlətin İnformasiya Təhlükəsizliyi üçün Big Data Texnologiyaları

Yadigar İmamverdiyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
yadigar@lan.ab.az

***Xülasə* — Big Data texnologiyaları bir sıra spesifik informasiya təhlükəsizliyi problemləri yaratmaqla yanaşı, informasiya təhlükəsizliyinin təmin edilməsi üçün yeni imkanlar da vəd edir. Bu işdə Big Data texnologiyalarının gətirdiyi yeni informasiya təhlükəsizliyi təhdidləri analiz edilir və bu texnologiyaların e-dövlətin informasiya təhlükəsizliyi sistemində tətbiqi imkanları qiymətləndirilir.**

Açar sözlər — e-dövlət; Big data; informasiya təhlükəsizliyi; informasiya təhlükəsizliyinin monitorinqi

I. GİRİŞ

Hazırda dövlət orqanları qarşısında dövlət təşkilatlarının fəaliyyətində şəffaflığın artırılması və iş məhsuldarlığının yüksəldilməsi problemləri durur. Buna görə də dövlətin əhali ilə tez, şəffaf və səmərəli şəkildə qarşılıqlı əlaqə saxlamağa, vətəndaşlara xidmət müddətini azaltmağa, əhalinin xidmətlərə tələbatını artırmağa, maliyyə effektivliyini yüksəltməyə, inzibati xərcləri azaltmağa imkan verən, korrupsiyanın aradan qaldırılmasına şərait yaradan informasiya sistemlərinin yaradılmasına və istismarına ehtiyacı vardır. Mövcud texnologiyalar artıq bu məsələləri həll edə bilmir. Bu bir neçə səbəblə şərtlənir.

Dövlət sektoru çox böyük həcmdə verilənlər toplayır, dünyada heç bir kommersiya strukturunun dövlət və hakimiyyət orqanları qədər verilənlər toplamaq imkanı yoxdur. Lakin bu verilənlərin 70%-i strukturlaşdırılmayıb və mətnlər və müxtəlif sənədlər şəklində saxlanır. Bu verilənlərin həcmi daim artır, dünya üzrə illik artım sürəti 30%-dir. Vətəndaşlar daim bu və ya digər şəkildə (o cümlədən, sosial şəbəkələr vasitəsilə) dövlət orqanları ilə ünsiyyətdə olurlar, elektron dövlətin formalaşdırılması gedişində elektron xidmətlər genişlənir, mövcud kağız informasiyasının elektronlaşdırılması həyata keçirilir.

Nəticədə verilənlər təkrarlanır və həcmləri artır, onların emalı getdikcə daha çox resurs tələb edir. Bu verilənləri sistemləşdirmək, arxiv verilənləri ilə işləmək, lazım olan informasiyanı arxivdə axtarıb tapmaq getdikcə çətinləşir. Strukturlaşdırılmamış şəkildə daxil olan informasiyanı sonrakı analiz məqsədilə saxlamaq üçün xüsusi emal etmək lazım gəlir. Bu çox zaman informasiyanın analizi tsiklini əhəmiyyətli dərəcədə yavaşıdır və verilənlərin emalına və saxlanmasına çəkilən xərcləri artırır.

Dövlət sektorunda çox böyük həcmdə toplanan verilənlərin saxlanılması və analizi zamanı bir sıra texnoloji problemlər də meydana çıxır. Böyük həcmdə və müxtəlif formatda toplanmış bu verilənlərin Big Data texnologiyaları olmadan, relyasion verilənlər bazaları texnologiyaları ilə emalı getdikcə çətinləşir və qeyri-mümkün olur. Nəticədə dövlət hakimiyyəti orqanları Big Data problemləri ilə qarşılaşırlar və bu vəziyyət yeni texnologiyaların tətbiq edilməsini tələb edir. Beləliklə, dövlət sektorunun vətəndaşlarla sürətli, şəffaf və səmərəli qarşılıqlı əlaqəsi üçün Big Data həllərinə böyük ehtiyacı vardır [1].

Lakin Big Data texnologiyaları e-dövlətin informasiya infrastrukturunda toplanmış fərdi məlumatların təhlükəsizliyinə bir sıra özünəməxsus təhlükələr yaradır. Bununla yanaşı, Big Data texnologiyalarının perspektiv tətbiq sahələrindən biri də e-dövlətin informasiya təhlükəsizliyinin təmin edilməsi sisteminin təkmilləşdirilməsidir [2]. Bu işdə Big Data texnologiyalarının yaratdığı spesifik informasiya təhlükəsizliyi təhdidləri analiz edilir və bu texnologiyaların e-dövlətin informasiya təhlükəsizliyinin idarə edilməsi sistemində tətbiqi problemlərinə baxılır.

II. E-DÖVLƏTDƏ BİG DATA TƏTBİQLƏRİ

Dövlət sektorunda Big Data tətbiqlərinin sayı artır, burada anbarlarda verilənlərin həcmi yüzlərlə terabayta və ya petabayta çata bilər. MGI mütəxəssislərinin tədqiqatlarına görə, Avropa İttifaqının dövlət sektoru yalnız verilənlərin şəffaflığının artırılması və Big Data qabaqcıl analitika texnologiyalarının istifadəsi hesabına inzibati xərcləri 15-20% (150-300 milyard yevro) azalda bilər [3].

IBM şirkəti dövlət sektorundan 28 İT-direktorun rəy sorğusu əsasında «Big Data potensialının reallaşdırılması» (“Realizing the Promise of Big Data”) adlı hesabat hazırlamışdır, hesabatda Big Data sahəsində dövlət layihələrinə xas olan 10 ümumi tendensiya müəyyən edilmişdir [4]. Tədqiqatlar göstərmişdir ki, dövlət təşkilatları Big Data tətbiqində hələlik başlanğıc mərhələdədirlər.

Gartner-in tədqiqatları Big Data investisiyalarına görə dünyada ən az inkişaf etmiş sektor olaraq dövləti müəyyən edir (cəmisi 16 %) [5]. Bir sıra ölkələr Big Data strategiyalarını müəyyən etməyə cəhd edirlər. Məsələn, Avstraliya hökuməti 2013-cü ilin avqustunda “Avstarliya dövlət xidmətləri üzrə Big Data strategiyası”nı qəbul etmişdir [6]. Bu strategiya mövcud xidmətləri genişləndirmək, yeni xidmətlər təklif etmək və daha yaxşı siyasəti təmin etmək və eyni zamanda gizliliyin qorunması üçün ən yaxşı təcrübəni tətbiq etmək və mövcud İKT investisiyalarını dəstəkləmək üçün Big Data sahəsində sistemli dövlət yanaşmasıdır. Bu sahədə ABŞ-ın təşəbbüsləri

də əhəmiyyətlidir [7]. Avropa İttifaqı da Big Data sahəsində öz strateji seçimini edib [8].

Son dövrlər hüquq-mühafizə orqanları və milli təhlükəsizliyi təmin edən strukturlar Big Data ideyalarından cinayətkar və terrorçu elementlərin daha dəqiq aşkarlanması üçün istifadə edilməsində maraqlıdırlar.

Big Data texnologiyalarından istifadə edən dövlət təşkilatları çox zaman verilənlərin qanuniliyi və keyfiyyəti, onların qeyri-müəyyənliyi, verilənlərin emalı və analizinin keyfiyyəti ilə bağlı problemlərlə üzləşirlər. Bu insanları böyük risk altında qoyan səhv qərarlara gətirib çıxara bilər. Big Data istifadəsinin mənfi nəticələri təkcə ayrı-ayrı insanlara deyil, ictimai həyatda, iqtisadiyyatda və siyasətdə də özünü göstərə bilər.

III. İNFORMASİYA TƏHLÜKƏSİZLİYİNDƏ PARADİQMA DƏYİŞİKLİYİ

İnformasiya təhlükəsizliyində paradiqma dəyişikliyi haqqında danışmaq üçün bir sıra şərtlər mövcuddur. Proqram məhsullarının və informasiya sistemlərinin mürəkkəbliyi artdıqca onlarda mövcud olan boşluqlar da artır və bədniyyətlilərin istifadəçilərə və istehsalçılara hələlik məlum olmayan boşluqlardan istifadəsi ehtimalı yüksəlir. İnformasiya texnologiyaları və informasiya təhlükəsizliyiin sahəsində mütəxəssislərin peşəkarlığı artır və asimmetrik aktorların yüksək ixtisaslı mütəxəssisləri bədniyyətli hərəkətlərə cəlb etməsi imkanları asanlaşır. Asimmetrik aktorlar hücum silahlarının inkişaf etdirilməsinə də yetərli investisiya yatıra bilərlər. Əsas məqsədi ziyan vurmaq olan ənənəvi kiber-hücumlardan fərqli olaraq müasir kiber-hücumlar konkret hədəfə yönəlir, daha yaxşı təşkil olunur və əlaqələndirilir. Bu hücumların əsas məqsədi tez aşkarlanan ziyan vurmaq deyil, qiymətli verilənlərə çıxış əldə etmək və onu mümkün olduqca uzun müddət – aylar, hətta illər ərzində saxlamaqdır. Qiymətli verilənlər dedikdə təşkilatın intellektual mülkiyyəti (proqram məhsulunun ilkin kodları, alqoritmlər, müştərilər bazası, istənilən digər korporativ sirlər) nəzərdə tutulur.

Hədəfə yönəlik hücumların böyük sinfini APT (Advanced Persistent Threat) hücumları təşkil edir [9]. Termin ABŞ-a yönəlmiş, xarici kəşfiyyat /dövlət tərəfindən dəstəklənən hücumları təsvir etmək üçün 2006-cı ildə ilk dəfə hərbçilər tərəfindən işlədilmişdi.

APT təhdidlərin xüsusi növüdür, konkret dövlət strukturlarına, şirkətlərə, təşkilatlara və hətta şəxslərə yönəlir. Bu növ təhdidlərin əsas fərqli cəhəti yalnız onların hədəflərinin konkretliyi deyil, həm də qabaqcıl informasiya texnologiyaları ilə yanaşı, psixologiya, sosial mühəndislik və s. metodlarına əsaslanan ən müasir yanaşmalardan istifadə etməsidir.

Bu mürəkkəb terminin hər bir sözü təsadüfi seçilməyib, xüsusi məna daşıyır:

- ***Advanced* (*məqsədyönlü*).** Hücumun mənbəyi – yaxşı maliyyələşdirilən təşkilatdır, onun sərəncamında hücumu həyata keçirmək üçün kifayət qədər resurs, hesablama texnikası və yaxşı təlim görmüş, kompüter sistemlərinə girmək vərdişlərinə malik yüksək ixtisaslı mütəxəssislər var.
- ***Persistent (davamlı).*** Hücum edən səbirlidir, müəyyən məqsədlə hərəkət edir və öz məqsədinə çatmaq üçün əhəmiyyətli səy göstərməyə hazırdır. Əgər bir hücum üsulu uğursuz olarsa, digər üsulla nəticə əldə etməyə cəhd edilir. Adi bədniyyətlidən fərqli olaraq, hədəf diqqətlə seçilir və hücum aylarla, hətta illərlə davam edə bilər.
- ***Threat (təhdid).*** Hücumun mənbəyi hədəfin maraqlarına təhdid təşkil edir.

Hazırda informasiya təhlükəsizliyi sistemlərinin çoxu siqnaturaların və məlum təhdidlərin davranış modellərinin axtarışına əsaslanır. Belə sistemlər siqnaturaları hələlik məlum olmayan yeni hücumlara, o cümlədən APT-hücumlarına qarşı gücsüzdürlər. Big Data texnologiyaları əsasında bu hücumların vaxtında aşkarlanmasına ümid edilir [9].

IV. BİG DATA VƏ YENİ İNFORMASİYA TƏHLÜKƏSİZLİYİ TƏHDİDLƏRİ

Big Data texnologiyalarının təhlükəsizliyi üzrə son illər bir neçə işçi qrup yaradılmışdır. Onlar arasında bir sıra aparıcı ABŞ universitetinin və şirkətinin tədqiqatçılarını birləşdirən Cloud Security Alliance (CSA) qrupu 2011-2013-cü illərdə Big Data texnologiyalarının yaratdığı informasiya təhlükəsizliyi problemlərini analiz etmiş və bir neçə analitik hesabat hazırlamışdır [10-12]. Aşağıda bu qrupun müəyyən etdiyi 10 informasiya təhlükəsizliyi təhdidi təsvir olunur [10, 11].

1) Paylanmış proqramlaşdırma platformalarında informasiya təhlükəsizliyi

Paylanmış proqramlaşdırma platformaları çox böyük həcmdə verilənləri emal etmək üçün hesablama və saxlama proseslərində paralel emaldan istifadə edirlər. Geniş yayılmış yanaşma MapReduce platformasıdır MapReduce-un işi iki addımdan ibarətdir: Map və Reduce. Map-addımda giriş verilənləri ilkin emal edilir. Bunun üçün kompüterlərdən biri (mepper – əsas qovşaq) məsələnin ilkin verilənlərini alır, onu hissələrə bölür və ilkin emal üçün paylanmış fayl sisteminin işçi qovşaqlarına paylayır. Reduce-addımda əsas qovşaq ilkin emal edilmiş verilənləri işçi qovşaqlardan toplayır, onları birləşdirir və məsələnin həllini formalaşdırır. Hücumların qarşısını almaq üçün burada iki əsas tədbir var: mepperin təhlükəsizliyi və etibar edilməyən mepper halında verilənlərin təhlükəsizliyi.

2) Qeyri-relyasyon verilənlər bazaları üçün ən yaxşı təhlükəsizlik qaydaları

NoSQL (not only SQL və ya no SQL) − ənənəvi relyasion verilənlər bazalarında istifadə olunan SQL vasitəsilə verilənlərə müraciətdən əhəmiyyətli dərəcədə fərqlənən verilənlər bazası modellərinin reallaşdırılmasına yönəlmiş bir sıra layihələri, yanaşmaları bildirən termindir (2009-cu ildə meydana çıxıb). NoSQL texnologiyası ilə yaradılan qeyri-relyasion verilənlər bazaları təhlükəsizlik infrastrukturu baxımından hələlik inkişaf mərhələsindədir. Məsələn, NoSQL-inyeksiya hücumlarına qarşı dayanıqlı həllər hələlik əlyetər deyil. Hər bir NoSQL verilənlər bazası analitikanın müxtəlif problemini həll etmək üçün yaradılmışdır və bu səbəbdən təhlükəsizliyə layihələndirmə mərhələsində heç vaxt modelin əsas hissəsi

kimi baxılmamışdır. NoSQL verilənlər bazasında təhlükəsizlik üçün heç bir dəstək təmin edilmir və belə bazalardan istifadə edərkən təhlükəsizlik adətən aralıq proqram təminatının öhdəsinə buraxılır.

3) Verilənlərin və tranzaksiya loqlarının təhlükəsiz saxlanması

Verilənlər və tranzaksiya loqları verilənlərin çoxsəviyyəli saxlama mühitində saxlanır. Verilənlərin səviyyələr arasında mexaniki köçürülməsi zamanı IT-menecer hansı verilənin nə zaman köçürülməsinə birbaşa nəzarət edə bilir. Lakin verilənlərin həcminin eksponensial artması verilənlərin səviyyələr arasında avtomatik köçürülməsini tələb edir. Avtomatik köçürmə həlləri verilənlərin harada saxlanılmasını izləmir və bu təhlükəsizliyi təmin etmək üçün yeni problemlər yaradır.

4) Son-nöqtədə daxiletmənin yoxlanması

Böyük verilənlərin istifadəsi zamanı verilənlər bir çox mənbədən toplanır. Məsələn, SIEM sistemi (Security Information and Event Management – təhlükəsizlik məlumatlarının və hadisələrinin idarə edilməsi) şəbəkədə minlərlə aparat və proqram təminatından hadisə loqlarını toplaya bilər. Məlumatların toplanması prosesində əsas problem toplanmış verilənlərə nə dərəcədə etibar edilə bilməsidir. Verilənlər mənbəyinin həqiqi, yoxsa bədniyyətli olmasını aydınlaşdırmaq, bədniyyətlə daxil edilmiş verilənləri həqiqi verilənlərdən ayırmaq, filtrasiya etmək tələb edilir.

5) Təhlükəsizliyin real zaman rejimində monitorinqi

Təhlükəsizliyin real zamanda monitorinqi təhlükəsizlik qurğularının yaratdığı siqnalların sayı baxımından həmişə böyük problem olmuşdur. Bu siqnalların bir çoxu hücum haqqında səhv siqnallar olur, onların öhdəsindən gəlmək mümkün olmadığından, sadəcə, nəzərə alınmırlar. Bu problem həcmi və sürəti artan Big Data nəticəsində daha da böyüyəcək. Lakin Big data texnologiyaları sürətli emal və məlumatların müxtəlif növlərinin analizi ilə müəyyən imkanlar da yaradır. Bu, məsələn, anomaliyaların real zamanda aşkarlanmasına imkan verə bilər.

6) Gizliliyi saxlayan Big Data analitikası

Big data şəxsi həyatın gizliliyinə, vətəndaş azadlıqlarının pozulmasına potensial təhdidlər yaradır, dövlət və korporativ nəzarət imkanlarını artırır. Şirkətlərin marketinq məqsədləri üçün Big Data analitikasından istifadə edərək şəxs barəsində gizli məlumatlar əldə edə bilərlər (məsələn, pərakəndə satış firması yeniyetmə qızın hamilə olduğunu onun ailəsindən əvvəl öyrənmişdi). Eynilə, analitika üçün verilənlərin anonimləşdirilməsi istifadəçi gizliliyini qorumaq üçün kifayət deyil. Məsələn, AOL tədqiqat məqsədləri üçün anonim axtarış loqlarını açmışdı, lakin istifadəçiləri axtarış sorğuları əsasında asanlıqla müəyyən etmək mümkün olmuşdu. Netflix də oxşar problemlə üzləşmişdi, IMDB balları ilə Netflix film ballarını əlaqələndirməklə Netflix anonim verilənlərindən istifadəçilər identifikasiya edilmişdi. Buna görə də, gizliliyin təsadüfən pozulmalarının qarşısını almaq üçün qaydalar və tövsiyələr yaratmaq vacibdir.

7) Kriptoqrafiya ilə həyata keçirilən girişə nəzarət və təhlükəsiz kommunikasiya

Daha həssas fərdi məlumatların tam təhlükəsizliyini və yalnız icazə verilmiş subyektlər tərəfindən müraciət edilməsini təmin etmək üçün verilənlər girişə nəzarət siyasəti əsasında şifrlənməlidir. Autentifikasiyanı və rəqəmsal imzanı təmin etmək üçün paylanmış subyektlər arasında təhlükəsiz kriptoqrafik kommunikasiya platforması reallaşdırılmalıdır.

8) Girişə qranulyar nəzarət

Girişə nəzarətin funksiyası verilənlərə giriş hüququ olmayan istifadəçilərin bu verilənlərə girişinin qarşısını almaqdır. Girişə qranulyar nəzarət nexanizmlərinin əsas problemi ondadır ki, paylaşıla bilən verilənlər təhlükəsizliyə zəmanət vermək üçün çox zaman daha məhdudlaşdırıcı kateqoriyalar altında saxlanır. Girişə qranulyar nəzarət isə təhlükəsizliyi pozmadan daha çox istifadəçiyə giriş hüququ verməyə çalışır.

9) Qranulyar audit

Təhlükəsizliyin real zamanda monitorinqində hücum baş verdiyi anda xəbər veriləcəyinə cəhd edilir. Əslində, bu həmişə belə olmur (məsələn, yeni hücumlar, hücumların aşkarlanmaması). Buraxılmış hücumları aşkarlamaq üçün audit məlumatları lazımdır. Bu yalnız nəyin baş verdiyini və nəyin yanlış olduğunu aşkarlamaqla bağlı deyil, standartlara uyğunluq və məhkəmə ekspertizası ilə də əlaqədar ola bilər. Bu baxımdan audit yeni bir şey deyil, lakin miqyas və qranulyarlıq fərqli ola bilər. Məsələn, baxılacaq verilənlər obyektləri çox ola bilər və onlar paylanmış ola bilərlər.

10) Verilənlərin mənşəyi

Bir sıra mühüm təhlükəsizlik proqramları verilənlərin mənbəyinin bilinməsini – onların yaradılması haqqında ətraflı məlumat əldə edilməsini tələb edir. Məsələn, maliyyə şirkətlərində insayder fəaliyyətinin aşkarlanması təhqiqatları verilənlər mənbəyinin dəqiqliyini tələb edir. Big Data tətbiqləri inkişaf etdikcə verilənlərin mənbəyi haqqında metaverilənlər də sürətlə artır və onların emalı üçün sürətli alqoritmlər tələb edilir.

V. İNFORMASİYA TƏHLÜKƏSİZLİYİ ÜÇÜN BİG DATA ANALİTİKASI

Big Data texnologiyaları informasiya təhlükəsizliyi sahəsində xeyli əvvəl tətbiq edilir. İlk SIEM sistemləri hələ 1996-cı ildə təqdim edilmişdi, bu sistemlər 2000-ci illərin ortalarından geniş yayılmağa başlayır. SIEM-sistemlərin nə üçün Big Data texnologiyalarına aid edilə bilməsi sualı, heç kimdə şübhə doğurmamalıdır: hadisələrin toplanması, normallaşdırılması və korrelyasiyası istənilən müasir Big Data sisteminin əsas və məcburi elementidir.

Ənənəvi SIEM-sistemlərinin bir sıra nöqsanları vardır:

- təhdidlərin aşkarlanması və təhqiqatı sahəsində imkanlar məhduddur;
- məhdud verilənlər mənbələri istifadə edilir;
- verilənlərin süzülməsi/normallaşdırılması zamanı verilənlərin itkisi qaçılmazdır;

– relyasion verilənlər bazaları miqyas və sürət problemləri yaradır;

– insidentlərin təhqiqatı və məhkəmə ekspertizası imkanları məhduddur.

SIEM-sistemlərin Big Data texnologiyaları ilə inteqrasiyasının bir neçə yolu movcuddur. Hazırda SIEM-sistemlərin ikinci nəslinin meydana çıxması haqqında danışmaq olar, lakin onların funksiyalarını bir çox cəhətdən intellektual təhlükəsizlik (Security Intelligence) sahəsinə aid etmək olar.

Gartner ekspertləri hesab edirlər ki, Big Data analitikası informasiya təhlükəsizliyi sistemlərini daha etibarlı etməyə və onlarda səhv həyəcan siqnallarını aradan qaldırmağa imkan verəcək [5]. Big Data analitikası üzrə həllərin böyük üstünlüyü təkcə informasiya təhlükəsizliyi sistemlərinin hadisələri arasında deyil, biznes-sistemlərin hadisələri ilə də korrelyasiyanın təmin etməsi imkanıdır.

İnformasiya təhlükəsizliyi həlləri təqdim edən bir çox şirkət Big Data texnologiyalarının informasiya təhlükəsizliyi sahəsində imkanlarını vurğulayan marketinq məlumatları çap etdirmişdir [13, 14]. CSA işçi qrupunun hesabatında Big Data-nın təhlükəsizlikdə rolu araşdırılır və mümkün tədqiqat istiqamətləri diqqətə çatdırılır [12]. İnformasiya təhlükəsizliyi sahəsində Big Data aşağıdakı imkanları təqdim edir [2, 12, 15]:

– kiber-hücumların aşkarlanması dəqiqliyinin yaxşılaşdırılması;

– real zaman ərzində korrelyasiya və anomaliyaların aşkarlanması;

– daha dərin məhkəmə ekspertizası ;

– böyük həcmdə verilənlərin vizuallaşdırılması və analizi üçün qrafik alətlər [16];

– daha yaxşı və sürətli qərarlar.

Sadalanan imkanların keyfiyyətli şəkildə reallaşdırılması müvafiq elmi və praktiki tədqiqatların aparılmasını tələb edir. Müxtəlif məlumat mənbələrindən alınmış çox böyük həcmdə verilənləri emel edə bilən aşkarlama alqoritmlərinin işlənməsi tələb edilir. Hazırda informasiya təhlükəsizliyi hadisələrinin aşkarlanması üçün Big Data analitikasından istifadə edən və perspektivli nəticələr göstərən az sayda iş vardır [17-20].

Big Data e-dövlətin informasiya təhlükəsizliyinin real zaman rejimində monitorinqi üçün müəyyən imkanlar vəd edir. Big Data texnologiyalarının e-dövlətin informasiya təhlükəsizliyinin idarə edilməsi sistemində tətbiqinə – Big Data əsasında təhlükəsizlik əməliyyatları mərkəzinin (Security Operations Center) qurulması məsələsinə [21]-də baxılır.

İnformasiya təhlükəsizliyi sahəsində Big Data texnologiyalarının uğurlu tətbiqi üçün uyğun ixtisaslı kadrların olması çox vacib elementdir. Bu baxımdan problemlərdən biri belə kadrların çatışmazlığıdır. Xüsusi bacarıqlara verilənlərin idarə edilməsi, verilənlərin analizi və təhlükələrin analizi bacarıqları daxildir. Bu bacarıqların hər üçünün bir şəxsdə olmasına az rast gəlinir, buna görə təşkilatların Big Data səylərində optimal nəticələrə nail olmaq üçün əməkdaşlıq edən mütəxəssis qruplarının yaradılması lazımdır.

NƏTİCƏ

Big Data böyük həcmdə heterogen verilənlərin analizi sahəsində böyük perspektivlər vəd edən paradiqma kimi meydana çıxmışdır. Big Data texnologiyaları özü ilə yeni informasiya təhlükəsizliyi təhdidləri gətirir, eləcə də təhlükəsizlik həllərində köklü dəyişikliklər vəd edir.

Lakin informasiya təhlükəsizliyinin təmin edilməsi üçün əhəmiyyətli imkanlar vəd etməsinə baxmayaraq, bu Big Data potensialından tam istifadə etmək üçün bu məqalədə göstərilmiş bir çox problem həll edilməlidir. Hədəfəyönəlik hücumların aşkarlanması, verilənlərin sızmasının aşkarlanması, məhkəmə ekspertizası, təhlükəsizlik kəşfiyyatı və vizuallaşdırma sahəsində tədqiqatlara yalnız indi diqqət yönəlməkdədir. Ona görə də tədqiqatçıların bu sahədə fundamental töhfələr verəcəyinə və innovasiyaların inkişafına əhəmiyyətli təsir göstərən kəşflər edəcəyinə böyük imkanlar və ümidlər vardır.

ƏDƏBİYYAT

[1] R. M. Əliquliyev, M. Ş. Hacırəhimova, “Big data” fenomeni: problemlər və imkanlar,” İnformasiya texnologiyaları problemləri, №2, s.3-16, 2014

[2] R.Alguliyev, Y.Imamverdiyev, “Big Data: Big promises for information security,” IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), 2014.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.

[4] K. C. Desouza, Realizing the promise of Big Data: Implementing Big Data projects. IBM Center for The Business of Government. 2014.
http://www.businessofgovernment.org/sites/default/files/Realizing%20the%20Promise%20of%20Big%20Data.pdf.

[5] L.Kart, N.Heudecker, F.Buytendijk, “Survey Analysis: Big Data Adoption in 2013 Shows Substance Behind the Hype.” September 2013. http://www.gartner.com/newsroom/id/2593815

[6] The Australian Public Service Big Data Strategy. August 2013. http://agict.gov.au/sites/default/files/Big%20Data%20Strategy.pdf

[7] Big Data Research and Development Initiative. March 29, 2012.
http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

[8] Communication on data-driven economy. http://ec.europa.eu/digital-agenda/en/news/communication-data-driven-economy

[9] A.K.Sood, R.J.Enbody, “Targeted cyberattacks: A superset of Advanced Persistent Threats,” IEEE Security & Privacy, Vol. 11, No. 1, pp. 54-61, 2013.

[10] Cloud Security Alliance (CSA): Top ten Big Data security and privacy challenges. 2012.

[11] Cloud Security Alliance (CSA): Expanded top ten Big Data security and privacy challenges. 2013.

[12] Cloud Security Alliance (CSA): Big Data Analytics for Security Intelligence. 2013. https://cloudsecurityalliance.org/download/big-data-analytics-for-security-intelligence

[13] J. Oltsik, “IBM: An Early Leader across the Big Data Security Analytics Continuum”. White paper, June 2013.

[14] M. Bouchard, “Big Data for advanced threat protection.” 2012.

[15] A. A. Cardenas, Manadhata P. K., and Rajan S. P., “Big Data Analytics for security,” IEEE Security & Privacy, Vol. 11, No. 6, pp.74-76, 2013.

[16] A. Shiravi, H. Shiravi, and A.A. Ghorbani, “A survey of visualization systems for network security,” IEEE Transactions on Visualization and Computer Graphics, Vol. 18, No. 8, pp. 1313-1329, 2012.

[17] Dumitras T., Shou D., “Toward a standard benchmark for computer security research: The Worldwide Intelligence Network Environment (WINE),” Proc. EuroSys BADGERS Workshop, pp. 89–96, 2011.

[18] J. François et al., “BotCloud: Detecting botnets using MapReduce,” Proc. Workshop Information Forensics and Security, pp. 1–6, 2011.

[19] T.-F. Yen et al., “Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks,” Proc. Annual Computer Security Applications Conference (ACSAC 13), pp. 199-208, 2013.

[20] Y. Lee, Y. Lee, "Toward scalable Internet traffic measurement and analysis with Hadoop," ACM SIGCOMM Computer Communication Review, Vol. 43, No. 1, pp. 5-13, 2013.

[21] Z.Fataliyev, Y.Imamverdiyev, H.Ko, “Security Operation Center architechture for e-government based on Big data Analysis”, Elektron dövlət quruculuğu problemləri I respublika elm-praktiki konfransı, s. 140-144, 2014.

Big Data Tətbiqinin Kibercinayətlə Mübarizədə Potensialı və Məhdudiyyətləri

Elvin Balacanov
Lids Universiteti, Lids, Böyük Britaniya və Şimali İrlandiya Birləşmiş Krallığı
Hüquq və Yeni Texnologiyalar Araşdırma Qrupu
lw11eb@leeds.ac.uk

***Xülasə* — Müasir dövrdə rəqəmsal innovativ texnologiyaların tətbiqi digər sahələrin inkişafı ilə yanaşı, cinayət mühakimə icraatının vəzifələrinin, xüsusilə də kibercinayətlə effektiv və səmərəli mübarizənin realizəsində mühüm rola və imkanlara malikdir. İnformasiya və kommunikasiya texnologiyalarının (İKT) insan həyatına inteqrasiyası fonunda artan kibercinayət halları ilə mübarizənin təşkili İKT-nin özünəməxsus xüsusiyyətlərinin mübarizə alət və vasitələrinin yaradılması, inkişafı və tətbiqi prosesində ehtiva olunması zərurətini yaratmışdır. Bu məqalədə kibercinayətlə mübarizədə Big Data-nın bəzi imkanları və perspektivləri, habelə tətbiqinin hüquqi və faktiki problemləri nəzərdən keciriləcəkdir.**

Açar sözlər — kibercinayət; İKT-spesifik mühafizə mexanizmləri; Big Data; Big Data Mining; Big Data analitika; şəxsi məlumatların toxunulmazlığı.

I. GİRİŞ

İnformasiya erasında rəqəmsal məlumatların böyük hissəsi Big Data formasında təşəkkül tapır. Yüksək emal sürəti, böyük həcm və forma müxtəlifliyi ilə xarakterizə olunan bu rəqəmsal informasiyanın (Big Data-nın) yaddaş qurğularında saxlanılması, effektiv istifadə və analiz olunması üçün tələb olunan səviyyə bu informasiyaya çıxışı olan subyektlərin müvafiq imkanlarını xeyli üstələyir. Nəticədə mövcud informasiya əsasında ənənəvi üsul və metodların tətbiqi vasitəsilə vaxtında və düzgün qərarların verilməsində bir sıra çətinliklər yaradır. Rəqəmsal texnologiyalar tərəfindən 2015-ci ildə emal olunan qlobal illik informasiya həcminin 2009-cu illə müqayisədə təxminən 10 dəfə artıq olduğunu, 2020-ci ildə isə bu göstəricinin 44 dəfədən daha yüksəkdə qərarlaşacağı proqnozunu [1] nəzərə aldıqda Big Data konsepsiyasının qaçılmazlığını, həmçinin mövcud uzlaşma fərqinin aradan qaldırılması yönündə sürətlə artan əhəmiyyətini və tətbiqi zərurətini anlamaq olar.

II. İKT: KİBERCİNAYƏTLƏ MÜBARƏZİNİN TƏRKİB ELEMENTİ KİMİ

Hüquq-mühafizə orqanları tərəfindən elm və kompüter texnologiyalarının imkanlarından analitik, əməliyyat-tətbiqi və qərarvermə kimi proseslərdə istifadə edilməsinə baxmayaraq, ümumilikdə “big data” konsepsiyasının ətraflı qəbulu və onun potensialından geniş istifadə olunması sahəsində özəl sektorla müqayisədə daha mühafizəkar bir mövqe tutulmuşdur.. Əlavə olaraq qeyd etmək lazımdır ki, hüquq-mühafizə orqanları məlumat toplama sahəsində geniş təcrübə və imkanlara malik olsalar da, toplanmış məlumatların analizi və istifadə oluna bilməsinin təmini zamanı hər bir halda yüksək səviyyədə səriştə nümayiş etdirmirlər [2]. Bu da öz-özlüyündə İKT-spesifik elementlərin mühafizə mexanizmlərinin yaradılması, inkişafı və tətbiqi zamanı nəzərə alınmasını şərtləndirir. Kibercinayətlə mübarizə isə bu tip alət və vasitələrin tətbiqindən xüsusilə asılıdır.

Kibercinayətlə mübarizə mexanizmlərinin yaradılması, inkişafı və tətbiqinin bütün mərhələlərində onun tərkib elementlərinin xüsusiyyətlərinin nəzərə alınması bu mexanizmlərin effektivliyinin və səmərəliliyinin təmini üçün zəruridir. Qeyd etmək lazımdır ki, Azərbaycan Respublikasının Cinayət Məcəlləsinin “Kibercinayətlər” fəslində öz əksini tapmış bütün cinayət əməllərinin obyektini birbaşa və ya bilavasitə kompüter sistemi, bu sistemin normal fəaliyyətinin təmini və ya kompüter məlumatlarının mühafizəsi sahəsində yaranan ictimai münasibətlər təşkil edir. Həmçinin, bu cinayətlərin predmeti də, ya kompüter sistemi (proqramları, qurğuları), ya da bu sistemin tərkibində olan məlumatlardır (və ya məlumat daşıyıcılarıdır). Kibercinayətlərin obyektiv cəhəti də hər bir halda bu və ya digər formada kompüterlə bağlı anlayışları ehtiva edir [3]. Bütün bunlar isə bu cinayətlərlə mübarizənin və cinayət mühakimə icraatının tamlığının təmin edilməsinin İKT-spesifik mexanizmlərdən asılılığı kimi şərh oluna bilər.

Bundan başqa, bu əməllərin subyektiv cəhətdən yalnız birbaşa qəsdlə törədilməsi halında cinayət əməli kimi tövsif olunması onun obyektiv cəhətinin ifadə olunduğu əməllərin icrasındakı kompleks və planlı fəaliyyət zərurətindən irəli gəlir. Belə ki, kibercinayətlər kompüter sisteminə qanunsuz daxil olma, məlumatların qanunsuz ələ keçirilməsi, saxtalaşdırılması, bu məqsədlərlə kompüter qurğu və ya proqramlarının yaradılması, dövriyyəsi formasında həyata keçirilir ki, bunlar da bir sıra digər cinayətlərin subyektlərindən fərqli olaraq, kibercinayətin subyektlərindən daha yüksək biliklərə və təcrübəyə malik olmağı tələb edir. Nəticədə isə bu cinayətlərlə mübarizə hüquq-mühafizəni təşkil edən subyektlərdən adekvat səviyyədə inkişaf etdirilmiş və daha spesifik rəqəmsal alət və vasitələrin tətbiqini tələb edir.

Əlavə olaraq qeyd olunmalıdır ki, İKT-spesifik mexanizmlərin rolu təkcə “Kibercinayətlər” fəslinə daxil

edilmiş müddəalarla bağlı cinayət-mühakimə icraatının həyata keçirilməsi ilə məhdudlaşmır. Müasir dövrdə digər bir sıra cinayət əməlləri ilə mübarizənin də tərkibində kiber elementlər ehtiva olunduğu üçün bu üsul və vasitələrin əhəmiyyəti və onlara olan zərurət günü-gündən daha da artır.

III. KİBERCİNAYƏTLƏ MÜBARİZƏDƏ BİG DATA

Kiberməkanda sürətli təkamül prosesi keçən informasiyanın həcmi, mürəkkəbliyi, emalı və dövriyyə sürəti verilənlərin ənənəvi üsul və metodlarla səmərəli analizinin həyata keçirilməsini az qala mümkünsüz edir. Hüquqi və faktiki cəhətdən tətbiqi məhdudiyyətlər və uyğunsuzluqlarla müşayiət olunsa da, Big Data kibercinayətlə mübarizədə bir sıra texniki imkanlara malikdir.

A. Texniki aspektlər

Kibercinayətlə mübarizənin IKT-spesifik mexanizmlərin tətbiqindən asılılığını nəzərə alaraq iddia etmək olar ki, Big Data-nın bu istiqamətdəki imkanlarından istifadə etməklə mübarizənin effektivliyini və səmərəliyini əhəmiyyətli dərəcədə artırmaq olar. Çünki Big Data konsepsiyası məhz İKT sahəsində vüsət alan innovasiyalar sayəsində böyük həcmli məlumatların yaradılması, emalı, yadda saxlanması və ötürülməsi önündə analoq məlumatlarla bağlı mövcud bir sıra çətinliklərin aradan qaldırması nəticəsində diqqət mərkəzinə keçmişdir [4] . Buna görə də, mürəkkəb və böyük həcmli (peta və ekzabaytlarla ölçüyə malik həcmdə) verilənlərin emalı və analizi məqsədilə inkişaf etdirilən texniki və metodoloji üsul və vasitələrin müasir kompüterlər və kompüter sistemlərinin gücü vasitəsilə tətbiqi və realizəsi həm cinayətkar əməllərlə mübarizə, həm də bu məqsədlə ayrılan resursların daha səmərəli istifadəsində böyük potensiala malikdir. Bu anlamda Big Data Mining və Big Data analitika diqqətəlayiq bir mövqeyə sahibdir.

Bu üsul və vasitələr həm vahid struktura malik, həm də natamam strukturlu və struktursuz formada mövcud olan verilənlərin avtomatlaşdırılmış emalı və analizi nəticəsində onlar arasında adi halda çətin izlənilə bilən münasibət və əlaqələrin müəyyən olunması məqsədilə istifadə oluna bilər [5]. Ümumilikdə, bu cür münasibət və əlaqələr ibtidai araşdırmanın həyata keçirilməsi zamanı zəruri araşdırma istqamətlərinin daha aydın müəyyən olunmasında, cinayətlərin inkişaf istiqamətlərinin və onlarla mümkün mübarizə üsulları ilə bağlı daha düzgün seçimlərin edilməsində, habelə bu məqsədlə resursların daha səmərəli şəkildə bölüşdürülməsində mühüm rol oynaya bilər.

Risk profilləşdirilməsi məqsədilə tətbiq olunduğu təqdirdə isə bu üsul və vasitələr əvəllər izlənilməmiş əlaqə və münasibətlərin aşkara çıxarılması yolu ilə ayrı-ayrı cinayətlərin subyektlərinin, onların daxil olduğu mümkün şəbəkələrin, habelə şübhəli şəxslərin və ya onların dairəsinin müəyyən olunmasında geniş imkanlara malikdir. Şəxsin cinayətin icraçısına və ya cinayətin digər formada iştirakçısına çevrilmə riski, cinayətin baş verməsinə potensial şərait yaradan hallar və zərərçəkənlərin dairəsinin müəyyən olunmasında da Big Data Mining və Big Data analitika tətbiq oluna bilər [6].

Bunlarla yanaşı, Big Data əsasında həyata keçirilən risk profilləşdirilməsi profillərlə bağlı obyektivliyin daha yuxarı səviyyədə təmin olunmasına imkan yaratmasına baxmayaraq, analitikaya yönləndirilmiş verilənlərin məhdudluğu, natamamlığı və qeyri-dəqiqliyi qərarın subyektivliyinə və beləliklə, fərdlər üzrə tətbiq olunduğu təqdirdə ayrı-seçkiliyə gətirib çıxara bilər [7].

Qeyd etmək lazımdır ki, kibercinayətlərin araşdırılması və istintaqı üçün zəruri proseslərdən sayılan elektron/rəqəmsal sübutların toplanılması, emalı və sənədləşdirilməsi prosesi müvafiq kompüter avadanlıqları və proqramlar vasitəsilə avtomatlaşdırıla bilir. Hal-hazırda istifadə olunan bu avadanlıq və proqramların verilənlərin yalnız informasiyaya çevrilməsində iştirak etməsini və bu informasiyanın faydalılığının təmini və düzgün qərarvermənin araşdırmanı həyata keçirən şəxsdən asılı qalması bu sahədə əsas problemlərdən biri kimi qəbul olunur [8]. Lakin bu çətinliyin aradan qaldırılmasında Big Data analitikanın rolu rəqəmsal ekspertizanın həyata keçirilməsində mühüm potensial imkanlara malik bir üsul kimi nəzərə çatdırılmalıdır. Belə ki, digər ənənəvi analitika vasitələrindən fərqli olaraq, Big data analitikası araşdırmanı həyata keçirən şəxsdən araşdırmanın obyekti barədə əvvəlcədən məlumatlara malik olmanı, verilənlər arasında konkret münasibət və əlaqələrin axtarılmasını tələb etmir. Bunun əksinə olaraq son model inkişaf etmiş rəqəmsal analitika texnologiyaları vasitəsilə tətbiq olunduğu təqdirdə Big Data analitika verilənlərin modelləşdirilməsini aparır və nəticədə lazımı məlumatlar nümayiş olunur. Yekun olaraq araşdırmanı həyata keçirən şəxslərin üzərinə yalnız bu məlumatların məqsədəuyğunluğunun müəyyən olunması və araşdırma üçün faydalı məlumatların seçilib götürülməsi düşür [9]. Əvəzində isə doğru proqram və həllərin tətbiqi, habelə onların tətbiqi sahəsində yüksək bilik və peşəkarlıq tələb olunur.

Hal-hazırda ən geniş yayılmış Big Data həlləri qismində NoSQL, Apache Cassandra və Hadoop, xüsusilə də IBM Crime Information Warehouse (CIW) Big Data analitika və Big Data mining sahəsində müəyyən imkanlar vəd edir. Lakin hüquq-mühafizə və təhlükəsizlik orqanları tərəfindən cinayətlərin proqnozlaşdırılması və cinayət hallarının müəyyən olunmasında mühüm rola malik olmasına baxmayaraq, Big Data tətbiqlərinin (applications) sırf bu məqsədlərlə istifadəsinin yüksək effektivliyi barədə təsdiqləyici təcrübi məlumatlar çox azdır [7]. Halbuki, texniki həllər kibercinayətlə mübarizədə qanuni üsul və vasitələr qədər vacibdir və yalnız bu iki komponentin vəhdəti hüquq-mühafizə mexanizminin tamlığını təmin edir. Bu mexanizmin çatışmazlığı isə cinayətin baş verməsinə şərait yaradan üç əsas səbəbdən biridir [10].

Əlavə olaraq qeyd olunmalıdır ki, hüquq-mühafizə orqanlarının fəaliyyətinin təşkilində və cinayətkarlıqla mübarizə alət və vasitələrinin yaradılması, inkişafı və tətbiqində innovasiyaların inteqrasiyası prosesi ümumilikdə

aşağı sürətlə gedir. Həmçinin, baxmayaraq ki, rəqəmsal ekspertiza alət və vasitələri cinayət-mühakimə icraatının vəzifələrinin daha effektiv və daha səmərəli formada icrası üçün imkanlar yaradır, müasir texnologiyaların tətbiqi həm öz-özlüyündə mürəkkəb olması, həm də bu texnologiyaların tətbiqi üçün zəruri bilik və bacarıqlara malik insan resurslarının çatışmazlığı səbəbindən daha da çətinləşmişdir [11].

B. Hüquqi aspektlər

İKT-nin insan həyatının bütün sferalarına sürətli inteqrasiyası öz təsirini cinayət əməllərinin də kiberməkana transformasiyası formasında göstərmişdir. Kiberməkanda isə həyata keçirilmiş fəaliyyətin bütün mərhələləri rəqəmsal izlərlə müşayiət olunur ki, bunlar da onların izlənilə bilənliyini təmin edir. Kibercinayətlə mübarizə aspektindən götürüldükdə bu izlər sübuti əhəmiyyətə malikdir.

Lakin, Big Data konsepsiyası araşdırma üçün lazımi məlumatlara limitsiz çıxışın əldə olunması və bütün onlayn fəaliyyətlərin izlənilməsi vasitəsi kimi qəbul edilməməlidir. Bir tərəfdən, onlayn fəaliyyətin izlənilə bilənliyi cinayətlərin sayını azaldan 3 prinsipial faktordan biri kimi mühüm əhəmiyyət daşıyır. Belə ki, bütün hallarda izlənilmə və müəyyən oluna bilmə imkanının yüksək olduğunu dərk etmə şəxsin cinayətkar əməlin icrasından çəkinməsilə nəticələnə bilər [12]. Digər tərəfdənsə bu, əsassız onlayn nəzarət və internet trafiki barədə məlumat bazalarındakı məlumatların asanlıqla ələ keçirilməsinə zəmin yaradır. Nəticədə yuxarıda qeyd olunmuş texniki imkanlarla yanaşı Big Data şəxsi məlumatların subyektiv və qərəzli məqsədlər üçün əldə olunması və istifadəsi üçün də potensiala çevrilmişdir. Edward Snowden tərəfindən ABŞ Milli Təhlükəsizlik Agentliyinin Big Data analitikasından fərdlərin və təşkilatların kommunikasiya və yazışmalarının sistemli şəkildə izənilməsi ilə bağlı ictimaiyyətə ötürülmüş məlumatlar bu cür sui-istifadə hallarına misal kimi göstərilə bilər. Onu da əlavə etmək lazımdır ki, Snowden tərəfindən verilmiş məlumatlar ümumilikdə Big Data analitikasının nüfuzuna da mənfi təsir göstərmişdir [13].

Qeyd olunduğu kimi, Big Data-nın yaratdığı imkanlardan sui-istifadə ayrı-seçkiliyin yaranmasına da gətirib çıxara bilər. Belə ki, diskresion səlahiyyətlərin icrası zamanı hüquq-mühafizə orqanları tərəfindən Big Data analitikası nəticəsində əldə olunmuş məlumatlar əsasında subyektiv və tam əsaslandırılmamış qərarların da qəbul edilməsi mümkündür. Bundan başqa, Big Data anonimlik və konfidensiallığın qorunması önündəki texniki maneələri aradan qaldırdığı üçün bu anlayışlar daha açıq bir hədəf formasını almışdır. Bu isə sui-istifadəyə zəmin yaradan əlavə hallar qismində çıxış etmə imkanının yaranması deməkdir. Big Data-nın bu mənfi cəhətlərinin aradan qaldırılmasında qanunvericiliyin tələbləri müəyyən rola malikdir. Eyni zamanda, bu tələblər Big Data-nın həm də qanuni məqsədlər naminə istifadəsi prosesini mürəkkəbləşdirir.

Belə ki, məhkəmələr və cinayət prosesinin iştirakçıları tərəfindən Azərbaycan Respublikası Konstitusiyasına və Cinayət-Prosessual Məcəlləsinə, Azərbaycan Respublikasının digər qanunlarına, habelə Azərbaycan Respublikasının tərəfdar çıxdığı beynəlxalq müqavilələrə bütün hallarda ciddi riayət olunması qanunçuluq prinsipinin əsasını təşkil edir. Habelə, dövlət hər kəsin yazışma, telefon danışıqları, poçt, teleqraf və digər rabitə vasitələri ilə ötürülən məlumatın sirrini saxlamaq hüququna təminat verir [14]. Cinayət-mühakimə icraatının vəzifələrinin icrası zamanı insan və vətəndaş hüquq və azadlıqlarının qanunsuz məhdudlaşdırılması hallarından müdafiə, habelə hər bir cinayət təqibinin qanuniliyi və əsaslılığının təmini onun əsas prinsipləri və şərtlərinin təyinatı qismində çıxış etdiyini də nəzərə almaq vacibdir [15]. Bunlar isə Big Data analitika və Big Data mining-in praktikada tətbiqi zamanı riayət olunmalı olan standartların yüksəkliyi barədə ilkin mənzərə yaradır.

Onu da qeyd etmək lazımdır ki, Cinayət-Prosessual Məcəllə ilə müəyyən edilmiş hallarda cinayət prosesi prinsiplərinin və ya şərtlərinin pozulması cinayət təqibi üzrə başa çatmış icraatın etibarsız sayılmasına, onun gedişində qəbul edilmiş qərarların ləğvinə, yaxud toplanmış materialların sübutedici qüvvəsinin olmaması qənaətinə gəlməyə səbəb ola bilər [15]. Bu isə o deməkdir ki, Big Data-nın tətbiqinin hər hansı bir elementi və ya mərhələsi qanunvericiliyin tələblərinin pozulması ilə müşayiət olunursa, bu istiqamətdə aparılan prosessual hərəkətlər nəticəsində əldə olunmuş sübut və materialların prosessual qanunvericiliyə görə hüquqi qüvvəsi olmur [15]. Bu isə Big Data–nın tətbiqi ilə bağlı tənzimləmə məsələlərinin milli qanunvercilikdə təsbitini şərtləndirir.

NƏTİCƏ

Big Data kibercinayətlə mübarizə sahəsində potensial imkanlara malik olsa da, onun tətbiqi prosesində bir sıra texniki və hüquqi maneələr ortaya çıxır. Bu maneələrin böyük hissəsi cəmiyyətin idarə olunması üçün yaradılmış mexanizmlərin kiberməkanda tətbiqi üçün tam uyğun olmaması nəticəsində meydana gəlir. Beləliklə də, resurs çatışmazlığı özünü bir neçə müstəvidə daha kəskin şəkildə göstərməyə başlamışdır. Başqa sözlə, texniki, institutsional, hüquqi və insan resurlarının çatışmazlığı İKT sahəsində idarəetmənin və vəziyyəti nəzarət altında saxlamağın çətinləşməsi və mürəkkəbləşməsilə nəticələnmişdir.

Kiberməkanda təşəkkül tapmış cinayətlərlə mübarizə mexanizmlərinin yaradılması bu resursların hər biri ilə bağlı çatışmazlığın aradan qaldırılması üçün kompleks tədbirlərin tətbiqini şərtləndirir. Bu cür tədbirlər analoji olaraq Big Datanın da kibercinayətlə mübarizədə effektivliyi və səmərəliliyinin artırılması məqsədilə tətbiq olunması üçün zəruridir.

Big Data-nın tətbiqinin önündə yaranmış maneələrin aradan qaldırılması üçün ilkin olaraq onun tətbiqinin qanuniliyinin təmini üçün addımlar atılmalıdır. Bunun üçün isə müvafiq qanunvericiliyin texnoloji innovasiyaların inkişafı

və tətbiqini, habelə praktikaya sürətli inteqrasiyasını dəstəkləyəcək istiqamətdə inkişaf etdirilməsi zəruridir. Qanunvericiliyin təkmilləşdirilməsi insan hüquq və azadlıqlarının yüksək səviyyədə təminatı məqsədinə xidmət etməklə bütün mərhələlərdə vətəndaşların da prosesdə yaxından iştirakı ilə müşayiət olunmalıdır.

Eyni zamanda, kibercinayətlə mübarizəni həyata keçirən hüquq-mühafizə və təhlükəsizlik orqanları subyektlərinin bu sahədə müvafiq bilik və bacarıqlarının artırılması məqsədilə mütəmadi tədbirlərin həyata keçirilməsi, habelə texnoloji innovasiyalar sahəsində onlar daim yenilənmiş məlumatlarla təmin olunmalıdır. Bu, həmçinin Big Data-nın analitika məqsədilə tətbiqi üçün cəlb olunmuş texniki resursların daha effektiv və daha səmərəli istifadəsi üçün zəruridir.

ƏDƏBİYYAT

[1] Computer Sciences Corporation, 'Data Universe Explosion & the Growth of Big Data', 2015. http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode .

[2] B. Akhgar, B. Gregory, Application of Big Data for National Security, Butterworth-Heinemann, 2015.

[3] Azərbaycan Respublikasının Cinayət Məcəlləsi, http://www.e-qanun.az/code/11

[4] R. Walker, From big data to big profits. New York, NY: Oxford Univ. Press, 2015.

[5] SAS, Managing the Intelligence Life Cycle: A More Effective Way to Tackle Crime 2013. http://www.sas.com/en_us/whitepapers/managing-intelligence-life-cycle-106831.html

[6] J. James and F. Breitinger, Digital Forensics and Cyber Crime 2015: 7th International Conference, (Seoul, South Korea, ICDF2C), 2015, pp. 126-139.

[7] Postnote 470, Big Data, Crime and Security, UK Parliament, 2014.

[8] J. James and P.Gladyshev, Challenges with Automation in Digital Forensic Investigations, 2013, CoRR, abs/1303.4498, http://arxiv.org/abs/1303.4498

[9] Sas.com, 'How big data analytics can be the difference for law enforcement', 2015. http://www.sas.com/en_us/insights/articles/risk-fraud/big-data-analytics-for-law-enforcement.html.

[10] L. E. Cohen, M. Felson, “Social Change And Crime Rate Trends: A Routine Activity Approach”, American Sociological Review, 1979, Vol. 44, pp. 588-608.

[11] Crcs.seas.harvard.edu, 'Simson Garfinkel: "Automated Digital Forensics"', 2015. http://crcs.seas.harvard.edu/event/simson-garfinkel-automated-digital-forensics.

[12] M. Ouimet, ‘Internet and crime trends’ in Frank Schmalleger and Michael Pittaro, Crimes Of The Internet, Prentice Hall, 2009.

[13] B. Marr, Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance, John Wiley & Sons, Ltd, 2015.

[14] Azərbaycan Respublikasının Konstitutsiyası, http://e-qanun.az/framework/897

[15] Azərbaycan Respublikasının Cinayət Prosessual Məcəlləsi, http://e-qanun.az/code/14

İnformasiya Təhlükəsizliyi Üzrə Tədqiqatlar üçün Böyük və Kiçik Verilənlər

Yadigar İmamverdiyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
yadigar@lan.ab.az

***Xülasə* — Big Data texnologiyaları informasiya təhlükəsizliyi üzrə tədqiqatlarda böyük verilənlər toplularından istifadə edilməsini tələb edir. Lakin belə verilənlərin toplanması, tədqiqatçıların geniş dairəsi üçün əlyetər edilməsi sahəsində bir sıra hüquqi, təhlükəsizlik, etik, intellektual mülkiyyət, rəqabət və digər problemlər mövcuddur. Məqalədə informasiya təhlükəsizliyi sahəsində aktual tədqiqat istiqamətləri üzrə mövcud verilənlər topluları haqqında məlumat verilir, bu toplulara müraciət modelləri analiz edilir.**

Açar sözlər — informasiya təhlükəsizliyi; Big Data; verilənlərin paylaşılması; verilənlər toplusu; botnet

I. GİRİŞ

Big data əsrində informasiya təhlükəsizliyi üzrə tədqiqatların aparılması üçün verilənlər bazis rolunu oynayır. Zəngin və etibarlı verilənlər toplularının əlyetər olması elmi-tədqiqat işinin aparılması üçün yaxşı bir başlanğıcdır.

Lakin informasiya təhlükəsizliyi üzrə elmi-tədqiqatlarda istifadə edilən verilənlərin toplanması kifayət qədər çətin məsələdir. Bu çətinlik belə verilənlərin tədqiqatçılar arasında paylaşılması üçün səbəblərdən biridir. Bununla yanaşı, burada bəlkə də daha vacib olan digər bir səbəb də var. Eksperimentlərdə istifadə edilən verilənlərin paylaşılması nəticələri yoxlamağa və metodları müqayisə etməyə imkan verir. Bu yaxşı əsaslandırılmış elmi iş üçün fundamentlərdən biridir.

İnformasiya təhlükəsizliyi üzrə elmi-tədqiqatlarda istifadə edilən verilənlərin paylaşılmasında bir sıra etik, hüquqi və texnoloji problemlər meydana çıxır, bu problemlər [1]-də araşdırılır.

Bu işdə informasiya təhlükəsizliyi üzrə tədqiqatlarda istifadə edilən verilənlər topluları analiz edilir, onların toplanması, saxlanması və istifadəsi üzrə hüquqi, texnoloji və təhlükəsizlik problemləri araşdırılır. Verilənlər AMEA İnformasiya Texnologiyaları İnstitutunda informasiya təhlükəsizliyi üzrə aparılan və aparılması planlaşdırılan aşağıdakı elmi-tədqiqat istiqamətləri üzrə təsnif edilir:

- kiberhücumların aşkarlanması;
- DDoS hücumların aşkarlanması;
- botnetlərin aşkarlanması;
- spamın aşkarlanması;
- zərərli proqramların aşkarlanması;
- veb-texnologiyaların təhlükəsizliyi;
- informasiya təhlükəsizliyi risklərinin idarə edilməsi;
- informasiya təhlükəsizliyinin qiymətləndirilməsi;
- informasiya təhlükəsizliyi insidentlərinin idarə edilməsi;
- simsiz şəbəkələrdə informasiya təhlükəsizliyi problemləri.

Qeyd edək ki, verilənlərlə bağlı oxşar problem informasiya təhlükəsizliyi sahəsində tədqiqat aparan digər tədqiqatçıları da narahat edir. Aparıcı tədqiqatçıların bir qrupu 2010-cu ildə informasiya təhlükəsizliyi elminin reprezentativ verilənlərə ehtiyaclarını vurğulayan “arzular siyahısı”nı tərtib etmişdi [2]. Siyahı müəllifləri hazırda əlyetər olmayan spesifik verilənlər üçün metaverilənlərin toplanmasını asanlaşdıran və hüquqi və gizlilik problemlərinin həllinə kömək edən məlumatların paylaşılması prosesinə tələbləri də müəyyən etməyə cəhd edirlər.

İşin strukturu aşağıdakı kimidir. İkinci bölmədə müdaxilələrin aşkarlanması sistemlərinin (ing. Intrusion Detection System, IDS) test edilməsi üçün mövcud verilənlər topluları analiz edilir. Üçüncü bölmədə verilənlərin paylaşımına yanaşmalardan biri kimi PREDICT modeli və dördüncü bölmədə WINE verilənlər toplusu araşdırılır. Beşinci bölmədə botnetlərin aşkarlanması üçün yeni işlənmiş verilənlər topluları və onlara yürüdülən tələblər təhlil olunur.

II. IDS SİSTEMLƏRİNİN TEST EDİLMƏSİ ÜÇÜN VERİLƏNLƏR

IDS sistemləri tarixən informasiya təhlükəsizliyi sahəsində verilənlərin intellektual analizinə əsaslanan ilk sistemlərdir. 1990-cı illərin ortalarından belə sistemlərin test edilməsi üçün etalon test verilənlərinin vacibliyi etiraf edildi və DARPA verilənləri topluları meydana çıxdı [3]. Ümumiyyətlə, IDS sistemləri üçün test toplularının kritik analizinə həsr olunmuş tədqiqat işləri yerinə yetirilmiş və yeni topluların yaradılması ilə bağlı xeyli təcrübə toplanmışdır. Bunu nəzərə alaraq bu bölmədə belə toplulların bir qədər ətraflı analizi aparılır.

IDS sistemlərinin test edilməsi üçün tədqiqatçılara bir çox verilənlər toplusu əlyetərdir: DARPA [3], KDD’99 [4], Internet Trafik Arxivi [5], LBNL [6], CAIDA [7], DEFCON [8,9], PREDICT [10], və ISCX 2012 IDS verilənlər topluları [11].

DARPA verilənlər topluları şəbəkə təhlükəsizliyi məqsədləri üçün orta ölçülü ABŞ aviabazasında müşühidə edilən şəbəkə trafiki imitasiya edilməklə yaradılmışdı. DARPA’98 və ’99 verilənlərinin diqqətli analizi nəticəsində bir sıra nöqsanlar aşkarlandı [12-14]: real trafikə tam oxşamaması,

verilənlər toplusuna süni yaradılmış verilənlərin daxil edilməsi nəticəsində meydana çıxan çoxsaylı uyğunsuzluqlar və s.

DARPA′98-də 7 həftəlik şəbəkə trafikinin təxminən 5 Qbaytlıq tcpdump verilənləri var. Onları hərəsi 100 bayt olmaqla 5 milyon bağlantı haqqında məlumatlara bölmək olar. Iki həftəlik test verilənlərində təxminən 2 milyon bağlantı məlumatı var. KDD təlim verilənləri toplusunda təxminən 4.9 milyon bağlantı vektoru var, hər bir vektorda 41 əlamət var və normal və ya müəyyən bir hücum növü ilə işarə edilib. İmitasiya edilmiş hücumlar dörd kateqoriyadan birinə aiddir:

1) Xidmətdən imtina hücumu (Denial of Service, DoS) – hücum edən müəyyən servisləri həddindən artıq yükləyir ki, qanuni istifadəçilərə xidmətdən imtina edilsin.
2) User to Root Attack (U2R): Hücum edən müəyyən eksploytlardan istifadə edərək sistemdə normal istifadəçi hesabından administrator hesabına yüksəlməyə cəhd edir.
3) Remote to Local Attack (R2L): hücum edən kompüterdə müəyyən boşluqlardan istifadə edərək həmin kompüterdə lokal istifadəçi hesabına giriş əldə etməyə cəhd edir.
4) Zondlama hücumu (ing. probing attack): informasiya təhlükəsizliyini pozmaq məqsədilə kompüter şəbəkəsi haqqında informasiya toplamağa cəhd edilir.

KDD′99-un test verilənlərinə təlim verilənlərində olmayan spesifik hücum növləri daxildir, bu aşkarlamanı reallığa daha yaxın edir. Təlim verilənlərində 24 hücum növü vardır ki, onlardan 14-ü test verilənlərində yoxdur. Təlim verilənlərindəki hücum növlərinin adları və ətraflı təsvirləri [4]-də verilir.

DARPA və KDD verilənlər topluları köhnəlmiş hesab edilir, bir sıra nöqsanlara malikdirlər, lakin hələ də istifadə edilirlər. Onların aşağıdakı nöqsanlarını göstərmək olar [12–14]:

1) Gizlilik naminə həm fon, həm də hücum verilənləri bir neçə ay ərzində ABŞ hərbi aviasiya bazalarında müşahidə edilən verilənlərə oxşar olan sintetik verilənlər generasiya edilmişdir. Lakin səhv aşkarlama xarakteristikalarının analitik və ya eksperimental yoxlamasına cəhd edilməmişdir. Eyni zamanda, sintez edilmiş verilənlər real şəbəkədəki trafikə oxşar deyil.

2) DARPA′98-də istifadə edilən tcpdump kimi trafik kollektorları ağır trafik yüklənməsi zamanı paketləri atırlar. Lakin atılmış paketlərin mümkünlüyü yoxlanılmamışdır.

3) Hücumların dəqiq tərifləri verilmir. Məsələn, əgər iterasiyaların sayı müəyyən həddi keçmirsə, onda zondlama əməliyyatı hücum hesab edilmir. Oxşar olaraq, bufer daşması yaradan paket də həmişə hücumu təmsil etmir. Buna görə də, bu şərtlər altında qiymətləndirənlə qiymətləndirilən arasında tərifləri barədə razılaşma olmalıdır. Lakin DARPA'98-də şəbəkə hücumlarının aydın tərifləri yoxdur.

4) İzafi məlumatların sayı olduqca çoxdur. Təlim və test verilənlərində məlumatların uyğun olaraq 78%-i və 75%-i təkrarlanır. Təlim verilənlərində böyük sayda təkrarlanma öyrənmə alqoritmlərində tezliyi çox olan yazılara doğru meyletməyə səbəb olur və daha zərərli ola bilən az tezlikli məlumatlardan öyrənmənin qarşısını alır. Test verilənlərində təkrarlanma isə daha yüksək aşkarlanma faizi olan yüksək tezlikli məlumatlara doğru meyletməyə səbəb olacaq.

5) Məlumatların çətinlik səviyyəsi aşağıdır. Ədəbiyyatda məlumat verilənlərin aşkarlanma metodlarının dəqiqliyi 98%, səhv aşkarlanma faizi isə 1% səviyyəsindədir. Lakin praktiki İDS sistemlərində bu metodlar istifadə edilmir, hesab edilir ki, onlar hələlik mükəmməl texnologiyalar deyil.

KDD’99 verilənlər toplusu da DARPA’98-də toplanmış verilənlərdən əldə edilmişdir və eyni problemlər qalır. Bu problemləri aradan qaldırmaq məqsədilə NSL_KDD verilənlər toplusu KDD`99-dan seçilmiş verilənlərdən təşkil edilib [14]:

- Təlim toplusunda izafi yazılar yoxdur.
- Test toplusunda təkrar yazılar yoxdur.
- Hər bir qrupda seçilmiş yazıların sayı ilkin KDD toplusundakı yazıların faiz payına tərs mütənasibdir.

CAIDA (Center for Applied Internet Data Analysis) [7] müxtəlif növ verilənlər toplayır və onları elmi-tədqiqatlar üçün paylaşır. CAIDA verilənlər toplularının əksəriyyəti xüsusi hadisələrlə və hücumlarla əlaqəlidir. Bu verilənlərin bir çoxu anonim magistral trafikidir və onlardan faydalı yük, bəzən protokol məlumatları, gediş ünvanı və s. tamamilə silinib. Verilənlər toplularının bu nöqsanları onların etalon test verilənləri kimi effektiv istifadəsinə təsir edir.

DEFCON verilənlər toplusu CTF (Capture The Flag) yarışları zamanı yığılmış trafikdir, bu verilənlər real şəbəkə verilənlərindən çox fərqlidir, çünki normal fon trafikindən fərqli olaraq əsasən müdaxilə trafikindən ibarətdir. Bu nöqsanına görə DEFCON verilənləri adətən həyəcan siqnallarının korrelyasiyası üsullarının qiymətləndirilməsi üçün istifadə edilir [8].

ISCX 2012 IDS Data Set – istifadəçilərin davranışlarını imitasiya edən real trafik (SSH, HTTP, SMTP) generasiya edən real qurğulardan istifadə edilməklə toplanmışdır [11].

Profil anlayışına müdaxilələrin ətraflı təsviri və tətbiqi proqramların, protokolların və aşağı səviyyə şəbəkə subyektlərinin abstrakt paylanma modelləri daxildir. Real trafik analiz edilərək trafik generasiya edən agentlərin profilləri yaradılmışdır. Real verilənlər generasiya etmək üçün qaydalar müəyyən edilmiş və profillə birlikdə istifadə edilmişdir. Bu qaydalar verilənlər toplusunun reallığı, qiymətləndirmə imkanları, tamlıq, ümumi toplama və zərərli fəaliyyət baxımdan effektivliyi üçün olduqca vacibdir. Daha sonra profillərdən istifadə edilərək verilənlər generasiya edilmişdir. Verilənlər toplusunda anomal fəaliyyətə aid verilənləri generasiya etmək üçün müxtəlif çoxmərhələli hücum ssenariləri yerinə yetirilmişdir.

III. PREDICT VERİLƏNLƏR TOPLUSU

PREDICT (Protected Repository for the Defense of Infrastructure Against Cyber Threats) ABŞ Milli Təhlükəsizlik Nazirliyi tərəfindən dəstəklənir, 2008-ci ildən fəaliyyətdədir və şəbəkə təhlükəsizliyi üzrə tədqiqatları inkişaf etdirmək üçün real verilənlərin paylaşımını həyata keçirir.

PREDICT ölkənin informasiya infrastrukturuna olan kiber-təhdidləri qiymətləndirmək və ölkənin kiber təhlükəsizlik potensialını artırmaq üçün yeni modellərin, texnologiyaların və məhsulların yaradılması üçün tədqiqat verilənləri bazasıdır (repozitarisidir).

Verilənlər provayderi (Data Provider) və verilənlər hostu (Data Host) anlayışları istifadə olunur. Verilənlər provayderi özlərinin sahibi olduqları və ya nəzarət etdikləri verilənləri PREDICT vasitəsilə icazə verilmiş tədqiqatçılara əlyetər edir. Verilənlər hostu PREDICT verilənlər toplularını saxlamaq və onların icazə verilmiş tədqiqatçılara ötürülməsini koordinasiya etmək üçün hesablama infrastrukturu təqdim edən təşkilatdır.

PREDICT verilənlər toplularının provayderləri Kolorado Dövlət Universiteti, Miçiqan Universiteti, Viskonzin Universiteti, Cənubi Kaliforniya Universiteti – İnformasiya Elmləri İnstitutu (USC–ISI), Kaliforniya Universiteti San-Diyeqo (UCSD), Corciya Texnologiya İnstitutu (Georgia Tech), Merit Networks, SKAION və Packet Clearing House-dur.

PREDICT repozitarisində hazırda 526 verilənlər toplusu var. Toplama müddətləri bir neçə saatdan günlərə və aylara, verilənlərin həcmi isə baytlardan terabaytlara (TB) çatır. Cədvəl 1-də toplanmış verilənlərin kateqoriyalarına və həcmlərinə aid məlumatlar verilir.

CƏDVƏL 1. PREDICT TOPLUSUNA DAXİL OLAN BƏZİ VERİLƏNLƏR

Verilənlər Hostu/Provayderi	**Verilənlərin kateqoriyası**	**Verilənlərin həcmi**
UCSD/CAIDA	Internet Topology Data	1612.1 TB
UCSD/CAIDA	Blackhole ünvan fəzası verilənləri	2113.1 TB
USC-ISI	Trafik axını, Internet topologiya, Ünvan paylanması verilənləri	42 TB
Kolorado Dövlət Universiteti	Trafik axını verilənləri, Spam loqları, IP Reputasiya siyahıları	90 TB
Miçiqan Universiteti	Tətbiqi proqramlar səviyyəsi təhlükəsizlik verilənləri	120.0 TB
Merit Networks	Trafik axını verilənləri	1400.0 TB
Georgia Tech	Sinkhoul verilənləri	0.01 TB
Viskonzin Universiteti	IDS və şəbəkə ekranı verilənləri	2.0 TB
Packet Clearing House	Sintetik verilənlər	7083.4 TB
SKAION	Sintetik verilənlər	119.2 TB

PREDICT paylanmış verilənlər repozitarisidir, onun verilənləri Verilənlər provayderləri tərəfindən təqdim olunur və verilənlər hostlarında saxlanır və müraciət edilir. PREDICT Koordinasiya Mərkəzi (PKM) verilənlərin repozitariyə daxil edilməsini idarə edir, bu verilənlərə müraciət xahişlərini nəzərdən keçirir və icazə verir. PKM verilənlərin tamlığını və konfidensiallığını, eləcə də müvafiq məqsədlərdə istifadəsini təmin edəcək protokollarla işləyir.

Hazırda toplanmış verilənlər kateqoriyasına aşağıdakılar daxildir: ünvan fəzasının paylanması, BGP marşrutlama verilənləri, Blackhole ünvan fəzası, DNS, IDS və şəbəkələrarası ekran verilənləri, infrastruktur, Internet topologiyası, IP paket başlıqları, sintetik genereasiya edilmiş verilənlər, trafik axını, e-poçt spamı. PREDICT-ə təqdim edilən hər bir verilənlər toplusu korporativ siyasətə və öhdəliklərə uyğunluğu təmin etmək üçün verilənlər provayderinin hüquqi və etik şurası tərəfindən baxılmalı və icazə verilməlidir.

Repozitaridəki verilənlər toplusu üçün qeyri-məhdud, kvazi-məhdud, və məhdud giriş hüquqları müəyyən olunur (qeyri-kommersiya məqsədləri üçün qeyri-məhdud və ya kvazi-məhdud hüquqları da vardır). Məhdud giriş sinfinə aid verilənlərə giriş əldə etmək üçün yazılı razılaşma memorandumu tələb edilir:

PREDICT portalı (www.predict.org) bu repozitariyə giriş nöqtəsidir və əlyetər verilənlər toplularını təsvir edən kataloqa malikdir. Lakin verilənlərə müraciət xahişi göndərmək üçün portalda istifadəçi hesabı yaratmaq tələb edilir.

İstifadəçi hesabı yaradılması üçün bütün müraciətlərə PKM tərəfindən baxılır və icazə verilir; lakin digər ölkələrdən olan tədqiqatçılara hesab açılmasına icazə verilməsi prosesində müraciətin edildiyi ölkədə olan PREDICT koordinatorunun rəyi və icazəsi tələb edilir. Verilənlər toplusundan istifadə etmək üçün kibertəhlükəsizlik sahəsində tədqiqat aparan ABŞ və bir neçə ölkə (Avstraliya, Birləşmiş Krallıq, Kanada, İsrail və Yaponiya) tədqiqatçıları müraciət edə bilərlər.

IV.WINE VERİLƏNLƏR TOPLUSU

Zərərli proqramların aşkarlanması sahəsində tədqiqatlar üçün yararlı olan real iş mühitində toplanmış verilənlər toplusuna nümunə Symantec Research Labs tərəfindən nəzarət edilən WINE toplusudur [15]. WINE verilənlər toplusuna zərərli proqram nümunələri, binar reputasiya, URL-reputasiya, e-poçt spamı, anti-virus telemetriya məlumatları daxildir (Cədvəl 2). Zərərli proqramlar kolleksiyasına siqnaturalarının yaradılmasında istifadə edilmiş zərərli proqram nümunələri (virus, soxulcan, bot və s.) daxildir. Binar reputasiya verilənləri naməlum binar fayllar haqqında məlumat təmin edir (məsələn, antivirus siqnatutası hələ yaradılmamış fayllar), bu fayllar Symantec reputasiya əsasında təhlükəsizlik proqramında iştirak edən istifadəçilər tərəfindən göndərilir. Anti-virus telemetriyası Symantec tərəfindən siqnaturası yaradılmış və antivirusların aşkarladığı məlum təhdidlərin baş verməsini də qeydə alır. WINE verilənlər toplusu informasiyanın nə zaman, harada və necə toplandığını göstərən metaverilənləri də müəyyən edir.

CƏDVƏL 2: WINE VERİLƏNLƏRİ

Verilənlər toplusu	**Mənbələr**
Binar reputasiya	50 milyon kompüter
Anti-virus telemetriyası	130 milyon kompüter
E-poçt spamı	2.5 milyon tələ-hesab
URL reputasiya	10 milyon domen
Zərərli proqram nümunələri	200 ölkə

WINE giriş modelində tədqiqatçı verilənlər toplusu kataloquna baxır və verilənlərə giriş üçün sorğu hazırlayır. Sorğuların doğruluğu və soruşulan verilənlərin əlyetərliyi daxili və xarici tədqiqatçıların məsləhət şurası tərəfindən qiymətləndirilir. Verilənlərə giriş icazəsi verilmiş tədqiqatçılar WINE infrastrukturu olan laboratoriyaya gəlməli və tədqiqatlarını orada aparmalıdır. WINE istifadə edilməklə yaradılan intellektual mülkiyyət tədqiqatçıların özlərinə məxsusdur və onlar öz nəticələrini məsuliyyətlə çap etmələri təşviq edilir.

V. BOTNET TRAFİKLƏRİ ÜZRƏ VERİLƏNLƏR

Botnetlərin aşkarlama metodlarının qiymətləndirilməsi real trafiki qəbul edilən səviyyədə imitasiya edən kifayət qədər qeyri-bircins olan verilənlərlə eksperimentlərin aparılmasını tələb edir. Botnet detektorlarının praktikada effektivliyi onların qiymətləndirilməsi zamanı istifadə edilən reallığa yaxın botnet trafikindən çox asılıdır. Belə verilənlərin yoxluğu botnetlərin aşkarlanmasını qiymətləndirmək üçün ədəbiyyatda dəfələrlə vurğulanmış problemlər səbəbindəndir [16,17].

Botnetlərin qiymətləndirilməsi üçün istifadə edilən verilənlər toplusu aşağıdakı problemlər nəzərə alınmaqla yaradılmalıdır [17]:

Ümumilik. Mövcud botnet verilənləri toplularında yalnız bir neçə (adətən 2-3) botnetdən götürülmüş verilənlər olur. Təbiətcə məhdud olan bu yanaşmalar qeyri-pratik və yeni təhdidlərə qarşı qeyri-dəqiqdir (belə verilənlərlə yaradılmış detektorlar çox spesifik olan botnet davranışlarının çox az sayda xarakteristikasını əks etdirir).

Realizm. Botnet trafiki nəzarət edilən mühitdə generasiya edilir və ya toplanır. Botnetin nəzərdə tutulmuş bütün zərərli fəaliyyətini yerinə yetirməsi üçün yararlı mühit (botnet tərəfindən aşkarlanmayan) təmin etmək asan məsələ deyil. Eyni zamanda, toplama müddəti kifayət qədər uzun olmalıdır ki, “yuxuya getmiş” botnetlər də öz funksiyalarını nümayiş etdirsinlər.

Reprezentivlik. Botnet verilənlərini toplanması zamanı daha bir problem detektorun praktiki istismarı zamanı qarşılaşacağı real mühiti əks etdirən şəbəkə trafikini toplamaq imkanıdır. Gizlilik tələbləri səbəbindən real istismar mühitində olan fon trafikini toplamaq bir çox halda mümkün deyil, nəticədə trafik ya imitasiya edilir, ya da nəzarət edilən mühitdə toplanılır.

UNB ISCX Botnet verilənləri toplusu yuxarıdakı meyarlar nəzərə alınmaqla yaradılıb və topluya aşağıdakılar daxildir:

- ISOT verilənlər toplusu müxtəlif verilənlər topluları birləşdirməklə yaradılıb [17]. Topluda həm zərərli (Storm və Zeus botnetlərindən), həm də zərərli olmayan trafik (HTTP trafiki, bittorrent) vardır.
- Malware Capture Facility Project tərəfindən generasiya edilmiş botnet trafiki [18] – uzunmüddətli botnet trafikinin generasiyası və toplanması üzrə tədqiqat layihəsidir. Bu topludan təlim verilənləri üçün 4 botnet izi (Neris, Rbot, Virut və NSIS) və test verilənləri üçün (Neris, Rbot, Virut, NSIS, Menti, Sogon və Murla) doqquz botnet izi seçilmişdir.

Bu botnet izlərini vahid topluda birləşdirmək üçün overley metodologiyası istifadə edilmişdir, sintetik verilənlər toplusunun yaradılması üçün ən populyar metoddur [17]. Zərərli verilənlər adətən honeypotlar və ya botnet binar kodu ilə yoluxmuş kompüterlər vasitəsilə nəzarət edilən mühitdə toplanılır. Botnet izlərini normal verilənlərlə birləşdirmək üçün zərərli verilənlər ya daxili, ya da xarici şəbəkədə mövcud olan kompüterlərə inikas etdirilir. Sonra TCPReplay istifadə edilməklə zərərli və normal trafik qarışdırılıb və TCPdump vasitəsilə vahid topluda toplanılıb.

NƏTİCƏ

Bu işdə informasiya təhlükəsizliyi üzrə tədqiqatlarda istifadə edilən verilənlər topluları analiz edilmişdir. Analizin nəticələrinə görə, bəzi topluların köhnəlmiş olmasına və eksperimentlərin nəticələrinə təsir edən nöqsanlara malik olmalarına baxmayaraq, yenə də istifadə edilir. Bir sıra mühüm verilənlər toplusuna isə yalnız müəyyən ölkə tədqiqatçılarının giriş imkanları vardır. Daha bir müşahidəyə görə, qabaqcıl tədqiqat qrupları bir qayda olaraq, verilənlərin toplanmasına və nişanlanmasına da yetərli resurslar ayırırlar.

ƏDƏBİYYAT

[1] S.E. Coull, E. Kenneally, “A qualitative risk assessment framework for sharing computer network data,” 2012 TRPC. http://dx.doi.org/10.2139/ssrn.2032315

[2] J.Camp et al. Data for Cybersecurity Research: Process and Wish List. NSF Workshop on Cyber Security Data for Experimentation, 2010. http://www.cc.gatech.edu/fac/feamster/papers/data-wishlist.pdf

[3] Lincoln Laboratory, M.I.T., 2011. DARPA Intrusion Detection Evaluation. http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html.

[4] University of California, 2011. KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[5] Lawrence Berkeley National Laboratory, 2010. The Internet Traffic Archive. http://ita.ee.lbl.gov/index.html..

[6] Lawrence Berkeley National Laboratory and ICSI, LBNL/ICSI Enterprise Tracing Project. www.icir.org/enterprise-tracing/.

[7] CAIDA Data. http://www.caida.org/data/

[8] The Shmoo Group, 2011. Defcon. http://cctf.shmoo.com/.

[9] B. Sangster, T.J. O'Connor, T. Cook, R. Fanelli, E. Dean, J. Adams, C. Morrell, and G. Conti, "Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets;" USENIX Security's Workshop on Cyber Security Experimentation and Test (CSET), 2009.

[10] PREDICT, 2011. http://www.predict.org/.

[11] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, “Towards developing a systematic approach to generate benchmark datasets for intrusion detection,” Computers & Security, Vol. 31, No. 3, pp. 357–374, 2012.

[12] J. McHugh, “Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system valuations as performed by Lincoln Laboratory,”ACM Transactions on Information and System Security, Vol. 3, No.4, pp. 262–294, November 2000.

[13] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Toward credible evaluation of anomaly-based intrusion-detection methods,” Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, Vol. 40, No. 5, pp. 516–524, 2010.

[14] S. Revathi, A. Malathi “A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection,” International Journal of Engineering Research & Technology (IJERT), Vol. 2, Issue 12, pp. 1848-1853, December 2013.

[15] T. Dumitras, D. Shou, “Toward a standard benchmark for computer security research: The Worldwide Intelligence Network Environment (WINE),” Proc. of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, pp. 89-96, 2011.

[16] A. J. Aviv and A. Haeberlen, “Challenges in experimenting with botnet detection systems,” in USENIX 4th CSET Workshop, 2011.

[17] D. Zhao, I. Traore, B. Sayed, W. Lu, S. Saad, A. Ghorbani, and D. Garant,“Botnet detection based on traffic behavior analysis and flow intervals,”Computers & Security, 2013.

[18] S.García, (2013). Malware Capture Facility Project. CVUT University. Retrieved February 03, 2013, from https://agents.fel.cvut.cz/malware-capture-facility.

Big Data və Fərdi Məlumatların Qorunması Problemləri

Rasim Mahmudov
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
rasimmahmudov@gmail.com

***Xülasə* - Məqalədə Big Data texnologiyalarının tətbiqi şəraitində fərdi məlumatların qorunması problemləri araşdırılır. Fərdi məlumatların toplanması və emalı ilə bağlı hüquq normalarının təkmilləşdirilməsi zərurəti göstərilir. Fərdi məlumatlara mülkiyyət hüququnun tanınması probleminə toxunulur. Fərdi məlumatların şəxssizləşdirilməsi, onların emalının legitimləşdirilməsi və məhdudlaşdırılması ilə bağlı hüquqi konsepsiyalarla Big Data texnologiyalarının tətbiqi konsepsiyaları arasındakı ziddiyyətlər şərh olunur.**

Açar sözlər* - *Big Data texnologiyaları, fərdi məlumatlar, fərdi məlumatların şəxssizləşdirilməsi, məlumatların emalının legitimləşdirilməsi, məlumatların emalının məhdudlaşdırılması.

I. GİRİŞ

Müasir dövrdə bir çox fəaliyyət sahələrində müvəffəqiyyət qazanmağın əsas amillərindən biri böyük həcmdə verilənlər kütləsini və informasiya axınlarını emal etmək imkanı ilə bağlıdır. Verilənləri analiz etməklə yeni xidmətlər və məhsullar yaratmaq, biznesi optimallaşdırmaq, bununla da, yeni gəlir mənbələri əldə etmək, mövcud gəlirləri artıqmaq olar. *Big Data* texnologiyalarının timsalında çox böyük həcmdə informasiyanı toplamaq, emal etmək, saxlamaq və analiz etmək üçün yeni, unikal bir vasitə meydana çıxıb.

Lakin *Big Data*-nın mühüm üstünlükləri ilə yanaşı, bir sıra problemləri də mövcuddur. Xüsusilə, müvafiq texnologiyalardan istifadə ilə bağlı bir sıra hüquqi problemlər meydana çıxır. O cümlədən intellektual mülkiyyət hüquqlarının qorunması, böyük verilənlərlə işləmək üçün lisenziyalaşdırma və müqavilələrin bağlanması, verilənlərin əlyetərliliyi və onlardan istifadə, şəxsi həyatın toxunulmazlığı ilə bağlı problemlər elmi aktuallıq kəsb edir.

Big Data ilə bağlı etimad, diskriminasiya, əlyetərliliyin qeyri-bərabərliyi, konfidensiallıq, istismar və manipulyasiya kimi bir sıra sosial və etik məsələlər böyük narahatlıqlar doğurur. Bu məsələlər *Big Data* obyekti olan məlumatların əsas sosial və etik dəyərlər sırasına aid edilməsinin qəbul olunmasını tələb edir.

Eyni zamanda, fərdi məlumatların qorunmasını nəzərdə tutan bir sıra hüquq normaları *Big Data* texnologiyalarının effektiv tətbiqi ilə bağlı problemlər yaradır. Bundan başqa, müvafiq hüquq sistemi sürətlə inkişaf edən informasiya texnologiyalarının artan imkanlarına adekvat reaksiya verə bilmir.

II. FƏRDİ MƏLUMATLARIN TOPLANMASI VƏ EMALI İLƏ BAĞLI PROBLEMLƏR

Müasir informasiya texnologiyaları şəxsi həyat anlayışı ilə bağlı təsəvvürlərdə əhəmiyyətli dəyişikliklərə səbəb oldu. Gerçək dünyada baş verən hadisə və proseslər - əmtəə və xidmətlərin əldə edilməsi, yaxın adamlarla ünsiyyət, dövlət orqanları ilə qarşılıqlı əlaqə və s. virtual mühitə keçərək bir qədər fərqli mahiyyət daşımağa başladı. Bunun nəticəsində insanların bilərəkdən, yaxud bilməyərəkdən açıqladığı və müxtəlif təşkilatların hədəfinə çevrilən fərdi məlumatlar misli görünməmiş miqyas aldı.

Xüsusilə, İnternet kütləvi sosial hadisəyə, cəmiyyət həyatının ayrılmaz hissəsinə çevrildikdən sonra insanlar bu texnologiyanın onların şəxsi həyatına təsirindən ehtiyat etməyə başladılar. Bu gün insanlar, narahatdırlar ki, onların sağlamlığı, maliyyə vəziyyəti, şəxsi münasibətləri, siyasi baxışları və s. ilə bağlı kimlərdəsə məlumatlar toplanır.

Amerika alimi R.Posner hesab edir ki, insanlar fərdi məlumatlarının tanımadıqları adamın əlinə keçməsini, başqalarının onları çılpaq vəziyyətdə görməsi situasiyası ilə eyniləşdirirlər [1].

Vətəndaşların narahatlığına səbəb olan fərdi məlumatlardan istifadə daha çox özəl sektorda müşahidə olunur. Lakin dövlət orqanları da fərdi məlumatları toplayır, mühafizə edir və müəyyən məqsədlər üçün analiz edir və nəticələrini tətbiq edir. Fərdi məlumatlardan istifadə zamanı şəxsi həyatın uğurla qorunması modeli hüquq sisteminə olan etimadla bağlıdır. Əgər bu şərt ödənilərsə, şəxsi həyatın qorunması modeli yalnız qanunun aliliyini möhkəmləndirər və dövlətin məsuliyyətini artırar.

Əgər tənzimləyicilər yalnız fərdi məlumatların toplanması ilə deyil, həm də onların istifadəsi ilə bağlı vətəndaşların qanuni mənafelərinin qorunmasına səy göstərərlərsə, onda insanlar daha arxayın olarlar.

Lakin bu narahatlıqlar texnoloji prosesləri əngəlləmək iqtidarında deyil. Hazırda fərdi məlumatların kütləvi şəkildə toplanması və saxlanması qaçılmazdır. Bu gün insanlar şüurlu şəkildə bir çox tşkilatlara, o cümlədən dövlər orqanlarına, İnternet-provayderlərə, səhiyyə müəssisələrinə, telekommunika-siya şirkətlərinə, maliyyə qurumlarına böyük həcmdə informasiya verir. İndi insan fəaliyyətinin elə bir sahəsini təsəvvür etmək olmaz ki, orada fərdi məlumatlar

lazımsız məhsul kimi zibil qutusuna atılsın. İnsanlar bilmirlər ki, onların məlumatları harada və necə saxlanılır, onlardan nə üçün istifadə edilir. Hər gün fərdi məlumatları analiz etmək və müəyyən nəticələr almaq üçün daha güclü prosessorlar və serverlər meydan gəlir [2].

Hazırkı qanun və normalar, əsasən, fərdi məlumatların toplanması və mühafizəsinə nəzarət edilməsinə yönəlib. Amma həmin məlumatlardan necə istifadə edilməsi məsələsinə diqqət yetirilmir .

Microsoft şirkətinin tədqiqat və strategiya üzrə keçmiş direktoru Kreyq Mandi hesab edir ki, fərdi məlumatlardan qanunsuz istifadə edən ağır cinayət əməli kimi dəyərləndirilməlidir [3].

Gartner analitik şirkətinin proqnozuna görə, işgüzar etikanın pozulması hallarının yarıya qədəri məhz *Big Data* analizi nəticəsində əldə edilən informasiyanın düzgün istifadə edilməməsi ilə bağlı olacaq. Bunun nəticəsində müvafiq müəssisələr reputasiyalarını, resurslarını, rəqabət qabiliyyətlərini itirə və hüquqi sanksiyalara məruz qala bilərlər [4].

Fərdi məlumatların qorunması ilə bağlı əsas problemlərdən biri odur ki, yaxın keçmişdə müvafiq məlumatların toplanması əlavə zəhmət tələb edirdi. İndi isə istifadəçilərin İnternetlə və digər kommunikasiya qurğuları ilə qarşılıqlı əlaqə prosesinin özü informasiyanı meydana çıxarır. Ona görə də şəxsi həyatın qorunmasının mövcud metodları artıq öz effektivliyini itirib və yeni metodlar işləyib hazırlamaq zərurəti yaranıb.

Almaniyanın məşhur *SAP* şirkətinin yaradıcılarından olan Hasso Plattner hesab edir ki, hakimiyyət orqanları vətəndaşların şəxsi həyat sirrinin mühafizəsini və insanlar haqqında informasiyanın toplanması və emalı üzrə ən müasir texniki imkanları nəzərə alan yeni qaydalar işləyib hazırlamalıdırlar. Onun fikrincə, *Big Data*-nın üstünlükləri mənfi tərəflərindən daha çoxdur. Ona görə də bu qənaətə gəlir ki, fərdi məlumatların mühafizəsi təmin edilməlidir, amma bu, *Big Data* texnologiyalarının tətbiqinin məhdudlaşdırılması hesabına olmamalıdır [5].

2014-cü ilin oktyabrında Mavriki adasında məlumatların və şəxsi həyatın qorunması məsələləri ilə bağlı beynəlxalq konfrans keçirilmişdir [6]. Konfransın məqsədi *Big Data*-nın və Əşyaların İnternetinin insanların gündəlik həyatına müsbət və mənfi təsirlərini müzakirə etmək, məlumatların toplanması və istifadəsi ilə bağlı risklərin azaldılması üçün prinsiplərin müəyyən edilməsi və tövsiyələrin hazırlanması idi.

Konfransda bəyan olunmuşdur ki, Əşyaların İnterneti qurğuları *Big Data* ilə birlikdə insan həyatını asanlaşdırır, amma bu texnologiyaların vətəndaşların şəxsi həyatına müdaxiləsi və onların hüquqlarının pozulaması ilə bağlı böyük təhlükələri mövcuddur.

Konfransın yekun bəyannaməsində *Big Data* istifadəçilərinə aşağıdakılar tövsiyə olunur:

- hansi məlumatların toplandığı, onların necə emal olunduğu, hansı məqsədlər üçün istifadə ediləcəyi, üçüncü şəxlərə verilib-verilməyəcəyi ilə bağlı məsələlər şəffaf olmalıdır;
- informasiyanın toplanmasının məqsədi bilavasitə onu toplayan zaman müəyyən olunmalıdır və informasiyanın təyin olunmuş məqsəddən kənar istifadəsi istisna edilməlidir;
- informasiyadan istifadə üçün razılıq alınmalıdır;
- informasiya qeyd olunan qanuni məqsədlərin reallaşdırılması üçün zəruri olan sayda toplanmalı və saxlanmalıdır;
- fərdi məlumatları əldə edilən şəxslər onlardan necə istifadə olunduğu barədə məlumatlandırılmalıdır;
- insanlara öz informasiyasına düzəliş etmək və onu idarə etməyə imkan yaradılmalıdır;
- informasiyanın toplanmasının insanların şəxsi həyatına necə təsir etdiyi analiz edilməlidir.

III. FƏRDİ MƏLUMATLARA MÜLKİYYƏT HÜQUQUNUN TANINMASI PROBLEMLƏRİ

Virtual dünyada insanların gündəlik həyatını əks etdirən informasiya bu gün tədqiqat üçün faydalı resurs hesab olunur. Öz növbəsində, hüquq normaları ilə bağlı daim çətinliklər meydana çıxır. Çünki bu tip informasiyanın onu generasiya edənə, təqdim edənə, toplayana, analiz edənə, zənginləşdirənə, satın alana, yaxud bütövlükdə cəmiyyətə məxsus olması hüquqi baxımdan öz həllini tapmayıb. Problemi daha da çətinləşdirən odur ki, informasiyaya eyni zamanda həm mülkiyyət, həm də bu və ya digər qanundan asılı olmayan müstəqil məlumat kimi baxıla bilər [7].

İnformasiyanın müasir texnologiyaların inkişafı ilə əlaqədar meydana çıxan yeni növləri mülkiyyət haqqında ənənvi qanunvericilikdə öz hüquqi statusunu almayıb, amma müəssisələrdə və ya inzibati orqanlarda generasiya olunur və mühafizə edilir. Belə bir vəziyyət, təbii ki, hüquqi baxımdan bir sıra problemlər yarada bilər. Ona görə də qanunvericilik sisteminin bu cür texnoloji yeniliklərə adekvat reaksiya verməsi zəruridir.

IV. FƏRDİ MƏLUMATLARIN ŞƏXSSİSZLƏŞDİRİLMƏSİ PROBLEMLƏRİ

Fərdi məlumatların şəxssizləşdirilməsi onların vətəndaşlara vurula biləcək ziyan riskinin minimallaşdırılması üsullarından biridir. Bu üsul insanın şəxsiyyətinin avtonomluğunu təmin etməklə mühüm sosial funksiyanı yerinə yetirir [8]. Bu üsul əlavə informasiya olmadan fərdi məlumatın kimə aid olduğunu müəyyən etməyin qarşısını alır. İnformasiyanın şəxssizləşdirilməsi metodları arasında aşağıdakıları qeyd etmək olar [9]:

- identifikatorların tətbiq edilməsi;
- tərkibin və ya semantikanın dəyişdirilməsi;
- dekompozisiya;
- qarışdırma;
- və s.

Konkret metod fərdi məlumatların emalının məqsəd və vəzifələrindən asılı olaraq seçilir. Bu zaman nəzərə alınır ki, fərdi məlumatların şəxssizləşdirilməsi təkcə onlardan qanunsuz istifadənin qarşısını almalı deyil, həm də onların emalını təmin etməlidir.

Lakin *Big Data* erasında bir neçə informasiya fraqmenti arasında korrelyasiyanın müəyən edilməsi vasitəsi ilə şəxsiyyətin identifikasiya edilməsinin mümkünlüyü həmin üsulun effektivliyini şübhə altına alır. Başqa sözlə, informasiyanın şəxssizləşdirilməsi artıq yeni texnoloji reallıqda şəxsi həyatı qorumaq iqtidarında deyil [10].

İnternetin meydana gəlməsi ilə bu cür identifikasiya daha da asanlaşmışdır. Çünki hər bir İnternet istifadəçisi onu identifikasiya etmək üçün kifayət qədər "rəqəmsal iz" buraxır.

Belə bir sual da ortaya çıxır ki, şəxssizləşdirilmiş məlumatlar fərdi məlumat kateqoriyasına daxil edilir, yoxsa fərdi məlumat rejiminin tətbiq edilmədiyi xüsusi məlumat növünə aid edilir? Avrop Şurasının fərdi məlumatlar üzrə işçi qrupu belə bir nəticəyə gəlib ki, əgər şəxssizləşdirilmiş məlumatlar ilkin vəziyyətinə dönmə qabiliyyətinə malikdirsə, onda fərdi məlumatlara aid edilməlidir. Nəzərə almaq lazımdır ki, yuxarıda göstərilən şəxssizləşdirmə metodlarının əksəriyyəti məhz dönmə qabiliyyətinə malikdir [8].

V. FƏRDİ MƏLUMATLARIN EMALININ LEGİTİMLƏŞDİRİLMƏSİ PROBLEMLƏRİ

Fərdi məlumatlarla bağlı hüquqi tənzimləmənin ideal modeli nəzərdə tutur ki, fərdi məlumatların subyekti onun informasiya sferasına (informasiyanın emalı, müvafiq risk və faydaları dair müstəqil qərar qəbul etmək imkanına malikdir. Bununla bağlı fərdi məlumatların subyektinin razılığı onların emalı üçün əsas hüquqi əsas sayılır. Fərdi məlumat subyektinin razılığının informativliyə, konkretliyə, şüurluluğa əsaslanması üçün ona fərdi məlumatlarının necə istifadə ediləcəyi: istifadənin məqsədi, əldə edilən məlumatların tərkibi və onların emal üsulları barədə ətraflı informasiya verilməlidir [11].

Lakin *Big Data* erasında fərdi məlumatların emalına dair məlumatlandırma razılığı aşağıdakı bir sıra səbəblərə görə öz effektivliyini əhəmiyyətli dərəcədə itirir [2]:

- fərdi məlumatların mümkün üsulları və emal məqsədləri haqqında müfəssəl məlumatların verilməsinin qeyri-mümkünlüyü;
- fərdi məlumat subyektinin bu cür informasiyanı adekvat şəkildə qavramasının qeyri-mümkünlüyü;
- müasir cəmiyyətdə fərdi məlumatların toplanması və emalı ilə məşğul olan böyük sayda qurumlarla fərdi qaydada qarşılıqlı əlaqə qurmağın qeyri-mümkünlüyü.

Beləliklə, məlumatlandırma razılığı fərdi məlumat subyektinin həmin məlumatların emalının ilkin mərhələsində bir sıra qərarlar qəbul etməsinin zəruriliyini nəzərdə tutur. Amma *Big Data* texnologiyalarının tətbiqinin xüsusiyyətləri bu cür qərarların nəticələrinin əvvəlcədən müəyyən etməyə imkan vermir. Ona görə də hesab etmək olar ki, məlumatlandırma razılaşması müasir şəraitdə fərdi məlumatların emalının legitimləşdirilməsi üçün hüquqi əsas sayıla bilməz.

VI. FƏRDİ MƏLUMATLARIN EMALININ MƏHDUDLAŞDIRILMASI PROBLEMLƏRİ

Fərdi məlumatların toplanması və emalının müəyyənliyi prinsipi şəxsin özü haqqında məhdud informasiya vermə hüququnun əsaslarından biridir. *Big Data* erasında isə məlumatların təkrar emalına üstünlük verilir. Belə ki, istisnasız olaraq bütün məlumatlar potensial dəyərə malikdir. Bu, eyni dərəcədə həm texniki, həm fərdi, həm də potensial olaraq fərdi məlumat kimi təsnifatlandırıla biləcək məlumatlara aiddir. Məsələn, müştərilərin alış-verişindən əldə edilən məlumatlar həm ünvanlı reklam göndərişi, həm bazarın tədqiqi, həm də müştərinin statusuna dair proqnozların hazırlanması üçün faydalıdır. Hər hansı şirkətdə nə qədər çox məlumat olarsa, *Big Data* texnologiyalarının tətbiqi üçün bir o qədər geniş meydan olar [11].

Big Data-nı okeandakı aysberqlə müqayisə etmək olar. Yəni ilk baxışda onun yalnız az bir hissəsini görmək mümkündür. İnnovasiya şirkətləri *Big Data*-nın görünməyən tərəflərini aşkara çıxarmaqla böyük faydalar əldə edirlər. O cümlədən müxtəlif məlumat bazalarını birləşdirməklə əlavə nəticələr qazanırlar.

Müasir texnologiyalar həmçinin fərdi məlumatların toplanması ilə bağlı olan bir sıra texniki məhdudiyyətləri aradan qaldırmaq iqtidarındadır. Belə ki, nəhəng informasiya massivlərinin saxlanmasının əhəmiyyətli dərəcədə ucuzlaşması əvvəlcədən müəyyən edilmiş məqsədlərlə fərdi məlumatların emalının məhdudlaşdırılması prinsipindən imtina etməyi stimullaşdırır [12].

Beləliklə, *Big Data* erasında müvafiq şirkətlər mümkün qədər daha çox məlumat toplamaqda maraqlıdır. Sözsüz ki, bu məlumatların arasında fərdi məlumatlar da mühüm yer tutacaq.

Bu səbəbdən də fərdi məlumatların emalı ilə bağlı milli qanunvericiliklərdəki mövcud məhdudiyyətlər, həmçinin müxtəlif məlumat bazalarının əvvəlcədən bəyan olunmuş bazalarla birləşdirilməsinin yolverilməzliyi və əvvəlcədən bəyan olunmuş emal məqsədlərinə uyğun olmaması müasir texnologiyaların mahiyyəti və biznes praktikası ilə ziddiyyət təşkil edir. Ona görə də *Big Data* erasında qanunvericiliyin müvafiq tələblərinə əməl etmək çox çətindir.

NƏTİCƏ

İnsanların etimadını qazanmaq, narahatlığına son qoymaq üçün fərdi məlumatların toplanması siyasəti şəffaf olmalıdır. Onlar informasiyanın hansı məqsədlə, necə toplandığını, nə cür istifadə edildiyini bilməlidirlər. Həmçinin insanlar özləri onlar haqqında toplanan və emal olunan informasiyadan istifadə etmək imkanı ilə təmin edilməlidirlər.

Eyni zamanda, fərdi məlumatların qorunması ilə bağlı bir sıra mövcud hüquq normaları *Big Data* texnologiyalarının

effektiv tətbiqində problemlər yaradır. Ona görə də bu texnologiyaların faydalarını nəzərə alaraq, qanunvericilik səviyyəsində şəxsi təhlükəsizlikə ictimai maraqlar arasında həssas balans yaratmaq lazımdır.

ƏDƏBİYYAT

[1] O.Tene, J. Polonetsky Big Data for All: Privacy and User Control in the Age of Analytics // Northwestern Journal of Technology and Intellectual Property, 2013, Volume 11, Issue 5,
http://scholarlycommons.law.northwestern.edu

[2] L.Bygrave Data Privacy Law: An International Perspective. Oxford University Press, 2014. 233 p.

[3] C.Mundie. Privacy Pragmatism. Focus on Data Use, Not Data Collection. Essay, March/April 2014, Issue,
https://www.foreignaffairs.com/articles/2014-02-12/privacy-pragmatism

[4] http://www.gartner.com/technology/topics/big-data.jsp

[5] http://www.blogs.saphana.com/2015/09/29/hasso-plattner-blog

[6] http://www.smartdatacollective.com

[7] Л.Джапек, Ф.Крёйтер, Мю.Берг и др. Отчёт AAPOR о больших данных. Американская ассоциация исследователей общественного мнения, Пер. с англ. М., 2015, 72 с.

[8] H.Nissenbaum A Contextual Approach to Privacy Online // Daedalus, 2011. No. 4, pp.32–48.

[9] Р.Жаринов, Ю.Трифонова Возможности обезличивания персональных данных в системах, использующих реляционные базы данных // Управление, вычислительная техника и информатика. № 2 (32), 2014, с. 188–194.

[10] P.Ohm Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization // UCLA Law Review No. 57. 2010, pp.1701–1777.

[11] И.Л.Бачило, Л.А.Сергиенко, Б.В.Кристальный, А.Г.Арешев Персональные данные в структуре информационных ресурсов. Основы правового регулирования. Минск: Беллитфонд, 2006. 474 с.

[12] J.Lane, V.Stodden, S.Bender, H.Nissenbaum Privacy, Big Data and the Public Good: Frameworks for Engagement. Cambridge University Press, 2014. 344 p.

Big Data və Təhlükəsizlik

Toğrul Qafarbəyli[1], Şəbnəm Ələkbərova[2]
[1,2]Azərbaycan Respublikası Rabitə və Yüksək Texnologiyalar Nazirliyi, Bakı, Azərbaycan
[1]*togrul_gafarbeyli@yahoo.com*

***Xülasə* - Bu məqalə günümüzdə İnternetin istifadəsi nəticəsində meydana çıxan ən böyük problemlərdən birinə - fərdi məlumatların qorunması problemlərinə həsr olunub.**

Açar sözlər - verilənlər, Big Data, təhlükəsizlik.

I. GİRİŞ

Big Data (Böyük verilənlər) texnologiyalarına tərəqqisi və istifadə sahələrinin artması nəticəsində tez-tez adını eşitdiyimiz və illərdir içərisində olduğumuz, lakin inkişafından heç də xəbərdar olmadığımız bir faktdır.

Texnologiyanın irəliləməsi və internetin inkişafı ilə birlikdə verilənlərin gücü də ön plana çıxdı ki, bununla birlikdə internet dünyasındakı bir çox fakt “məlumat tullantıları” olaraq xatırlanmağa başladı. Bu “tullantılardan” mənalı məlumatların da əldə olunacağını düşünən proqram şirkətləri araşdırmalarını bu yöndə apararaq, “Big Data” olaraq adlandırdığımız faktı ortaya çıxardılar [10].

Big Data (Böyük verilənlər) ictimai media paylaşımları, fotoşəkil arxivlərimiz, Veb serverlərin loqları, internet verilənləri, bloqlar, mikrobloqlar, iqlim verilənləri və bənzər məlumatlar, GSM operatorlarından əldə edilən axtarış qeydləri kimi müxtəlif mənbələrdən əldə etdiyimiz bütün bu məlumatların əhəmiyyətli və emal edilə bilər hala çevrilmiş formasıdır [8].

Big Data düzgün analiz metodlarından istifadə edildiyində şirkətlərin əhəmiyyətli strateji qərarlar qəbul etməsinə, riskləri daha yaxşı idarə etmələrinə və innovasiyalar yaratmalarına imkan yaradır.

Hazırda verilənlər bazası mütəxəssisləri mövcud məlumatları relyasyon verilənlər bazasında strukturlaşdırılmış şəkildə təsnifatlandırır [1] və şirkət rəhbərləri də bu məlumat bazalarındakı hesabatlar sayəsində qərarlar qəbul edirlər. İndiyə qədər bütün bu məlumatlar tullantı kimi qəbul olunurdu, çünki hesabatların verilənlər bazasında saxlanılması və istifadəsi çox çətin idi.

Qısaca olaraq, Big Data-nın məqsədi tullantılardan işimizə yaraya biləcək qiymətli məlumatlar ortaya çıxarmaqdır.

II. BİG DATA VƏ KOMPONENTLƏRİ

Kütləvi istifadə nəticəsində gündəlik həyatımızda internetdən əldə etdiyimiz xidmətlər, xüsusilə də satış sonrası müştəri məmnuniyyətini qoruyub saxlamaq üçün istifadəçi ilə bağlı məlumatların yaddaşa qeyd olunması vacib məsələ kimi ortaya çıxdı. Bu nəticə olaraq “Big Data” istifadəsini meydana gətirdi.

Big Data-nın meydana gəlməsində 5 komponent vardır. Bu komponentlər ingilis dilində sırasıyla - variety, velocity, volume, verification və value-dir ki, ümumi olaraq “5V” şəklində adlandırılmaqdadır.

Variety (müxtəliflik): Çıxarılan məlumatlar ümumi olaraq strukturlu olmadığı və bir çox fərqli mühitdən əldə edilən məlumatların formatlarından yarandığına görə bütöv və bir-birlərinə əlavə edilə biləcək halda olmalıdır.

Velocity (sürət): Big Data istehsalı hər keçən gün sürətinə sürət qatmaqda və bu məlumatlar saniyədə inanılmaz ölçülərə çatmaqdadır. Sürətlə böyüyən verilənlər, o məlumata möhtac olan əməliyyat sayının və müxtəlifliyinin də eyni sürətlə artması nəticəsini ortaya çıxarmaqdadır və nəticə olaraq, həm proqram təminatı, həm də xarici qurğular olaraq bu sıxlığa davam gətirməliyik.

Volume (Həcm): Big Data kimi adlandırdığımız məlumatların sürətlə artması gələcəkdə qarşılaşacağımız problemləri ön plana keçirərək, bu məlumat yığını ilə necə başa çıxacağımız və planlarımızı bu istiqamət üzrə həyata keçirəcəyimizi düşünmək məcburiyyətində qoyur.

IDC (International Data Corporation) statistikalarına görə 2020-ci ildə əldə olunacaq məlumat miqdarı 2009-cu ilə nisbətən 44 dəfə aratacaqdır [7]. 2010-cu ildə dünyadakı ümumi informasiya xərcləri ildə 5%, çıxarılan məlumat miqdarı isə 40% artır [11].

2012-ci il statistikasına görə dünyada hər gün 2.5 kentrilyon bayt məlumat əldə olunur [4].

Verification (doğrulama): Bu qədər sürətli böyüyən məlumatların axını zamanı gələn məlumatların etibarlı olub-olmadığına nəzarət etmək lazım olduğu hallarda da digər məlumatların komponenti olaraq verification (doğrulama) meydana çıxır. Bu verilənləri hər kəs görə bilməz və ya ümumiyyətlə gizli qala bilər.

Value (dəyər): Bəlkə də ən əhəmiyyətli qatlardan biri "dəyər" qatıdır ki, yuxarıdakı komponentlər vasitəsilə süzüldükdən sonra Big Data-nın istehsalı və emalı qatlarında əldə edilən məlumatlar bizim üçün əhəmiyyət kəsb etməyə başlayır.

III. BİG DATA VƏ TƏHLÜKƏSİZLİK

Bu gün tez-tez qarşımıza çıxan problemlərdən biri də korporativ məlumat bazalarının təhlükəsizliyinin təmin olunması sahəsində olan mövcud boşluqlardır. Məsələn, Home Depot, Target, Niemen Marcus və Ashley Madison kimi “protokollara” sahib data-kollektorların əksəriyyəti öz istifadəçilərinin şəxsi məlumatlarını qoruya biləcək səviyyədə deyillər [2].

Təhlükəsizlik protokoluna ciddi dəyişikliklər etmək qaçınılmazdır. Unutmayaq ki, Big Data inkişaf etdikcə hakerlər də öz işlərinə davam edirlər.

Şəbəkə ekranları və antivirus proqramları kimi ənənəvi mexanizmlər böyük məlumat bazalarının təhlükəsizliyinin təmin edilməsi üçün hər zaman yetərli olmur. Çoxtərəfli “giriş nöqtələri” və irihəcmli məlumatları özündə saxlayan bazaların təhlükəsizliyini təmin etmək üçün görülən işlər daha operativ və sürətli olmalıdır [5].

Could Security Alliance – bulud texnologiyalarının təhlükəsizliyini artırmaq, standardlar işləyib hazırlamaq sahəsində lider qeyri-kommersiya təşkilatıdır [6]. Bu təşkilat böyük həcmli məlumatların saxlanılması zamanı yaranan problemlər və həlli yolları haqqında aşağıdakı məlumatları verir:

Müəyyən proqram sistemlərində hesablamaların təhlükəsizliyi. Bir neçə mərhələdə hesablama həyata keçirən proqramların təhlükəsizliyi də bir neçə səviyyədə həyata keçirilməlidir.

Əlaqəli məlumat bazasının təhlükəsizliyi. Elmi dildə NoSQL adlanan bu məlumat bazası aktiv olaraq inkişaf etdiyindən müvafiq təhlükəsizlik tədbirləri də genişləndirilməlidir.

Məlumatların saxlanılmasının təhlükəsizliyi. Bu günə qədər məlumatların köçürülməsinə İT - menecerlər nəzarət edirdilər. Hazırda Big Data fərdi nəzarət sistemi öz iş qabiliyyətini itirib. Müxtəlif səviyyələr üzrə məlumatların ötürülməsi əlavə təhlükəsizlik tədbirlərini tələb edir.

Dəqiqliyin yoxlanılması. Sistemə milyonlarla məlumat daxil olduğu zaman hər bit informasiyanın dəqiqliyinin yoxlanılması vacibdir.

Real zamanda təhlükəsizlik üzrə monitorinq. Real zamanda təhlükəsizlik təhlükələrin aşkar olunmasında öz yüksək səviyyəsi ilə fəxr edə bilməz. Çünki gün ərzində sistemə daxil olan minlərlə yanlış məlumatlar nəticəsində mütəxəssislərin qarşısına çıxan problemlərin həllində çətinliklər yaradır.

Məxfiliyi qoruyan “data mining” və analitika. Big Data istifadəçinin razılığı və xəbəri olmadan şəxsi xarakterli məlumatların intensiv yığımı imkanıdır.

Rəqəmsal idarəetmə və təhlükəsizliyin təmini. Məlumatların tam təhlükəsizliyi üçün onlar əvvəldən axıra qədər kodlaşdırılmalıdır. Lakin eyni zamanda bu məlumatların oxunulması təmin olunmalıdır.

Fraqmentar idarəetmə. Hər məlumat konfidensial xarakter daşımır və şirkət bu məlumatları müxtəlif gizlilik səviyyələrində kodlaşdırmalıdır.

Dəqiq audit. Təhlükəsizlik sistemindəki qayda pozuntularını aşkar etmək üçün dəqiq auditə ehtiyac var. Lakin məlumatlar bazasının ölçülərini nəzərə alaraq, bu dəqiqlik baş verən insidentin ölçülərinə müvafiq olmalıdır.

Məlumatların mənşəyi. Məlumatların təhlükəsizliyini təmin edərkən əsas diqqət məlumatların ortaya çıxdığı mənbələrə yönəldilməlidir [12].

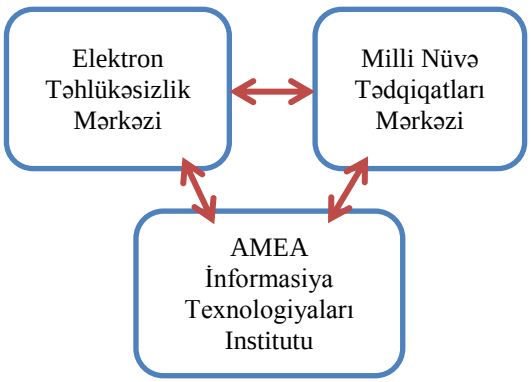
NƏTİCƏ

Yaşadığımız “böyük məlumatlar” çağında investorlar, texnologiya təşəbbüskarları, media və məsləhət şirkətləri Big Data mövzusuna xüsusi diqqət göstərərək yeni fürsətlər əldə etməyə çalışırlar.

Yaxın gələcəyin ən əhəmiyyətli texnologiya bazarının Big Data üzərində meydana gəlməsi ilə önümüzdəki beş il içərisində bu bazarın gəlirinin 50 milyard dolları aşması gözlənilir. Dünya səviyyəsində illik məlumatların həcmindəki artımın 59% və böyümənin artaraq davam etməsi gözlənilir [9]. Bu böyümənin mərkəzində həm ənənəvi, həm də yeni xəbər buraxılışları dayanır. IDC (International Data Corporation) rəqəmsal məlumatların bu ilin sonunda 1.2 milyon zetabayta (10^{21} bayt) çatacağını, önümüzdəki on il içində də 44 dəfə artacağını təxmin edir [3].

Beləliklə, böyük məlumatların saxlanılması və istifadəsi ilə eyni zamanda ortaya çıxan ən mühüm problem fərdi məlumatların məxfiliyinin qorunub saxlanılmasıdır. Bu sahə üzrə də araşdırmalar davam etməkdə və xüsusi bir ixtisas halına çevrilməkdədir.

Təklif olaraq - hazırda respublikamızda kiberhücum məlumatlarını analiz edərək, (xarici hücumlar və daxildən xaricə edilən hücumlar) kiberhücumlara qarşı tədbirlərin görülməsində və analitik təhlillərin aparılmasında Rabitə və Yüksək Texnologiyalar Nazirliyi tərkibində yaradılmış Elektron Təhlükəsizlik Mərkəzinin xüsusi rolu vardır. Elektron Təhlükəsizlik Mərkəzi tərkibində qurulan Big Data texnologiyası həmçinin Rabitə və Yüksək Texnologiyalar Nazirliyi nəzdində yaradılan Milli Nüvə Tədqiqatları Mərkəzi ilə birlikdə gələcəkdə nüvə sahəsində əldə olunacaq məlumatların təhlil olunması, nüvədən dinc məqsədlər üçün istifadə edilməsində yeni imkanlar yarada bilər. Təklifimizin davamı olaraq, unutmaq lazım deyil ki, adı çəkilən resurslardan istifadə edilməsi və alınan məlumatların emalı insan potensialına bağlıdır. Azərbaycan Milli Elmlər Akademiyasının İnformasiya Texnologiyaları İnstitutu illər ərzində yetişdirdiyi mütəxəssislər elm və texnologiya sahəsində birgə əməkdaşlığı zəruri edir.



Bu tipli üçbucağın yaradılması zəruridir. Belə ki, bu əməkdaşlıq respublikamızda Big Datanın inkişafına müsbət təsir göstərəcək, elm, səhiyyə, yüksək texnologiyalar,

informasiya texnologiyaları və kənd təsərrüfatı kimi sahələrdə də geniş tətbiq olunacaqdır.

Bütün bunları nəzərə alaraq xüsusi qeyd etmək istərdik ki, *Big Data gələcək deyil, artıq bu gündür.*

ƏDƏBİYYAT

[1] A. Barske, (2013-2014) Big Data Business Guide, https://www.academia.edu/6520364/Big_Data_Business_Guide_Complete_Big_Data_Overview_for_Enterprises_Example_of_Big_data_Projects/
[2] Big Data impact in the World.
[3] (25 February 2010). “Data, data everywhere”. The Economist.
[4] C. Göksu. İş analitiği çözümləri 2012 presentation. http://www.slideshare.net/CuneytGoksubig-data-nedir-ve-ibm-zmleri
[5] D. Kusnetzky, “What is “Big Data”?”. ZDNet^ Cukier, K.
[6] Y. Ersöz, “Bulut Barındırma”, TDD S.149.
[7] http://rtbinsight.ru/articles/big-data-security.html
[8] http://www.yac.mx/ru/mobile-security/android/where-big-data-and-security-align.html
[9] http://www.cisco.com/web/RU/news/releases/txt/2013/04/041913d.html
[10] http://www.osp.ru/iz/bigdata/articles/13044422
[11] http://www.wikiwand.com/en/Big_data
[12] https://en.wikipedia.org/wiki/Big_data

Biometrik Texnologiyalarda Big Data Problemləri

Şəfəqət Mahmudova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
shafagat_57@mail.ru

Xülasə **— Bu iş biometrik texnologiyalarda Big Data problemlərinə həsr olunmuşdur. Milli təhlükəsizlik, hüquq-mühafizə sisteminin və müdafiə orqanlarının qarşısında duran məsələlərin geniş spektrinin həllində Big Data texnologiyalarının istifadə edilməsinin əhəmiyyəti göstərilmişdir. Biometrik texnologiyalar sahəsində Big Data və buludların müasir texnologiyaları haqqında məlumat verilmişdir.**

Açar sözlər — biometrik texnologiyalar; Big Data, böyük verilənlər; bulud texnologiyası

I. GİRİŞ

Böyük verilənlər (Big Data) böyük və mürəkkəb məlumat yığımlarının emalı üçün istifadə olunur. Adi verilənlər bazalarının idarəetmə sistemləri strukturlaşdırılmamış verilənlərin böyük həcmlərini idarə etmək imkanına malik deyil. Böyük məlumatlar, bir qayda olaraq, böyük və mürəkkəb verilənlər yığımlarının məcmusu kimi təyin edilir və ondan məlumat bazalarının idarə edilməsi və ya məlumatların işlənməsi üçün ənənəvi proqramlardan istifadə etdikdə çətinlik yaranan zaman istifadə olunur.

Böyük verilənlər analitikası verilənlər üzrə elmin (Data science) diqqət mərkəzində durur. Bir çox dövlət və özəl təşkilatlar böyük miqdarda müxtəlif verilənləri yığmağa başladılar, belə ki, bunlar problemlər haqqında faydalı məlumatları özündə saxlaya bilir. Məsələn, milli kəşfiyyat, kiber-təhlükəsizlik, marketinq və tibbi informatikada böyük verilənlər texnologiyasından istifadə edilməsi əhəmiyyətli oldu [1].

Yaponiyada Big Data texnologiyalarından istifadə etməklə bəzi mağazalar oğrular və ya "şikayətçilər" kimi müəyyən edilmiş alıcıların qara siyahısında olan şəxslərin tanınmasının proqram təminatını tətbiq etməyə başlamışdır.

Florida ştatının Tampa şəhərində polislər sifətin skan olunmuş təsvirləri əsasında insanın tanınması üçün Big data texnologiyası olaraq Superbowl XXXV böyük proqram təminatından istifadə edirlər və onun vasitəsilə cinayətkarları müəyyənləşdirə bilirlər.

ABŞ-ın bir çox ştatlarındakı universitetlərdə və aeroportlarda biometrik texnologiyalar sahəsində Big Data-dan insanların tanınmasında istifadə olunur.

Lambda Labs proqram təminatını Google Glass üçün sifətin tanınması üçün istifadə edilən Big Data texnologiyalarından hesab etmək olar.

Face Recognition və Big Data Analysis-dən istifadə etməklə hüquq-mühafizə orqanlarında insanın tanınması məsələlərində işin səmərəliliyinin yüksədilməsinə səbəb olur.

İstifadəçinin şəxsiyyətini müəyyən edən müxtəlif sistemlərin təhlükəsizliyinin təmin edilməsində biometrik texnologiyalardan istifadə edilməsi əvəzolunmazdır.

Vahid platformada müxtəlif biometrik əlamətlər üzrə identifikasiya alqoritmlərinin reallaşdırılmasının üstünlüyü ondan ibarətdir ki, bütün əlamətlər üzrə identifikasiyanın nəticələri vahid formada operatora verilir. Nəticələrin verilməsi, onların qiymətləndirməsi və qərarların qəbul edilməsi üçün bəzi yanaşmalar biometrik identifikasiya alqoritmlərinin təşkilində Big Data texnologiyalarından istifadə edilməsi üçün nəzərdə tutulmuşdur.

Məlum olduğu kimi şəkillər, audio, video verilənləri kompüterin yaddaşında böyük yer tutur və onların işlənməsində böyük çətinliklər yaranır. Artıq adi serverlər onların öhdəsindən gələ bilmir. Bu halda Big Data texnologiyalarına böyük ehtiyac yaranır.

Hər bir sahədə çatışmamazlıqlar olduğu kimi Big Data texnologiyalarındada risklərlə və konfidensiallıqla bağlı müəyyən problemlər vardır.

II. BİOMETRİK VƏ BİG DATA TEXNOLOGİYALARI

Daxili işlər orqanlarında olan biometrik məlumatlar, kriminalistika bankları və məlumat bazaları kifayət qədər dağınıq və əsasən lokal səviyyədədir. Regionlararası cinayətkarlığa qarşı mübarizədə bu müəyyən çətinliklərə səbəb olur. Bu da şübhəlilərin üzə çıxardılması təcrübəsində axtarışın imkanlarını məhdudlaşdırılır, yəni şəxsin identifikasiya prosesi həyata keçmədən sona çatır.

Bu problemlərin həllində geniş imkana malik biometrik Big Data texnologiyalarının imkanlarından aktiv şəkildə istifadə oluna bilər [2].

Biometrik texnologiyalar: autentifikasiya (authentication), biokriptoqrafiya (Biocryptography) və Bulud arxitekturasından (Cloud-Based Architecture) istifadə edilməsi təhlükəsizlik məsələləri üçün idealdır. Proqram təminatı kimi it SaaS, Tygart, MXSERVER və s. misal göstərmək olar [3].

MXSERVER güclü server əsasında şəxsin tanınması sistemidir. Böyük həcmli videolar, şəkil kollleksiyaları və s. üçün istifadə olunur.

Tygart texnologiyasına əsaslanan proqram təminatı mətn, video və fotoların uyğun seqmentlərini insanların marağına uygun olaraq qruplaşdıra və silə bilər [3].

Biometrik texnologiyalarda Big Data problemlərinə həsr olunmuş konfrans və seminrlar keçirilmişdir.

2014-cü ilin iyununda Vashingtonda keçirilən “Biometrics Big Data Symposium” böyük maraq doğurmuşdur. Bu simposiumda bometrik texnologiyalar sahəsindəki bəzi problemlər müzakirə olunmuşdur. Onlardan bəziləri aşağıda qeyd olunmuşdur:

- Gələcəkdə dövlətlərin böyük verilənlərin analizi üzrə əsas biometrik proqramlarının və təkmilləşdirilmiş standartlarının olması;
- Biometrik böyük verilənlərin kompakt şəkildə idarə edilməsi və onun öhdəsindən gəlmək üçün üsulların tapılması;
- Genişmiqyaslı analitiklərin "böyük verilənlər"ə göstərəcəkləri köməklik və bu işdə "bulud" texnologiyasının rolu;
- Biometrik böyük verilənlərin analizi və qarşıya qoyulan məsələlərin həllində aktiv mobil mühitdə identifikasiya və autentifikasiya problemlərin həlli üçün irəliləyişin əldə edilməsi;
- Big Data-dan istifadə etməklə tələbə görə identifikasiyanın perspektivlərinin müəyyənləşdirilməsi;
- Biometrik texnologiyalar sahəsində Big Data-dan istifadə edərkən qarşıya çıxan problemlərin və xətaların azaldılmasının həlli yollarının araşdırılması və s. [4].

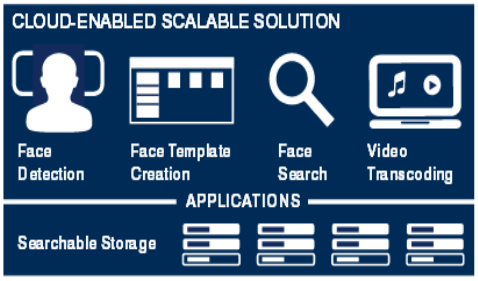
III. BİOMETRİK, BİG DATA VƏ BULUD TEXNOLOGİYALARI

Bulud texnologiyaları geniş sahəni əhatə edən paylanmış resurslara vahid nöqtədən sorğu üzrə giriş imkanı verən yeni informasiya texnologiyalarıdır.

Bu texnologiyaların xüsusiyyətləri sırasına istifadəçilərin bulud servislərlə sərbəst işləmək bacarığının olması, buluda girişin istənilən məkandan, istənilən vaxt, istənilən qurğu (smartfon, noutbuk, planşet, leptop və s.) vasitəsilə mümkünlüyüdür. Bundan əlavə, resurslar külliyatının istifadəçilərə servislər menyusu şəklində təqdim olunması, istifadəçiyə resursların həçmini sərbəst şəkildə artırıb-azaltmaq imkanının verilməsi bu texnologiyanın geniş tətbiqinə yol açan amillərdəndir.

Biometrik texnologiyalarda buluddan istifadə edilməsi təhlükəsizlik və məxfililik problemlərini nəzərə carpacaq dərəcədə azaldır və istənilən xoşagəlməz hadisələrin tez bir zamanda qarşısının alınmasına kömək edir.

MXSERVER ™ server sistemidir, belə ki, kompüterlər, mobil telefonlar, SIM-kartlar və resursların axtarışı üçün nəzərdə tutulmuş videomüşahidə sistemlərindən alınmış fayllardan istifadə etməklə böyük həcmli videoları və foto kolleksiyalarını emal edir (şəkil 1).

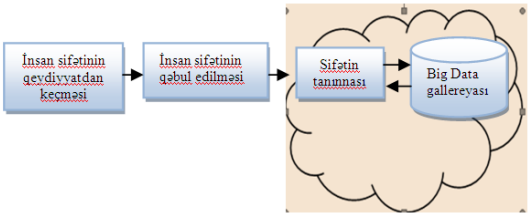


Şəkil 1. MXSERVER ™ video və fotonun analizi sistemi

Tanınmada dəqiqliyin əhəmiyyətli dərəcədə artırılması biometrik texnologiyalarda böyük verilənlərin analizinin əsas üstünlüklərindən biridir.

Biometrik texnologiyaların başqa formalarından fərqli olaraq, şəxsin tanınması proqramı məsafədən idarə oluna bilər. O İnternetdə şəkilləri və ya videonu skan edə bilir ki, buda Facebook və ya Google-da insanların şəkillərinin üzərində nişan qoymaq üçün istifadə olunur.

Big Data texnologiyalarına Apache Hive, Apache Giraph, Horton, Hadoop və s. aiddir. Bu texnologiyalardan istifadə etməklə yüksək keyfiyyətli tanınmanı həyata keçirtmək olar (şəkil 2).



Şəkil 2. Cloud və Big Data texnologiyasından istifadə etməklə insanın tanınması

Şəkil 2-dən göründüyü kimi İnsan sifətinin tanınma prosesi aşağıdakı qaydada baş verir.

Proqram təminatının köməyi ilə sistemə daxil olan yeni insanın sifəti qeydə alınır, kodlaşdırılır və buluda göndərilir, insanın sifəti əvvəlcədən verilmiş sifətin tanınması (face recognition) alqoritmi əsasında işlənilir, bu bir neçə mərhələdən ibarətdir (sifətin müəyyən edilməsi və s.) Sonrakı mərhələdə insan sifəti qalereyada olan təsvirlərlə müqayisə edilir və tanınma baş verir.

MXSERVER məhkəmə ekspertləri, müstəntiqlər və təhlükəsizlik xidməti əməkdaşları üçün faydalıdır (şəkil 3).

Şəkil 3. MXSERVER texnologiyadan istifadə etməklə insanın tanınması

Bulud texnologiyası yüksək miqyaslanmanı asanlaşdırır və sistem geniş istifadəçiyə yüksək səviyyədə xidmət etməyə zəmanət verir.

Bu sahədə etibarlı mexanizmlər vardır, məsələn, Cloud-based FR systemi nəzdində Animetrics, BioID və robust Face texnologiyalarını misal göstərmək olar [5,6].

Data.İT biometrik böyük verilənlərin ənənəvi statistik analiz üsullarından istifadə etməsi problemləri sadə şəkildə həll etmə qabiliyyətinə malikdir [7].

NƏTİCƏ

Bu işdə biometrik texnologiyalar sahəsində Dig Data problemlərindən bəhs edilir. Biometrik texnologiyaların istifadəsində Big Data və bulud texnologiyalarının qarşılıqlı istifadəsinin bəzi aspektləri izah edilir. Nümunə kimi bu texnologiyalardan insan sifətinin tanınmasında istifadə edilməsi barədə məlumat verilir.

Yaxın vaxtlarda Big Data texnologiyasından istifadə etməklə Google Glas üçün proqram təminatı yaradılmışdır. Proqram vasitəsilə istifadəçi küçədə yanından keçən hər hansı bir şəxsin adı və şəklini Facebook vasitəsilə (əgər orda varsa) ekrana çıxara biləcək. Hal-hazırda bu texnologiyalar uzrə işlər aparılır [8]. Gələcəkdə biometrik texnologiyalar sahəsində Big Data-dan geniş şəkildə istifadə edilməsi nəzərdə tutulur.

ƏDƏBİYYAT

[1] M. Najafabadi**,** F.Villanustre**,** T. Khoshgoftaar**,** N. Seliya**,** R. Wald and E. Muharemagic, **“**Deep learning applications and challenges in big data analytics”, Journal of Big Data, USA, 2015, vol 2, №1, pp. 1-21.

[2] Sh Mahmudova, “The Role of Biometric Networks in Recognition of Person // Communications”, USA, 2015, Vol. 3, No. 5, pp. 81-85.

[3] http://findbiometrics.com/face-recognition-and-big-data-analysis-bringing-efficiency-to-law-enforcement/

[4] https://www.asdevents.com/event.asp?id=3225&desc=Biometrics+Big+Data+Symposium

[5] P. Peter, J.Bule, J.Gros and V.Štruc, "Building cloud-based biometric services." Informatica an International Journal of Computing and Informatics, Ljubljanska ,2013, vol 37, No. 1, pp. 115-122.

[6] A. Vinay, S. Vinay, J. Rituparna, T. Aggrawal, K. Balasubramanya and S. Natarajan, “Cloud Based Big Data Analytics Framework for Face Recognition in Social Networks using Machine Learning”, 2nd International Symposium on Big Data and Cloud Computing (ISBCC’15), VIT University, Chennai, India, Mart 12-13, 2015, vol 50, pp. 623-630.

[7] Y. Sun, Y. Chen, X. Wang and X. Tang, "Deep learning face representation by joint identification-verification", In Advances in Neural Information Processing Systems, 2014, pp. 1988-1996.

[8] https://www.linkedin.com/pulse/20140717062548-64875646-why-your-face-matters-more-than-ever

Big Data və Fərdi Məlumatların Təhlükəsizliyi

Yadigar İmamverdiyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
yadigar@lan.ab.az

***Xülasə* — Big data üçün fərdi məlumatların təhlükəsizliyi kritik əhəmiyyətli məsələdir. Bu işdə Big data texnologiyalarının fərdi məlumatlara yaratdığı təhdidlər və bu təhdidləri qarşılamaq üçün fərdi məlumatlar sahəsində qanunvericiliyin təkmilləşdirilməsi çağırışları analiz edilir. Big data texnologiyalarında fərdi məlumatların təhlükəsizliyi üçün mövcud texnoloji yanaşmalar analiz edilir və müvafiq elmi tədqiqat istiqamətləri haqqında məlumat verilir.**

Açar sözlər — Big data; fərdi məlumatlar; fərdi məlumatların təhlükəsizliyi; diferensial gizlilik

I. GİRİŞ

Fərdi məlumatlar – şəxsin kimliyini birbaşa və ya dolayısı ilə müəyyənləşdirməyə imkan verən istənilən məlumatlardır [1]. Fərdi məlumatlar toplandığı andan mühafizə olunmalı və yalnız toplandığı məqsədlər üçün istifadə edilməlidir. Konfidensial fərdi məlumatlar qanunla müəyyən olunmuş hallar istisna olmaqla, üçüncü şəxslərə yalnız subyektin razılığı əsasında verilə bilər.

Big Data texnologiyaları fərdlər barəsində toplanan məlumatların dairəsini və mənbələrini hədsiz dərəcədə genişləndirir. Müxtəlif sosial-iqtisadi aktorlar istifadəçilər haqqında olduqca çox fərdi məlumatlar – istifadəçinin veb-saytlardakı və sosial şəbəkələrdəki davranışı, istifadəçi ilə əlaqəsi olan şəxslərin davranışları və əlaqələri, onların alış-veriş davranışları və s. Toplayırlar. Bundan başqa, sosial şəbəkədə ünsiyyətin emosional çalarlarını da analiz etmək mümkündür. İstifadəçinin sosial mediada nə yazdığı ilə yanaşı, necə yazdığı da analiz edilir. Bir sözlə, potensial istifadəçi barəsində vacib və ya lazımsız görünən hər bir məlumat toplanır və onun “360 dərəcəlik” profili yaradılır [2-4].

Nəticədə Big Data texnologiyaları şəxsi həyatın toxunulmazlığı baxımından ciddi problemlər yaradır. Lakin universal, heç nə ilə məhdudlaşdırılmayan hüquqlar yoxdur. Xüsusi halda, vətəndaşların şəxsi həyatının toxunulmazlığı hüquqları, məsələn, təhlükəsizliyin təmin olunması və ciddi cinayətlərlə mübarizə sahəsində cəmiyyətin qanuni maraqları ilə məhdudlaşdırılır.

Hələ 1967-ci ildə filosof Marşall Maklyuen qeyd edirdi ki, «insanın beşikdən qəbrədək hərtərəfli izlənilməsinə xidmət edən elektrik informasiya qurğuları insanın şəxsi sirlərini qorumaq cəhdləri ilə cəmiyyətin məlumat ehtiyacları arasında olduqca ciddi dilemma yaradır» [5]. ABŞ Milli Təhlükəsizlik Agentliyinin fəaliyyəti haqqında E. Snoudenin açıqlamaları göstərdi ki, Maklyuenin «ciddi dilemması» reallığa çevrilib – müasir izləmə texnologiyaları və Big Data analitikası sayəsində insanların şəxsi həyatının toxunulmazlığı böyük təhlükə altındadır [6].

Big Data texnologiyaları strateji, taktiki və operativ səviyyələrdə yaxşı əsaslandırılmış mürəkkəb qərarların qəbul edilməsinə və müxtəlif funksional sahələrdə yeni həllərin və məhsulların yaradılmasına şərait yaradır. Buna görə Big Data texnologiyaları, onların biznes perspektivləri, tətbiqləri və böyük miqyaslı yüksək məhsuldarlıqlı intellektual analiz sahəsində bir çox tədqiqat işləri aparılmışdır. Lakin Big Data müstəsna faydalı məlumatlar versə də, verilənlərin toplanması, saxlanması, saxlanma müddəti, təhlükəsizliyi və gizliliyi ilə bağlı yeni problemlər də yaradır və bu problemlər yalnız son illər tədqiqatçıların diqqətini cəlb etməkdədir [7,8].

Bu işin məqsədi Big Data baxımından fərdi məlumatların təhlükəsizliyi üzrə elmi-praktiki tədqiqatların müasir vəziyyətini analiz etmək və aktual tədqiqat istiqamətlərini müəyyən etməkdir.

II. BİG DATA: FƏRDİ MƏLUMATLARIN MƏNBƏLƏRİ

Ənənəvi informasiya texnologiyaları strukturlaşdırılmış verilənlərdən istifadə edir və bunun nəticəsində yaxın vaxtlara kimi IT-bazarın liderləri verilənlər bazalarını idarəetmə sistemlərinin istehsalçıları idi (Oracle, SAP və s.). Lakin hazırda mühüm yerdəyişmə baş verməkdədir – verilənlərin çox sayda yeni mənbələri meydana çıxır, təkcə sosial şəbəkə servislərini və mobil texnologiyaları yada salmaq kifayətdir. Bu verilənlər çox zaman strukturlaşdırılmayıb, olduqca müxtəlif formatlara malikdirlər, böyük sürətlə yaranırlar və həcmləri kəskin sıçrayışla artır. Belə informasiya axınının mövcud texnologiyalarla emalı mümkün deyil və yeni texnologiyalar tələb edilir. Big Data texnologiyaları müxtəlif mənbələrdən alınmış strukturlaşdırılmamış verilənlərin real zamanda emalını mümkün edir.

Biliyin, informasiyanın əsas qeyri-maddi aktiv olduğu sahələr – maliyyə, satış, marketinq, idarəetmə sahələri böyük həcmdə toplanan fərdi məlumatların əsas istehlakçılarıdır. Son dövrlər hüquq-mühafizə orqanları və milli təhlükəsizliyi təmin edən strukturlar cinayətkar və terrorçu elementlərin daha dəqiq aşkarlanması üçün Big Data ideyalarından istifadə edilməsində maraqlıdırlar. Müşahidə kameralarının məlumatları, sosial şəbəkələrdəki yazışmalar və paylaşılan multimedia məlumatları, veb-saytların loq-faylları, mobil və ənənəvi rabitə ilə danışıqlar, elektron xidmətlərin tranzaksiyaları yaxşı məlum olan Big Data mənbələridir.

Big Data texnologiyalarının marketinq məqsədləri üçün istifadəsinə yol açanlar onlayn-topdansatış mağazalarıdır.

Oflayn mağazalardan fərqli olaraq, onların İnternet sayəsində öz alıcılarının davranışı ilə bağlı bütün verilənləri analiz etmək üçün unikal imkanları var. Onlar olduqca müxtəlif verilənlər – müştərinin veb-saytlardakı və sosial şəbəkələrdəki davranışı, müştərinin əlaqəli olduğu şəxslərin davranışları, alış-veriş davranışları, kredit tarixçələri və s. haqqında müxtəlif məlumatlar toplayırlar.

Big data verilənlərinin 15-20%-i "Əşyaların İnterneti", o cümlədən, çoxsaylı telefonlar, planşetlər və digər qurğular tərəfindən generasiya edilir. "Əşyaların İnterneti" tərəfindən generasiya edilən verilənlərin payı 2020-ci ildə 40%-ə çatacaq.

Müasir tibb texnologiyaları tibbi yardımın göstərilməsi ilə bağlı böyük həcmdə verilənlər generasiya edir (şəkillər, video, real vaxt rejimində monitorinq). Artıq bu gün daşınan elektronika və sensor texnologiyaları insanın sağlamlıq vəziyyəti haqqında verilənləri real zamanda təqdim edə bilirlər. Sensor texnologiyaları mütəxəssislərə xroniki xəstəliyi olan pasiyentlərin vəziyyətini uzaqdan monitorinq etməyə və onlara vaxtında kömək göstərməyə imkan verir [9].

Tibbi sənədlərin ənənəvi formalardan elektron formalara keçirilməsi nəticəsində də böyük həcmdə fərdi məlumatlar informasiya sistemlərinə daxil olur.

Dövlət sektoru da çox böyük həcmdə verilənlər toplayır və bu verilənlərin 70%-i strukturlaşdırılmayıb, müxtəlif sənədlər – planlar, hesabatlar, ərizələr, şikayətlər, təkliflər… şəklində saxlanır. Bu verilənlərin həcmi daim artır və dünya üzrə illik artım sürəti 30%-dir. Vətəndaşlar daim bu və ya digər şəkildə dövlət orqanları ilə ünsiyyətdə olurlar, elektron dövlətin formalaşdırılması sayəsində bu ünsiyyətin intensivliyi və nəticədə e-xidmət sistemlərində toplanan fərdi məlumatların həcmi kəskin şəkildə artır.

III. BİG DATA VƏ FƏRDİ MƏLUMATLARA TƏHLÜKƏLƏR

Big Data texnologiyalarından istifadə etmək imkanı olan müxtəlif subyektlər fərdi məlumatları toplayırlar. Bu zaman bir sıra suallar meydana çıxır. Məsələn, şirkətlərin qanuni yolla hansı verilənləri toplamaq hüququ var?

Fərdi verilənlər çox zaman fərdlərin razılığı olmadan toplanır və çox nadir hallarda fərdlərin bu verilənlərin aqreqasiyasına və başqa məqsədlər üçün istifadə edilməsinə razılığı alınır. Bu fərdi məlumatların sui-istifadə ssenariləri olduqca rəngarəngdir.

İnsanların öz fərdi məlumatlarına nəzarət etmək və idarə etmək imkanları yoxdur. Fərdi məlumatların həcmi artdıqca istifadəçilərin hansı məlumatları, nə zaman, kiminlə paylaşdıqlarına nəzarət etmələrinə imkan verən mexanizmlər əlyetər olmalıdır.

Daha bir təhlükə ondadır ki, Big Data texnologiyaları sayəsində şirkətlər haqqımızda həssas və şəxsi faktları bizdən almadan, müxtəlif mənbələrdən topladıqları məlumatları birləşdirməklə birbaşa nəticə çıxararaq əldə edə bilərlər.

Big Data texnologiyaları məhsuldarlığı yüksəltmək və yeni biznes-prosesləri təkmilləşdirmək məqsədilə əməkdaşlar haqqında müfəssəl məlumatlar toplamaq üçün istifadə edilir. Lakin əməkdaşların daim izlənməsi təşkilatlarda gərgin atmosfer yarada bilər.

Bundan başqa, böyük həcmdə multimedia verilənlərinin yayılması genişləndikcə açıq və konfidensial verilənlər arasında sərhədlər yox olur. Video-verilənləri sosial şəbəkələrə yükləməyə imkan verən onlayn proqramlar insanı mahiyyətcə videokameraya çevirir. Lakin video-müşahidə kameralarından fərqli olaraq smartfonlar, məsələn, təsadüfən kadra düşən şəxslər üçün konfidensiallığın qorunmasını təmin etmirlər. Məsələn, Bostonda partlayışların təhqiqatı zamanı bir neçə nəfər terror aktının törədildiyi yerdə çəkilmiş fotoşəkillərin sosial şəbəkə saytlarında yerləşdirilməsindən sonra şübhəlilər sırasına düşmüşdü.

De-identifikasiya yanaşması təklikdə fərdi məlumatların təhlükəsizliyi üçün yetərli həll deyil. Əlavə mənbələrdən verilənlər cəlb etməklə məlumatların re-identifikasiya edilməsi imkanları da kifayət qədər realdır [10].

Bəzi fərdi məlumatlar şəxs barəsində qərar qəbul edilməsi, onun diskriminasiyası üçün istifadə edilə bilər.

Big Data fərq qoymadan bütün verilənləri toplamağa imkan verir. Verilənlər yeni kəşflər ediləcəyini, yeni məhsullar və servislər yaradılacağını vəd edirlər. Lakin bu verilənlərin gigiyenası üzrə “birinci qanuna” ziddir: verilənləri, xüsusilə də fərdi məlumatları ehtiyac olmadan toplamayın və saxlamayın. Bu qanunu Big Data əsrində ləğv etmək lazımdır, yoxsa hər yerdə hər veriləni toplamaq lazımdır; yaxud verilənlərin toplanmasını məhdudlaşdırmaq Big Data potensialını məhdudlaşdırırmı?

Big Data və fərdi məlumatların təhlükəsizliyi problemləri üzrə Gizliliyin Gələcəyi Forumunun və Stenford Internet və Cəmiyyət Mərkəzinin birgə təşkil etdiyi “Big Data and Privacy: Making Ends Meet” seminarının əsərlər toplusu yaxşı mənbədir [11]. Bu toplu aşağıdakı əsas suallardan bəzilərinə ünvanlanıb: Big Data yeni çağırışlar gətirirmi, yoxsa sadəcə verilənlərin tənzimlənməsi müzakirələrini yenidən gündəmə gətirir? Big data vətəndaş azadlıqlarına təhdid törədən prinsipcə yeni imkanlar yaradırmı? De-identifikasiya və digər texnoloji və ya inzibati tədbirlər gizlilik risklərini yetərincə azaldırmı? Verilənlərin minimallaşdırılması kimi gizlilik konsepsiyaları Big Data sahəsində hansı rolu oynamalıdır? Digər sahələrdə alınmış hansı dərslər tətbiq edilə bilər?

Massaçusets Texnologiya İnstitutunda “Big Data və fərdi məlumatların təhlükəsizliyi problemləri üzrə işçi qrup” (MIT Big Data Privacy Working Group) 2013-cü ildən başlayaraq bir neçə seminar keçirmişdir. Həmin işçi qrup fərdi məlumatların təhlükəsizliyində Big data texnologiyalarının təbiətindən irəli gələn aşağıdakı xüsusiyyətləri diqqətə çatdırır [12]:

Miqyas: Verilənlərin həcminin nəhəng olması gizlilik siyasətlərinin yaradılmasına, idarə edilməsinə və tətbiqinə ciddi problemlər yaradır.

Müxtəliflik: Big Data toplanmasına, idarə edilməsinə və istifadəsinə müxtəlif iştirakçılar cəlb edilir və onların məqsəd və hədəflərindəki fərqlər ziddiyyətlərə gətirib çıxarır.

İnteqrasiya: Verilənlərin idarə edilməsi texnologiyalarının inkişafı ilə (məsələn, bulud servisləri) müxtəlif verilənlər

topluları arasında inteqrasiya artır və məntiqi çıxarışlar əsasında fərdlər və onların davranışları haqqında yeni informasiya almaq imkanı genişlənir.

Digər iştirakçılara təsiri: Məlumatların bir çox hissəsi təkcə hədəf subyektə deyil, həm də digər, çox zaman istəməyərəkdən iştirak edən şəxslərə aid ola bildiyindən, məntiqi çıxarış nəticəsində alınmış yeni informasiya da gizliliyin subyekti olmayan digər şəxslərə aid ola bilər.

Yeni informasiya üçün yeni siyasətin zəruriliyi: Birləşdirilən verilənlər yeni məntiqi nəticələr verdikdə yeni informasiya və ya anlayışlar ortaya çıxır. Hər bir verilənlər toplusu mövcud gizlilik siyasətinə və həyata keçirilmə mexanizmlərinə malik ola bilər, lakin yeni məlumatlar üçün rekvizitlərin, uyğun gizlilik siyasətlərinin və müvafiq reallaşdırma mexanizmlərinin avtomatik yaradılmasına ehtiyac vardır.

Təşkilat daxilində fərdi məlumatların təhlükəsizliyi ilə bağlı aşağıdakı tövsiyələri vermək olar [12]:

- fərdi məlumatların təhlükəsizliyi risklərini qiymətləndirmək üçün təşkilat daxilində yeni rollar yaradılmalıdır;
- Big Data layihələrinin başlanğıc mərhələsində fərdi məlumatların təhlükəsizliyi riskləri qiymətləndirilməlidir;
- ilkin mərhələdən başlayaraq istifadəçilər üçün şəffaflıq təmin edilməlidir – verilənlərin toplanmasının məqsədlərini və gələcək istifadə potensialını aydın göstərmək zəruridir;
- verilənlərin saxlanması mühiti müntəzəm qiymətləndirilməli və istifadə modellərinin öyrənilməsi təmin edilməlidir;
- təhlükəsizlik mexanizmləri verilənlərə yaxın yerləşdirilməli və fərdi məlumatların bütün növləri üçün şifrələmədən istifadə edilməlidir;
- təşkilatlar mövcud və potensial müştərilərinin etimadını qorumaq üçün fərdi məlumatların qanun və etik normalar çərçivəsində istifadə edilməsini və istənilən növ konfidensial fərdi məlumatlara müraciətlərin qeydiyyata alınmasını və monitorinq edilməsini təmin etməlidirlər.

IV. BİG DATA VƏ FƏRDİ MƏLUMATLAR ÜZRƏ QANUNVERİCİLİYİN TƏKMİLLƏŞDİRİLMƏSİ

Big Data texnologiyalarının inkişafı fərdi məlumatların təhlükəsizliyinin təmin edilməsinin qanunvericilik bazasına yenidən baxılmasını zəruri edir [13]. Bu sahədə Avropa təşəbbüslərinə nəzər salmaq vacibdir.

Avropa Komissiyası 2015-ci ilin sonuna qədər fərdi məlumarların qorunması sahəsində yeni Avropa qanununun (General Data Protection Regulation, GDPR) qəbulu üçün lazım olan bütün razılaşmalara nail olmağı planlaşdırır [14]. Qüvvədə olan "EU Data Protection Directive 95/46/EC" təlimatından fərqli olaraq, yeni qanun Avropa Birliyinə üzv-ölkələrdə səsvermə keçirilmədən "birbaşa" qüvvəyə minəcək və icrası məcburi olacaq.

Fərdi məlumatların qorunması sahəsində dəyişikliklərə ehtiyac həm də Avropa iqtisadiyyatının rəqabət qabiliyyətinin yüksəldilməsi ilə əlaqəlidir. Bəzi qiymətləndirilmələrə görə, Avropa vətəndaşlarının fərdi məlumatlarının illik dəyəri 2020-ci ildə təxminən 1 trilyon avroya çatacaq. Ona görə fərdi məlumatların emalı sahəsində yüksək Avropa standartlarının bərqərar edilməsi biznesin inkişafına xidmət etməyə yönəlib. Məqsəd Big Data və bulud servislərinin Avropada inkişafını dəstəkləməkdir. Hansı ölkədən gəlmələrinə baxmayaraq, Avropada fəaliyyət göstərən bütün şirkətlərə fərdi məlumatların emalı üzrə eyni qaydalar tətbiq ediləcək. Eyni zamanda, kiçik şirkətlərin fərdi məlumatlar bazarına girişi də asanlaşdırılır. Qanunun fərdi məlumatların emalı sahəsində pərakəndəliyi aradan qaldıracağı və inzibati xərcləri ildə təxminən 2,3 milyard avro azaldacağı da gözlənilir [14].

Avropa Birliyinə üzv-ölkələr üçün yeni, vahid qanunvericilik fərdi məlumatların qorunması üzrə tələblərin icrası ilə bağlı xərcləri, xüsusilə kiçik və orta müəssisələr üçün əhəmiyyətli dərəcədə azaltmağa imkan verəcək. Beləliklə, vahid qanunvericilik təşkilatların hakimiyyətlə qarşılıqlı əlaqəsini elə təşkil etməyə imkan verəcək ki, fərdi məlumatların qorunması üzrə qeydiyyat ölkəsindəki təşkilat vahid əlaqə nöqtəsinə çevriləcək. Müvafiq olaraq, yeni qanunvericilik Avropa Birliyinə üzv-ölkələrin güc strukturları arasında fərdi məlumatların mübadiləsini də sadələşdirəcək.

Vətəndaşlara öz fərdi məlumatlarına baxmaq və düzəliş etmək hüququ, verilənlərin emalına etiraz etmək hüququ, verilənlərin sızması baş verdikdə məlumatlandırılmaq hüququ, unudulmuş olmaq hüququ (yəni, fərdi məlumatlarının operator tərəfindən dərhal, izafi rəsmiyyət olmadan məhv edilməsini tələb etmək hüququ) verilir. Bu hüquqların başqa şəxslərin və təşkilatların hüquq və azadlıqlarına, bütövlükdə cəmiyyətin maraqlarına zidd olmaması gözləniləcək. Vətəndaşlar üçün əsas faydalar vətəndaşların fərdi məlumat operatorlarına etibarının artmasından yaranacaq, bu onlara daha geniş onlayn-xidmətlərdən istifadə etməyə imkan verəcək (və eyni zamanda bu iqtisadiyyatın "elektron" sektorunun artmasına imkan yaradacaq).

Autentifikasiyanın gücləndirilməsi yeni qanunun tələblərinin qanunauyğun nəticəsidir. Fərdi məlumatların itkisi haqqında məcburi məlumat verilməsi və fərdi məlumatların itkisinə görə yüksək cərimələr qaçılmaz olaraq "fərdi məlumatların emalı zamanı daha enerjili müdafiə tədbirlərinin" görülməsinə gətirib çıxaracaq [15].

GDPR layihəsinin mətnində göstərilir ki, yeni qanunu pozmuş təşkilatlar ümumi gəlirin 2 %-inə kimi cərimələnə bilərlər. Əgər şirkət məlumat itkisinin qarşısını almaq üçün bütün mümkün tədbirləri gördüyünü sübut edə bilsə, onda daha yumşaq cəza tətbiq ediləcək.

Ehtimal olunur ki, fərdi məlumatların qorunması sahəsində Avropa Birliyinin yeni islahatları sanksiyalar rejiminin xeyli sərtləşməsinə gətirib çıxaracaq. Bununla əlaqədar olaraq, biznes dairələri Avropa qanunvericilərindən və nəzarət orqanlarından yeni qanunun daha aydın şərhini, xüsusi halda təhlükəsizlik tələblərinə uyğunluq üçün çoxpilləli

autentifikasiya və ya digər müdafiə tədbirləri daxil olmaqla, hansı tədbirlərin qəbul edilməsinin zəruri olduğunu aydınlaşdırmağı tələb edəcəklər.

Son illər ölkəmizdə fərdi məlumatlar sahəsində bir sıra normativ hüquqi aktlar qəbul edilmişdir, o cümlədən “Fərdi məlumatların avtomatlaşdırılmış qaydada işlənməsi ilə əlaqədar şəxslərin qorunması haqqında” 1981-ci il yanvarın 28-də Strasburq şəhərində imzalanmış Konvensiya müvafiq bəyanatlarla təsdiq edilmiş (2009-cu il) və “Fərdi məlumatlar haqqında” Azərbaycan Respublikası Qanunu qəbul edilmişdir (2010-cu il). Bununla yanaşı, Big Data reallıqlarını nəzərə almaqla fərdi məlumatlar üzrə milli qanunvericiliyin təkmilləşdirilməsinə zərurət yaranır.

V. FƏRDİ MƏLUMATLARIN TƏHLÜKƏSİZLİYİ ÜZRƏ TEXNOLOJİ YANAŞMALAR

Analiz göstərir ki, aşağıdakı istiqamətlərdə praktiki həllərin işlənilməsi zəruridir [16]: 1) Big Data verilənlərinin təhlükəsiz və gizliliyin qorunması ilə toplanması və emalı; 2) Big Data analizinin təhlükəsiz mühitdə və gizliliyin qorunduğu şəkildə həyata keçirilməsi; və 3) Big Data sistemləri üçün verilənlərin saxlanılması siyasətinin təhlükəsiz (və gizlilik rejimində) tətbiqi. Əgər bu həllər lazımi səviyyədə olmasa, insanların öz fərdi məlumatlarını Big Data sistemlərinə təqdim etməsi istəyi azalır.

“Fərdi məlumatların təhlükəsizliyinə cavabdeh təşkilatlar adətən de-identifikasya üsullarından, o cümlədən, anonimlik (*anonymization*), təxəllüs (*pseudonymization*), şifrləmə (*encryption*), açar-kodlaşdırma (*key-coding*) və s. istifadə edirlər. Anonimlik ad, ünvan və sosial təhlükəsizlik nömrələrini silməklə gizliliyi təmin edirsə, təxəllüs bu informasiyanı ləqəb və süni identifikasiya ilə əvəz edir. Açarla kodlaşdırma fərdi məlumatları kodlaşdırır və onların dekodlaşdırılması üçün açar yaradır” [17,18].

Tibbi məlumatların de-identifikasiyası üçün ABŞ federal qanunu HIPAA (Health Insurance Portability and Accountability Act – tibbi sığortanın portativliyi və hesabatlılığı) iki metod irəli sürür [9]:

- ***Təhlükəsiz liman:*** fərdin və ya qohumlarının, ailə üzvlərinin və ya işə götürənlərin 18 spesifik identifikatorunun silinməsi tələb edilir: adlar, ünvanlar, tarixlər, telefon nömrələri, faks nömrələri, tibbi sığorta nömrələri, elektron poçt ünvanları, sosial sığorta nömrələri, tibbi sənədləşmə nömrələri, hesab nömrələri, lisenziya/sertifikat nömrələri, maşın identifikatorları və seriya nömrələri, qurğu identifikatorları və seriya nömrələri, URL (universal resource locator), IP-ünvanlar, biometrik identifikatorlar, üzün tam fotoşəkilləri və identifikasiyanın başqa unikal nömrəsi, xarakteristikası və ya kodu.
- ***Statistik metod:*** müvafiq təlim görmüş şəxs identifikatorların kifayət qədər silindiyini yoxlayır.

Texnoloji yanaşmalardan biri də bu sahədə yeni kriptoqrafik alqoritmlərin işlənməsidir. Klassik şifrləmə üsulları artıq yetərli deyil və şifrlənmiş verilənlər üzərində hesablamalar aparmağa imkan verən üsulların işlənməsi tələb edilir:

Funksional şifrləmə – şifrlənmiş məlumat əsasında müəyyən f funksiyasının qiyməti hesablanır: m açıq məlumatının şifrlənmiş mətninə əsasən $f(m)$ hesablanır, bu zaman m haqqında (o cümlədən, f-in hesablanması zamanı aralıq nəticələr haqqında) hər hansı əlavə məlumat əldə etmək mümkün olmamalıdır. İlk funksional şifrləmə sxemi [19] yalnız funksiyaların bəzi siniflərində işləyirdi (monoton funksiyalar və skalyar hasil), hazırda istənilən funksiyaya tətbiq edilə bilən funksional şifrləmə sxemləri təklif edilib [20].

Diferensial gizlilik metodu fərdi məlumatlara müraciətləri alqoritm və interfeys (etibarlı kurator kimi çıxış edir) vasitəsilə idarə edir [21]. Tədqiqatçı verilənləri analiz etmək üçün kuratora sorğu göndərir və kurator da təsadüfi küy əlavə etməklə verilənlərin konfidensiallığını qorumağa və eyni zamanda sorğulara düzgün cavab verməyə çalışır.

Diferensial gizliliyin tərifində iki D və D' verilənlər çoxluğuna baxılır. D-də verilmiş fərdin məlumatları var və bu məlumatlar D'-də silinib (və ya hər hansı əlaqəsi olmayan məlumatlarla əvəzlənib). Tələb olunur ki, prosesin sonunda D və D'-ni bir-birindən fərqləndirmək mümkün olmasın. Fərqləndirməmə ε parametrindən istifadə etməklə ölçülür. ε parametri D və D'-dən alınmış nəticələrin paylanmasının nə dərəcədə yaxın olmasını ölçür. ε kiçik olduqda gizliliyin təhlükəsizliyi səviyyəsi yüksək olur.

Verilənlərin müxtəlif növləri üzərində analitik məsələlər üçün diferensial gizlilik alqoritmlərinin işlənməsi sahəsində bir çox tədqiqatlar mövcuddur [21,22].

Homomorf şifrləmə sistemi şifrlənmiş məlumatlar üzərində məlumatları deşifrləmədən riyazi əməliyyatlar (məsələn, toplama, çıxma, birləşmə, kəsişmə) aparmağa imkan verir. Bu şifrləmə sistemi fərdi məlumatların qorunması, elektron səsvermə, maliyyə məlumatlarının emalı, tövsiyə sistemlərində tətbiq edilə bilər. Ən uğurlu hesab edilən homomorf şifrləmə sistemi 2009-cu ildə IBM şirkətinin kriptoqrafları tərəfindən təklif edilib və IBM tərəfindən patentləşdirilib [23].

CryptoDB – şifrlənmiş verilənlər bazasında SQL sorğularının yerinə yetirilməsi üçün verilənlər bazasını idarəetmə sistemidir (VBİS). CryptoDB ənənəvi VBİS-lərdə (MySQL, Postgres) və tətbiqi proqramlarda dəyişiklik edilməsini tələb etmir. Eyni zamanda, CryptoDB-nin gətirdiyi əlavə yüklənmə də çox deyil: MySQL ilə müqayisədə etalon testlərə görə (TPC-C performance) məhsuldarlıq itkisi 26%-dir [24].

NƏTİCƏ

Big Data texnologiyaları fərdi məlumatların təhlükəsizliyi baxımından ciddi problemlər yaradır. Bu texnologiyalar sayəsində fərdi məlumatlardan sui-istifadə edilməsinin mənfi nəticələri təkcə ayrı-ayrı insanların şəxsi həyatlarının gizliliyinin pozulmasında deyil, ictimai həyatda, iqtisadiyyatda və siyasətdə də özünü göstərə bilər. Bu problemlərin aradan qaldırılması üçün Big Data texnologiyalarının insan həyatında oynamağa başladığı rolu yenidən qiymətləndirmək tələb edilir.

Eyni zamanda, fərdi məlumatlar sahəsində milli qanunvericiliyin təkmilləşdirilməsi və fərdi məlumatların konfidensiallığını dəstəkləyən texnoloji alətlərin yaradılması olduqca zəruridir. Big Data sahəsində fərdi məlumatların təhlükəsizliyi üzrə bir sıra çətin texnoloji problemlərin həlli: səmərəli şifrləmə və deşifrləmə alqoritmlərinin, şifrlənmiş informasiyada axtarış metodlarının, atribut əsasında şifrləmə alqoritmlərinin işlənməsi, Big Data əlyetənliyinə, etibarlığına və tamlığına hücumların aşkarlanması metodlarının işlənməsi aktual tədqiqat problemləridir.

ƏDƏBİYYAT

[1] Fərdi məlumatlar haqqında Azərbaycan Respublikasının Qanunu. 11 may 2010-cu il.

[2] T. M. Payton, T. Claypoole. Privacy in the age of Big Data: Recognizing threats, defending your rights, and protecting your family. Rowman& Littlefield Publishers. 2014.

[3] C.L. Borgman. Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press. 2015.

[4] M.Smith, C.Szongott, B.Henne, G. von Voigt, “Big data privacy issues in public social media,” 6th IEEE International Conference on Digital Ecosystems Technologies (DEST), pp.1-6, 18-20 June 2012.

[5] M. McLuhan, Q. Fiore. The Medium is the Massage: An Inventory of Effects. Random House, 1967.

[6] L. Li, M. F. Goodchild and S. Barbara, “ Is Privacy Still an Issue in the Era of Big Data — Location disclosure in spatial footprints”, Proceedings of 21st International conference on Geoinformatics, IEEE, pp.1-4, 2013.

[7] CSA: Top Ten Security and Privacy Issues. November 2012.

[8] R.Alguliyev, Y.Imamverdiyev, “Big Data: Big Promises for Information Security,” IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), 2014.

[9] M. Мамедова, “Проблемы информационной безопасности персональных данных в условиях электронной медицины,” Beynəlxalq Telekommunikasiya İttifaqının 150 illiyinə həsr olunmuş “İnformasiya təhlükəsizliyinin multidissiplinar problemləri üzrə II respublika elmi-praktiki konfransı” nın əsərləri, 14 may 2015, s.52-55.

[10] J. Sedayao, R. Bhardwaj and N. Gorade, “Making Big Data, Privacy, and anonymization work together in the enterprise: Experiences and issues”, IEEE International Congress on Big Data, pp.1-7, 2014.

[11] O. Tene, J. Polonetsky, “Big Data for All: Privacy and User Control in the Age of Analytics,” Northwestern Journal of Technology and Intellectual Property, Vol. 11, Issue 5, Article 1, 2013.

[12] “Big Data and Privacy: Making Ends Meet” workshop http://www.futureofprivacy.org/wp-content/uploads/Big-Data-and-Privacy-Paper-Collection.pdf

[13] Big Data Privacy Working Group. Big Data Privacy Scenarios. September 2015.

[14] Reform of the data protection legal framework in the EU. http://ec.europa.eu/justice/data-protection/reform/index_en.htm

[15] E. J. Kennedy, C. Millard, “Data security and multi-factor authentication: Analysis of requirements under EU law and in selected EU member states,” Queen Mary School of Law Legal Studies Research Paper No. 194/2015.

[16] President’s Council of Advisors on Science and Technology (PCAST): Big Data and Privacy: A Technological Perspective. 2014.

[17] M. Hacırəhimova, “Big Data texnologiyalarının təhlükələri,” Beynəlxalq Telekommunikasiya İttifaqının 150 illiyinə həsr olunmuş “İnformasiya təhlükəsizliyinin multidissiplinar problemləri üzrə II respublika elmi-praktiki konfransı” nın əsərləri, səh. 248-251, 2015.

[18] F. F. Zhao, L. F. Dong, K. Wang, Y. Li, “Study on Privacy Protection Algorithm Based on K-Anonymity,” Physics Procedia, Vol.33, pp. 483-490, 2012.

[19] A. Sahai, and B. Waters, “Fuzzy identity-based encryption,” Eurocrypt, pp. 457–473, 2005.

[20] S. Goldwasser, Y. Kalai; R. A. Popa; V. Vaikuntanathan, N. Zeldovich, “Reusable Garbled Circuits and Succinct Functional Encryption,” Proc. of the 35th annual ACM Symposium on Theory of Computing (STOC’13), pp. 555-564. 2013.

[21] C. Gentry, “Fully homomorphic encryption using ideal lattices,” Proceedings of the 41st annual ACM Symposium on Theory of Computing (STOC’09), pp. 169-178, 2009.

[22] C.Dwork, “Differential Privacy: A Survey of Results,” TAMC 2008, LNCS 4978, pp. 1–19, 2008.

[23] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, “Differentially private data release for data mining,” Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 493-501, 2011.

[24] R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan, “CryptDB: Protecting Confidentiality with Encrypted Query Processing, Proc. of the 23rd ACM Symposium on Operating Systems Principles (SOSP’11), pp. 85-100, October 2011.

Vacib İnfrastukturların İdarəetmə Sistemlərinin Bulud Texnologiyaları Əsasında Təşkili Problemləri

Zəfər Cəfərov[1], Rauf Həsənov[2], Rüfət Talıbov[3]
[1,2,3]Azərbaycan Texniki Universiteti, Bakı, Azərbaycan
[1]*c.zafar@mail.ru* ,[3]*r.talibov@delta-telecom.net*

***Xülasə* — Təqdim olunan işdə vacib ifrastrukturların idarəetmə sistemlərininin bulud texnologiyaları əsasında təşkil edilməsi problemləri təhlil olunur. Göstərilir ki, idarəetmə proseslərinin dayanıqlı və təhlükəsiz yerinə yetirilməsi üçün bu cür informasiya sistemləri daha səmərəlidir.**

Açar sözlər — vacib ifrastruktur, idarəetmə sistemləri, bulud texnologiyası, verilənlərin analizi

I. GİRİŞ

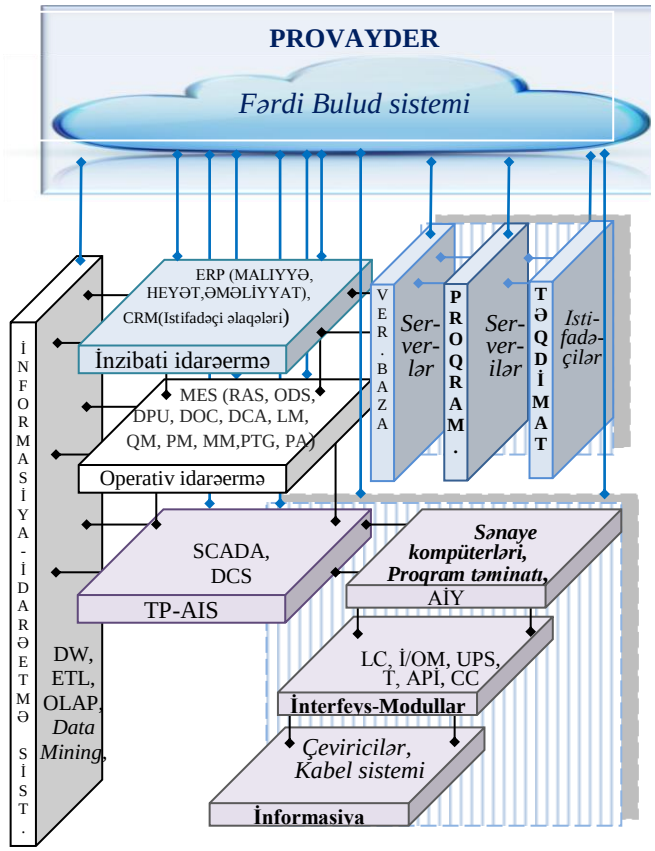
Əksəriyyət vacib ifrastrukturların informasiya sistemləri yüz minlərlə istifadəçilərə xidmət təqdim edən qərarlar dövriyyəsini təsvir edir. Bu cür informasiya sistemləri xüsusilə idarəetmə proseslərini şəffaf, vaxtlı və təhlükəsiz həyata keçirilməli olan dövlət və ya kommersiya müəssisələrinə aiddir. Texniki qəza və ya digər səbəblər üzündən informasiyanın sızması, itkisi və ya korlanmasına görə müraciətlərin gecikməsi və ya imtina olunması ciddi zərərli nəticələrə gətirib çıxara bilər. Vacib ifrastrukturların idarəetmə sistemlərində emal olunan informasiyanın həcmi və yükü dinamik olaraq artır. Əksəriyyətini struktursuz informasiya təşkil edən bu verilənlərin *Data Base* formatında saxlanması zərurəti yaranır. İnformasiyanın müxtəlif mənbələrdə yerləşərək mütəmadi yeniləşməsi, fərqli tərkibə, formata və iri həcmə malik olması bu problemi kəskinləşdirir. Bu cür informasiya «Böyük verilənlər» (big data) adlandırılmışdır.

Vacib ifrastrukturların idarəetmə sistemi, avadanlıqların dayanıqlı və təhlükəsiz istismarı, verilənlərin tamlığı üzrə cavabdehlik daşıyan provayderlər üçün minimal cavab reaksiyalı xidmət zəmanətinə malik, müasir data texnologiyaları əsasınnda bulud konsepsiyasının istifadə edilməsi ən səmərəli qərar hesab olunur.

II. VACİB İNFRASTRUKTURLARIN İDARƏETMƏ SİSTEMİNİN FUNKSİONAL MODELİ

Vacib ifrastrukturları idarəetmə sisteminin funksional modelini (şək. 1) müxtəlif idarəetmə səviyyəli müstəvilərin konstruksiyası kimi təsvir etmək olar [1].

Ən aşağı səviyyədə *texnoloji prosesləri idarəetmə sistemi* (TP AİS), aralıqda *operativ idarəetmə* sistemi və yuxarı hissədə *inzibati idarəetmə sistemi* yerləşir. Bu model avadanlıqlardan inzibati idarəetməyə qədər dövr olunan informasiya mübadiləsini təsvir edir.



Şəkil 1. Vacib ifrastrukturları idarəetmə sisteminin fuksional modeli

Texnoloyi proseslər səviyyəsindəki qurğulardan siqnallar operativ idarəetmə səviyyəsində yerləşən MES (*Manufacturing Execution System*) sisteminə daxil olur. Burada emal olunmuş informasiya vacib infrastrukrun inzibati idarə olunması üzrə ERP (*Enterprise Resource Planning*) sisteminə təqdim edilir. Bu sistemdən təlimat məlumatlarını qəbul edən orta səviyyə tələbatlara uyğun olaraq cari prosesləri yerinə yetirir. Aralıq səviyyəsi (MES-sistemi) tətbiqi avadanlıqların boş dayanma vaxtlarını və texnoloyi əməliyyatlar üçün sərf olunan müddəti əhəmiyyətli dərəcədə azaldır.

Inzibati idarəetmə səviyyəsinin əsasını ERP və CRM (*Customer Relationship Management*) sistemləri təşkil edir [2]. ERP sistemi əsasən maliyyə, kadr və əməliyyat modullarından təşkil olunur. Bu sistemdə həm də mərkəzi informasiya sistemi kimi vacib infrastrukturun fəaliyyəti üzrə zəruri olan informasiya saxlanılır. Bu məlumatların bəziləri sirr daşıyır və onların açıqlanması vacib infrastruktur sisteminə əhəmiyyətli dərəcədə zərər gətirə bilər. Buna görə də, informasiya təhlükəsizliyi problemləri ERP-sistemləri üçün xüsusilə aktualdır.

ERP sistemi operatorun inzibati fəaliyyətininin əsas aspektlərini dəstəkləyən inteqrallaşdırılmış proqramlar dəstidir: xidmətlərin planlaşdırılması, xidmətin təqdim olunması üçün zəruri resursların plamnlaşdırılması, planların icrasının operativ idarə olunması, inzibati fəaliyyətin nəticələrinin qeydiyyatı və təhlili. ERP sistemləri üzərinə qoyulmuş əsas tələblərdən biri verilənlər bazalarının mərkəzləşdirilməsi, vacib infrastrukturun fəaliyyəti üçün zəruri olan informasiyaların cəmləşdirməsi prosesini avtomatlaşdırmaqdır.

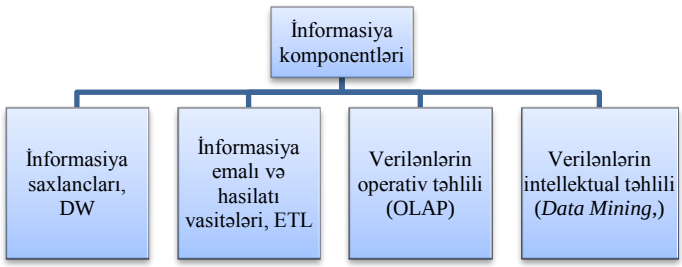
İnzibati idarəetmə sistemi verilənlər bazası, tətbiqi proqramlar və xidmətin təqdimatı səviyyələrindən ibarət üç-pilləli kliyent-server (bulud texnologiyasinda kliyent-virtual server) arxitekturuna malikdir (şək. 1). Onların mühafizəsini ayrı-ayrılıqda deyil bir buludda qurmaq daha təhlükəsiz hesab olunur.

CRM-sistemi istifadəçilərlə strateji əlaqələri idarəetmə üzrə xüsusi proqram təminatıdır. Onun vasitəsilə xidmət keyfiyyətinin yüksədilməsi, istifadiçilərlə əlaqələrin tarixi haqqında məlumat saxlamaqla inzibati fəaliyyətin təkmilləşdirilməsi və nəticələrin gələcək təhlili təmin olunur.

Operativ idarəetmə səviyyəsi MES əsasında təşkil edilir. MES-sistemi istehsalat prosesini idarəetmə üzrə xüsusi proqram təminatıdır. Onun vasitəsilə istismar zamanı koordinasiya, sinxronlaşdırma, optimallaşdırma və təhlil problemlərinin həll olunması nəzərdə tutulmuşdur. Bu sistemin əsas modulları bunlardır: RAS (Resource Allocation and Status) - resursların bölüşdürülməsi və vəziyyətinin təyini; ODS (Operations/Detail Scheduling) - Operativ/təfsilati planlaşdırma; DPU (Dispatching Production Units) -istehsalatın planlaşması; DOC (Document Control) - Sənədlərin idarə olunması; DCA (Data Collection/Acquisition) -informasiyanın toplanması və saxlanması; LM (Labor Management) -İnsan resurslarının idarə edilməsi; QM (Quality Management) - Keyfiyyətin idarə edilməsi; PM (Process Management) - istismar proseslərinin idarə edilməsi; MM (Maintenance Management) -Texniki xidmət və təmirin idarə edilməsi; PTG (Product Tracking and Genealogy) -məhsulların izlənmə və şəcərəsi; PA (Performance Analysis) - məhsuladrlığın təhlili və s. MES sistemi müəssisəsinn struktur bölmələrini idarəetmə sinfinə aid edilir.

TP-AİS də şərti olaraq üç səviyyəli model ilə təsvir edilir. Yuxarı səviyyədə sistem operatorlarının avtomatlaşdırılmış iş yerləri (AİY) - vizuallaşdırma və nəzarət sistemləri üçün xüsusi proqram (*Supervisory Control And Data Acquisition*, SCADA və *Distributed Control Systems*, DCS) təminatına malik sənaye təyinatlı kompüterlər yerləşir. Aralıq səviyyədə məntiqi kontrollerlər (LC), informasiyanın giriş-çıxış modulları (İ/OM), tətbiqi proqram interfeysləri (APİ), fasiləsiz elektrik enerjisi təchizatı (UPS), terminallar (T), rabitə kabelləri (CC) yerləşir. Bu komponentlər giriş və çıxış siqnalların sayından asılı olaraq bir və ya bir neçə kaskadda konfiqurasiya edilir. TP-AİS modelinin aşağı səviyyəsində "sahə qurğuları" da adlandırılan çeviricilər (sensorlar, ölçmə-nəzarət və avtomatlaşdırma avadanlıqları), kabel sistemi yerləşdirilir.

Vacib ifrastrukturların informasiya-idarəetmə müstəvisində idarəetmə sistemlərinin əsas informasiya komponentləri yerləşdirilir. İnformasiya saxlancları (*Data Warehouse*, DW), informasiya emalı və hasilatı vasitələri (*Extract, Transform, Load*, ETL), verilənlərin operativ təhlili (*On-Line Analytical Processing*, OLAP), verilənlərin intellektual təhlili (*Data Mining*, DM) və s. müasir vacib ifrastrukturların idarəetmə sistemlərinin əsas informasiya komponentləridir (şək. 2).



Şəkil 2. Vacib ifrastrukturların informasiya komponentləri

Data Mining texnologiyasının tətbiqi ilə həll olunan məsələlər aşağıdakılardır [3]:

- Təsnifatlaşdırma (Classification)
- Klasterləşdirmə (Clustering)
- Assosasiyalaşdırma (Associations)
- Proqnozlaşdırma (Forecasting)
- Qiymətləndirmə (Estimation)

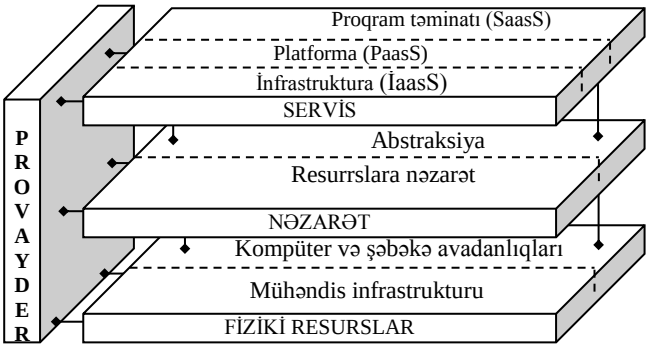
Vacib ifrastrukturların idarəetmə proseslərinin mahiyyətinə görə bu istiqamətlərin hər biri perspektivə malikdir. Məsələn, Data Mining tətbiqinin ən perspektivli istiqamətlərindən biri - bu texnologiyanın analitik CRM sistemlərində istifadə olmasiıdır. Data Mining texnologiyasının tətbiqi nəticəsində istifadəçilərin (müştərilərin) seqmentləşdirilməsi məsələsi həll olunur. Data Mining texnologiyası həmçinin bu seqmentlərin müəyyən təkliflərə reaksiaylarını proqnozlaşdırmaq üçün tətbiq oluna bilər.

III. VACİB INFRASTRUKTURLARIN İDARƏETMƏ SİSTEMİNİN İNFORMASİYA İNFRASTRUKTURU

Provayderlər vacib ifrastrukturların idarəetmə sisteminin informasiya infrastrukturunu bulud tenologiyasınım etalon modeli əsasında təşkil edir. Bu modeli müxtəlif funksional səviyyələrdə yerləşən müstəvilərin çoxluğu kimi təsvir etmək olar (şək. 3). Onun mərkəzində *servis*, *nəzarət* və *fiziki resurlar* səviyyələri, sol tərəfində bulud sisteminin təşkilatçısı olan *provayder* müstəvisi yerləşir.

Servis səviyyəsində (*Service Layer)* provayderin təqdim etdiyi bulud servisləri təyin olunur. Bu servislər üç baza modeli SPI (Software, Platform, Infrastructure as a Service) və törəmə modellər əsasında təqdim olunur [4].

Resursların abstraktsiya və nəzarəti səviyyəsində (*Resurs Abstraction və Control Level*) hipervizor, virtual yaddaş və bulud servislərini təqdim etməyə imkan verən infrastrukturun təşkili üçün proqram təminatı elementləri yerləşdirilir. Əsasən virtual maşın texnologiyası tətbiq olunmasına baxmayaraq bu səviyyəyə proqram təminatının abstraksiyası üzrə vasitələr əlavə oluna bilər.



Şəkil 3. Bulud hesablama sisteminin etalon modeli

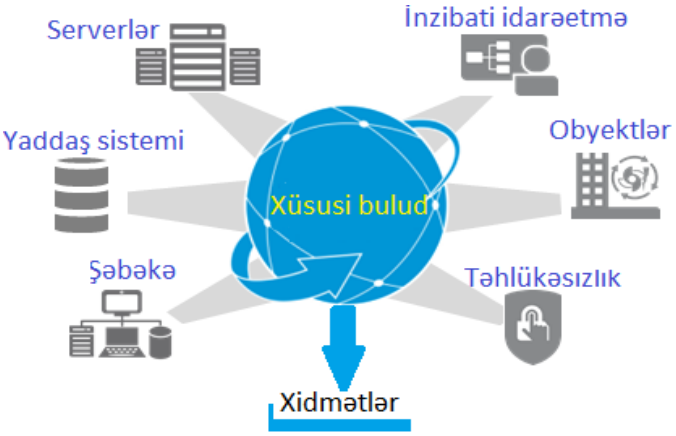
Fiziki resurslar (Physical Level) səviyyəsində zəruri texniki avadanlıqlar yerləşir:

✓ *Kompüter avadanlığı (Hardware):* Kompüter (CPU, HDD), şəbəkə (routerlər, firewall, kommutarorlar, rabitə kanalları, interfeyslər və s.), informasiya daşıyıcıları və hesablama infrastrukturunun digər fiziki elementləri.

✓ *Mühəndis infrastrukturu (Facilities):* kondisioner sistemləri (HVAC), elektrik qida qurğuları (UPS daxil olmaqla), kommunikasiya vasitələri və fiziki platformanın (sahənin) təşkili üzrə digər zəruri elementlər.

Bulud servislərini təqdim etmək üçün müxtəlif infrastruktur modelləri tətbiq olunur: *ümumi, fərdi, qrup və hibrid*.

Bütün resursların yalnız bir təşkilata məxsus olub, müəssisə daxilində təhlükəsiz fəaliyyət göstərən IT-infrastrukturu yaratmaq üçün ən səmərəli layihə xüsusi bulud (*private cloud*) texnologiyasıdır (şək. 4). Xüsusi buludun ümumi sistemdən üstünlüyü məxsusi müəssisə tərəfindən təşkil, nəzarət və idarə olunmasıdır. Bu, etibarlıq və təhlükəsizlik baxımdan daha idealdır. Təşkili baha və mürəkkəb olsa da xüsusi bulud sistemi müəssisədə maliyyə stabilliyi və yüksək səviyyəli İT xidmətləri təqdim edir [3].



Şəkil 4. Xüsusi bulud sistemi

NƏTİCƏ

Bulud sistemləri buraxıcılıq qabiliyyəti, idarəedici və zəruri məlumatların mərkəzləşdirilməsi, etibarlı saxlanması və emali hesabına vacib infrastruktur sisteminin daha effektli idarə olunmasını təmin edir. Data Mining texnologiyasının tətbiqi nəticəsində vacib ifrastrukturların hər bir idarəetmə səviyyəsi üzrə funksiyalrı daha səmərəli yerinə yetirilir və rəqabət davamlığı artır.

ƏDƏBİYYAT

[1] Головко М.В. Проекты ИС для крупных предприятий: от бессистемного управления к системам управления знаниями //Директору информационной службы, № 4, 2000. - С.

[2] 2.Управление производством на базе стандарта MPR II : 2-е изд. / Д. А. Гаврилов. – СПб. : Питер, 2008 – 368с.

[3] Bradley, P., Fayyad, U., Reina, C. Scaling Clustering Algorithms to Large Databases, Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, Calif., 1998.

[4] Cloud. http://resources.idgenterprise.com/original/AST-0032300_Understanding_the_Cloud_Computing_Stack.pdf, 2011,

AzScienceNet Şəbəkəsində İnternet-Resurslardan İstifadənin Monitorinqi Problemləri

Yadigar İmamverdiyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
yadigar@lan.ab.az

***Xülasə* — İnternet-resurslardan istifadə edilməsi təşkilatlarda əməkdaşların əsas iş alətinə çevrilib. İş vaxtı ərzində əməkdaşlar çoxsaylı İnternet-resurslara müraciət edirlər və bəzən bu resursların heç də hamısı bilavasitə iş məqsədləri ilə əlaqədar olmur. Analizlər göstərir ki, İnternet-resurslardan istifadənin effektivliyi məsələsi qlobal informasiya cəmiyyətində bütün təşkilatları maraqlandırır, bir sıra hüquqi, sosial və etik problemləri əhatə edir. İnternetdən istifadəyə müəyyən maneələr yaratmaqla bu məsələni effektiv həll etmək mümkün deyil və ən məqbul yanaşma mövcud qanunvericilik əsasında təşkilatın İnternetdən istifadə siyasətini müəyyən etmək və onun əsasında İnternet-resurslardan istifadənin monitorinqini həyata keçirməkdir. Bu işdə təşkilatlarda İnternet-resurslardan istifadə ilə bağlı problemlər və onların həllinə mövcud yanaşmalar analiz edilir, AzScienceNet şəbəkəsində İnternet-resurslardan istifadənin monitorinqi məsələləri üçün verilənlərin intellektual analizi metodlarının tətbiqi imkanları araşdırılır.**

Açar sözlər — İnternet resursları; monitorinq; tematik profil; vebdən istifadənin intellektual analizi

I. GİRİŞ

İnternet-resurslardan istifadə edilməsi təhsil, elm, dövlət və kommersiya təşkilatlarının fəaliyyətində səmərəli iş alətinə çevrilib. Əksər hallarda İnternetdən istifadə zamanı təşkilat əməkdaşlarının fəaliyyətinin aparıcı motivləri yeni bilik əldə etmək və əsas məqsədləri isə informasiya axtarışı olur. Lakin bəzi əməkdaşlar müxtəlif şəxsi məqsədlər üçün İnternet “gəzintisinə” iş vaxtının əhəmiyyətli hissəsini sərf edirlər, nəticədə təşkilatın ümumi əmək məhsuldarlığı aşağı düşür. Üstəlik, əməkdaşların İnternetdə “əyləncəsini” də təşkilat ödəyir. Dünyada iş vaxtı əməkdaşların İnternetdən işdənkənar məqsədlərlə istifadə etməsi hadisəsi kifayət qədər geniş yayılıb və bu hadisə üçün hələ 2000-ci illərdən xüsusi termin də – kiberslekinq (ing. cyberslacking) işlədilir [1]. Eyni zamanda, əməkdaşların İnternetdən tənzimlənməyən istifadəsi ciddi informasiya təhlükəsizliyi təhdidləri də yarada bilər, o cümlədən qanunvericiliklə qadağan olunmuş İnternet-resurslardan istifadə edilməsi və istifadəçilərin arzuolunmaz kontentlə təmasları da baş verə bilər.

AzScienceNet elm-kompüter şəbəkəsi Azərbaycan Milli Elmlər Akademiyasının (AMEA) institut və təşkilatlarının fəaliyyətinin səmərəli təşkilinə, alim və mütəxəssislərin dünya elmi mühitinə inteqrasiya olunmasına, dünyanın tanınmış beynəlxalq elmi bazalarından geniş istifadəsinə imkan yaradılması məqsədilə fəaliyyət göstərir AzScienceNet AMEA institut və təşkilatlarını elmi-tədqiqat, elmi-praktiki və tədris məsələlərinin həyata keçirilməsi üçün zəruri olan müasir şəbəkə xidmətləri ilə təmin edir. Hazırda AzScienceNet şəbəkəsi vasitəsilə İnternetə qoşulan istifadəçilərin sayı 5000-dən çoxdur [2].

AzScienceNet-də şəbəkə xidmətlərinin keyfiyyət göstəricilərinin yaxşılaşdırılması, şəbəkə resurslarının daha səmərəli idarə edilməsi və istifadə olunması, informasiya təhlükəsizliyinin təmin olunması vəziyyətinə nəzarətin həyata keçirilməsi üçün istifadəçilərin İnternet-resurslardan istifadəsinin operativ monitorinqi həyata keçirilir.

Bu işdə elmi ədəbiyyatın analizi əsasında təşkilatlarda İnternet-resurslardan istifadə edilməsinin monitorinqi məsələsinə mövcud yanaşmalar araşdırılır, monitorinqin effektivliyini artırmaq üçün veb-resurslara müraciətlərin loq-fayllarında toplanmış verilənlərə intellektual analiz üsullarının tətbiqi məsələsinə baxılır.

II. TƏŞKİLATLARDA İNTERNETİN İSTİFADƏSİ İLƏ BAĞLI PROBLEMLƏR

İnternetə idarə olunmayan, nəzarətsiz giriş kollektivdə əmək məhsuldarlığını xeyli aşağı salır, çünki iş vaxtının müəyyən hissəsi işdənkənar resurslara sərf olunur [3]. Qeyd edək ki, iş yerlərində smartfonların həddindən artıq istifadəsi də əmək məhsuldarlığının aşağı düşməsinə səbəb olur [4].

Əməkdaşlar sürət böyük olduğu üçün işdə İnternetə evdəkindən daha tez-tez müraciət edirlər. IDC şirkətinin məlumatına görə, iş yerində İnternetdən istifadənin 30-40%-i işlə əlaqədar deyil [3] və təşkilatların əməkdaşları iş vaxtının təxminən 1/3-ni onların işi ilə əlaqədar olmayan məqsədlərlə İnternetdə keçirirlər (şəbəkə oyunları, internet mağazalar, informasiya axtarışı). Əməkdaşların 30%-i işdə idman yarışlarını onlayn izləyir, 24%-i onlayn alış-verişi işdə edirlər. Onlayn birja satışlarının 92%-i iş zamanı iş yerindən edilir.

İnternetdən nəzarətsiz istifadənin daha bir nəticəsi – şəbəkədə sürətin (buraxılış qabiliyyətinin) aşağı düşməsidir. Statistikaya görə, təşkilatların əməkdaşlarının 44%-i korporativ resurslardan videoya baxmaq, audioyazıları dinləmək, şəbəkə oyunları oynamaq, böyük həcmli faylları yükləmək (filmlər, musiqi, şəkillər və s.) üçün istifadə edirlər. Nəticədə şəbəkəyə əhəmiyyətli yük düşür.

İnternet zərərli proqramların (virus, troyan və s.) yayılmasının əsas mənbəyidir. Hazırda zərərli proqramların yayılması üçün İnternet (e-poçt, veb-saytlar, sosial şəbəkələr) geniş istifadə edilir.

İnternet təşkilatın lokal şəbəkələrinə, serverlərinə və kompüterlərinə hücumların həyata keçirilməsi kanalıdır. İnternet-resursların çoxunda müxtəlif proqram – JavaScript, Flash, ActiveX və s. kodları olur. Bədniyyətlilər korporativ şəbəkələrə və istifadəçi kompüterlərinə hücumu təşkil etmək üçün bu kodlardan istifadə edə bilərlər [5].

İnternet konfidensial informasiyanın sızmasının əsas kanallarından biridir. Məsələn, əməkdaşların pulsuz e-poçt xidmətlərindən istifadə etməsi təşkilatın informasiya resurslarının konfidensiallığına ciddi təhdidlər yarada bilər. Statistikaya əsasən, əməkdaşların 80%-i pulsuz e-poçt xidmətlərindən istifadə edir, lakin bu halın təşkilatda konfidensiallığın təmin edilməsinə yaratdığı təhlükədən xəbərdar olsalar da, təşkilatların yalnız dörddə biri pulsuz e-poçtdan istifadəni qadağan edir.

III. TƏŞKİLATLARDA İNTERNETDƏN İSTİFADƏ PROBLEMİNİN HƏLLİNƏ YANAŞMALAR

Veb-resursların sürətli inkişafının hələ lap başlanğıcında əməkdaşların İnternetdən istifadəsinin effektiv tənzimlənməsi təşkilatlarda problemə çevrilmişdi. Təşkilatlar müxtəlif nəzarət mexanizmlərindən istifadə etməyə çalışırdılar [3,6-9].

Bəzi təşkilatlarda İnternetdən işdənkənar istifadə üçün cərimələr tətbiq edilirdi. Cərimələrin məbləği əməkdaşın şəxsi məqsədlər üçün İnternetdə sərf etdiyi vaxtdan və şirkətin əməkdaşa ödədiyi iş vaxtının dəyərindən asılı idi. Əməkdaşların İnternetdən istifadəsinin detallı uçotu aparılır (hansı kontentə baxılıb, hansı məlumat, hansı istifadəçi tərəfindən, nə zaman yüklənib və s.) və bunun əsasında rəhbərlik işdən çıxarılma da daxil olmaqla inzibati tənbeh üsullarından istifadə edirdi [3].

KLegal hüquq şirkəti ilə Personnel Today jurnalı tərəfindən 2002-ci ildə aparılmış tədqiqat göstərmişdi ki, İnternetin və e-poçtun işdənkənar məqsədlər üçün istifadə edilməsi əməkdaşların işdən çıxarılmasının ən geniş yayılmış səbəbidir. Tədqiqatlar Böyük Britaniyada 212 şirkəti əhatə edirdi. Hər on şirkətdən birində rəhbərlik İnternetə çıxışı tamamilə bağlamağa məcbur olmuşdu, təşkilatların 30 %-ində iş zamanı İnternetin istifadəsi ciddi məhdudlaşdırılmışdı [10].

Aydındır ki, cərimələr və işdən çıxarılmalar heç də ən effektiv tədbir deyil (texniki dəstək xidmətindən alınmış siyahıların daim nəzərdən keçirilməsi zərurəti də rəhbərliyə əlavə yükə çevrilir və tədricən lazımsız olur).

Coker B. tərəfindən 2011-ci ildə təsadüfi seçilmiş 700 ofis əməkdaşı arasında aparılmış sorğuda bəzi maraqlı qanunauyğunluqlar aşkarlanmışdı. Sorğunun nəticələrinə görə, İnternetdən istifadəyə iş vaxtının 12%-dən çoxu sərf edilmədikdə İnternetdən istifadə əmək məhsuldarlığı ilə müsbət korrelyasiyaya malik olur. İş yerində İnternetdən istifadə edən əməkdaşların əmək məhsuldarlığı iş yerində İnternetdən istifadə etməyən həmkarları ilə müqayisədə 9% artır [11]. Lakin vebdən həddindən artıq istifadə əmək məhsuldarlığının aşağı düşməsinə səbəb olur, çünki əməkdaşlar tapşırılmış işi planlaşdırılmış vaxt ərzində başa çatdıra bilmirlər.

Analizlər göstərir ki, təşkilatlarda İnternetdən təhlükəsiz və məhsuldar istifadə probleminin həllinə iki yanaşma mövcuddur:

- zərurət olmadan İnternetin istifadəsinin qadağan edilməsi (“radikal” yanaşma);
- istifadəçiyə “qadağan edilməyən hər şeyə icazə verilib” prinsipi ilə hərəkət etməyə icazə verilməsi. İstifadəçi İnternet-resurslardan istifadə edir, lakin onun hərəkətlərinə nəzarət (monitorinq) edilir.

Beləliklə, təşkilatlarda İnternet-resurslardan effektiv istifadə edilməsi probleminin həllinə kompleks yanaşma vacibdir: normativ-hüquqi sənədlərin işlənməsi və qəbul edilməsi; istifadəçilərin maarifləndirilməsi, texniki nəzarət və inzibati tədbirlər. Məsələnin texniki həlli – İnternetdən istifadə üzrə reqlamentin müəyyən edilməsi, internetə çıxışın tənzimlənməsi, hər bir əməkdaşın şəbəkə resurslarından istifadəsinin monitorinq edilməsindən ibarətdir. Nəzarətin konkret reallaşdırılması təşkilatın daxili siyasətindən və korporativ mədəniyyətindən asılıdır:

- iş yerindən İnternet şəbəkəsinə giriş məhdudlaşdırılır, yalnız fasilə zamanı və işdən sonra əməkdaşlara İnternetə sərbəst girişinə icazə verilə bilər;
- iş vaxtında idman, xəbər və əyləncə saytlarına giriş qadağan olunur;
- müəyyən veb-səhifələrə və xidmətlərə girişə icazə verilir;
- kənar e-poçt sistemlərinə girişə məhdudiyyət qoyulur;
- qeyri-məhdud girişə icazə verilir, lakin monitorinq aparılır;
- hər bir istifadəçi üçün konkret giriş növünə icazə verilir;
- gün ərzində istifadə edilən trafikin həcminə məhdudiyyət qoyulur.

İnternet-resurslardan istifadəyə çevik nəzarəti təmin etmək üçün təşkilatda İnternet-resurslardan istifadə siyasətinin işlənməsi və həyata keçirilməsi zəruridir. Bu siyasətin həyata keçirilməsi istifadəçilərin fəaliyyətinin monitorinqini tələb edir və kontent-filtrasiya sistemlərinin istifadəsi bu məsələnin həlli üçün effektiv vasitə hesab edilir. Bu sistemlər istifadəçilərin İnternet-resurslara tematik girişini təmin etməyə yönəlir və İnternet-resursların analizinə və kateqoriyalar üzrə təsnif edilməsinə əsaslanır (URL-filtrasiya). Kateqoriyalar üzrə təsnif edilmiş resursların keyfiyyətli bazası əsasında bu sistemlər ingilis dilli resurslarla yaxşı işləyə bilir. Digər dillər üçün belə sistemlərdə URL-bazasının natamamlığı və operativ yenilənməməsinə görə problemlər yarana bilər.

Qeyd edək ki, AzScienceNet şəbəkəsində istifadəçilərin İnternet-resurslardan istifadəsinə hər hansı bir məhdudiyyət yoxdur. AzScienceNet şəbəkəsinin əsas vəzifəsi günün 24 saatı ərzində istifadəçilərinin qlobal İnternet şəbəkəsinə yüksək sürətli, təhlükəsiz çıxışını təmin etməkdir və bu vəzifənin səmərəli həyata keçirilməsi üçün “AzScienc

eNet şəbəkəsindən istifadə siyasəti” [12] və “AzScienceNet şəbəkəsində informasiya təhlükəsizliyi siyasəti” [13] qəbul edilmişdir. Bu siyasətlərin həyata keçirilməsi vəziyyətini qiymətləndirmək üçün AzScienceNet Monitorinq və İnformasiya Təhlükəsizliyi Xidməti tərəfindən vaxtaşırı monitorinq və analiz işləri aparılır.

IV. İNTERNET-RESURSLARDAN İSTİFADƏNİN MONİTORİNQİ ÜÇÜN ALƏTLƏR

İnternet-resurslardan istifadə vəziyyətini izləməyə və bir sıra statistik hesabatlar almağa imkan verən xüsusi proqram təminatı istifadə edilir. Fərdi kompüterlər (NetPolice, iCencor və s.) və təşkilatın korporativ şəbəkəsi üçün həllər və proqram təminatı (UserGate Web Filter, ICS-Ideco, Internet Access Monitor, Websense Web Security Gateway) vardır. Veb-brauzerlərdə də kontent-filtrasiya imkanları mövcuddur.

İnternet-resurslardan istifadənin monitorinqi sistemlərində aşağıdakı avtomatik məhdudiyyətlər də ola bilər:

- müəyyən saytlara girişin qadağan edilməsi;
- müəyyən İnternet-protokolların qadağan edilməsi;
- İnternet-trafikin antivirusla yoxlanması və viruslu faylı qəbul etməkdən avtomatik imtina;
- faylların müəyyən tiplərinin yüklənməsinin qadağan edilməsi;
- günlük, həftəlik və aylıq limitlərə çatdıqda istifadəçilərin və (ya) istifadəçi qruplarının ödənişli resurslardan avtomatik ayrılması.

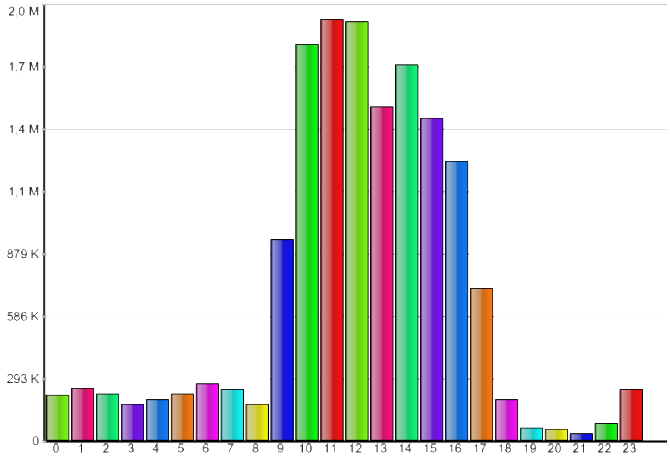
Bir sıra həllərdə ISS (Internet Security Systems) şirkətinin təqdim etdiyi İnternet-resursların kateqoriyalar bazası istifadə edilir, bu baza daha tam və aktual URL-baza hesab edilir. URL-bazada milyonlarla veb-saytın ünvanı var və hər gün yenilənir. Veb-saytın kateqoriyası kontent-filtrasiyanın ilkin mərhələlərində müəyyən edilir və çox vaxt tələb etmir, kifayət qədər effektivdir və verilənlərin emal vaxtını azaldır.

Müasir təşkilatların İnternetə çıxışının tipik üsulu xüsusi-proqramlardan – proksi-serverlərdən istifadə olunmasıdır. Proksi-server İnternet bağlantını bütün əməkdaşlar arasında bölməyə imkan verir. Bu proqramın yaratdığı loq-faylı analiz etməklə kimin nə vaxt, hansı sayta müraciət etməsi haqqında məlumat əldə etmək olar.

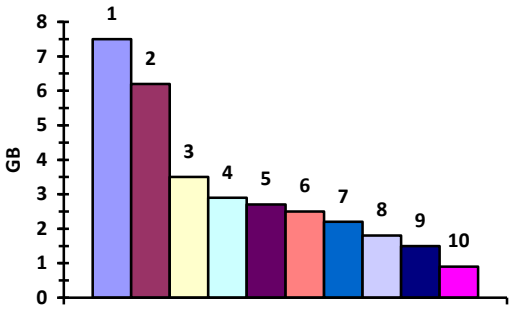
AzScienceNet şəbəkəsində istifadəçilərin veb saytlara müraciətlərinin statistikasını əldə etmək üçün Squid proksi-serverinin loq fayllarının analizi üçün WebSpy Vantage proqram təminatı istifadə edilir. Proqram aşağıdakı hesabatları almağa imkan verir:

- trafikin verilənlərin tipinə görə (şəkil, video, audio, mətn) paylanması;
- trafikin veb-saytların kateqoriyaları üzrə paylanması;
- trafikin istifadəçilər, istifadəçilər qrupu (təşkilatlar), protokollar üzrə paylanması;
- trafikin saatlar, həftənin günləri, aylar və tarixlər üzrə paylanması;
- trafikin istifadəçinin müraciət etdiyi saytlar üzrə paylanması;
- istifadəçinin İnternetə giriş və çıxış zamanları.

Göründüyü ki, proqram təminatı loq-fayl əsasında sadə statistik hesabatlar almağa imkan verir. Şəkil 1-də AMEA təşkilatlarından birində istifadəçilərin müraciətlərinin saatlar üzrə paylanması göstərilir (ordinat oxunda K – min, M – milyon bildirir, müraciətlərin sayı olaraq hitlərin sayı götürülüb). Şəkil 2-də isə həmin təşkilatda bəzi istifadəçilərin bir həftə ərzində İnternetdən yüklədikləri informasiyanın həcmi (qiqabaytlarla –GB) göstərilmişdir, istifadəçilər adsızlaşdırılıb və rəqəmlərlə göstərilir. Loq-fayllarda toplanmış informasiyanın daha dərindən analizi üçün verilənlərin intellektual analizi metodlarından istifadə edilməlidir. Bu məqsədlə növbəti bölmədə vebdən istifadənin intellektual analizi metodlarının qısa icmalına baxılır. Qeyd edək ki, bu işdə veb-saytların kontentinin analizinə baxılmır (Deep Content Inspection).



Şəkil 1. İstifadəçilərin müraciətlərinin saatlar üzrə paylanması



Şəkil 2. İnternetdən ən çox informasiya yükləyən istifadəçilər

V. VEB-DƏN İSTİFADƏNİN İNTELLEKTUAL ANALİZİ METODLARI

Veb-də toplanmış verilənlərin intellektual analizi metodları (ing. Web Mining) 1990-cı illərin ortalarından inkişaf edir və hazırda aktual tədqiqat istiqamətlərindən biridir [14,15]. Web Mining üç istiqamətə ayrılır: veb kontentin, veb strukturun və vebdən istifadənin intellektual analizi. Vebdən istifadənin intellektual analizi (Web Usage Mining, WUM) vebdən istifadə şablonlarını aşkarlamaq üçün istifadəçinin veblə qarşılıqlı əlaqəsi zamanı toplanmış verilənlərə intellektual analiz üsullarının tətbiqidir [16]. Analiz üçün veb məlumatları əsasən veb-serverin və ya proksi-serverin (bəzən veb-brauzerin) loq-fayllarından alınır. Vebdən istifadənin intellektual analizi veb-saytların dizaynını təkmilləşdirmək, istifadəçiləri cəlb etmək və daimi istifadəçilərə fərdi və adaptiv xidmətlər təklif etmək üçün istifadəçilərin davranışını aşkarlamağa və proqnozlaşdırmağa çalışır [17,18,19]. WUM yüklənmənin balanslaşdırılması və veb-trafikin effektiv idarə edilməsi, kiber-hücumlarla əlaqədar şübhəli aktivliyin aşkarlanması [20], istifadəçilərin müxtəlif qrupları üçün populyar mövzuların aşkarlanması, elm-tədqiqat prosesləri baxımındam veb-saytların qiymətləndirilməsi üçün də istifadə edilə bilər.

AMEA İnformasiya Texnologiyaları İnstitutunda vebdən istifadənin intellektual analizi üzrə bir sıra tədqiqt işləri aparılmışdır. Məsələn, [21]-də veb-trafikin intellektual analizi metodlarına baxılır, trafikin klaster analizi nəticəsində veb-istifadəçilərin aktivliyinin modelləşdirilməsi məsələsi həll edilir. [22]-də veb-istifadəçilərin maraqlarının avtomatik identifikasiyası məsələsinə baxılır. [23,24]-də veb-serverlərin loq-fayllarının analizi əsasında veb-istifadəçilərin klasterizasiyası modeli təklif edilir.

[25]-də istifadəçilərin müraciət etdikləri İnternet-resursları haqqında verilənlərin analizi əsasında qurulmuş tematik profillərin klasterizasiyası metodu təklif olunur. Məqalədə ilkin verilənlər, əlamətlərin qiymətləndirilməsi və seçilməsi prosesləri, aqlomerativ iyerarxik klasterizasiya metodu və k-means metodu istifadə edilməklə aparılmış eksperimentlər ətraflı təsvir olunur.

VI. TƏKLİF EDİLMİŞ YANAŞMANIN MAHİYYƏTİ

Vebdən istifadənin intellektual analizi üç mərhələdən ibarətdir: verilənlərin ilkin emalı, şablonun aşkarlanması və şablonun analizi.

Verilənlərin ilkin emalı mərhələsində proksi-serverin loq-faylından alınmış verilənlər əsasında hər bir istifadəçi üçün əlamətlər vektoru formalaşdırılır. Proksi-serverin loq-faylında hər bir müraciətin zamanı, müddəti, istifadəçinin IP-ünvanı, müraciət edilən veb-resursun IP-ünvanı və ya DNS-adı, İnternetdən alınmış informasiyanın həcmi haqqında məlumat yazılır. Beləliklə, loq-fayldan intellektual analiz üçün zəruri olan verilənləri almaq olar. Şəkil 3-də Squid proksi-serverinin loq-faylından bir fraqment göstərilib.

```
1444776131.155    515 10.100.80.23 TCP_MISS/403 810 GET
http://data.icecentury.com/product/wifi/ - HIER_DIRECT/54.169.165.185
text/html1444776131.187    547 10.100.80.23 TCP_MISS/403 813 GET
http://data.icecentury.com/product/cleaner/ -
HIER_DIRECT/54.169.165.185 text/html1444776131.198    560
10.100.80.23 TCP_MISS/301 568 GET http://v.icecentury.com/upgrade? -
HIER_DIRECT/54.169.165.185 text/html 1444776131.244    528
10.100.80.23 TCP_MISS/302 250 GET
http://market.voga360.com/frontend/checkpc.htm? -
HIER_DIRECT/54.255.133.240 –
```

Şəkil 3. Squid loq-faylından bir fraqment

Əlamətlər vektorunu formalaşdırmaq üçün istifadəçinin hitlər axını seanslara bölünür. Seans – İnternetdə fasiləsiz aktivlik intervallarıdır. Seansları müəyyən etmək üçün bir sıra yanaşmalar var, bu işdə aşağıdakı iki sadə qayda istifadə edilir:

1) əgər iki hit arasındakı interval 20 dəqiqədən çox deyilsə, onda onlar bir seansa aiddir.

2) Seans axırıncı hitdən sonra 20 dəqiqə davam edir.

Alınmış seanslar əsasında istifadəçilərin veb-resurslara müraciətlər xronologiyasını matrisə çevirmək olar. Matrisin sətirlərində istifadəçilər, sütunlarında əlamətlər verilir. Əlamətlər kimi veb-saytların tematik-kateqoriyaları götürülür (məsələn, ISS-in URL bazasına əsasən).

Konkret istifadəçi üçün əlamətin qiyməti olaraq bu əlamətin rast gəldiyi istifadəçi seanslarının sayı götürülür. Əlamətlərin qiymətini istifadəçinin hər bir tematik kateqoriyaya daxil olan veb saytlara müraciət tezlikləri və daxil olan trafikin həcmi əsasında da hesablamaq olar. Modelin keyfiyyətini yaxşılaşdırmaq üçün əlamətlərin qiymətləri [0...1] intervalında normallaşdırılır.

Əlamətlərin qiymətləri hesablandıqdan sonra (feature engineering) modelin qurulması üçün ən informativ və statistik etibarlı əlamətlər seçilməlidir. Bu emal edilən informasiyanın həcmini azaltmağa, izafi təlimdən qaçmağa və bütövlükdə modelin keyfiyyətini yaxşılaşdırmağa xidmət edir. Yalnız bir istifadəçidə rast gəlinən əlamətlər qeyri-informativ hesab olunur.

Nəticədə yalnız informativ əlamətlərdən ibarət olan “istifadəçi/kateqoriya” matrisi alınır. Adətən, bu matrisin ölçüsü böyük olur, həm də bu matris seyrək də ola bilər (sparse matrix). Buna görə belə matrislərlə işləyə bilən maşın təlimi alqoritmlərindən istifadə edilməlidir. Alınmış matris klasterizasiya metodları ilə emal olunur [26]. Eksperimentlər üçün WEKA proqram təminatı istifadə edilir [27].

İstifadəçilərin tematik profilləri müəyyən edildikdən sonra veb-saytların tematik kateqoriyalarına uyğun çəkilər verməklə istifadəçiləri tematik profillərinə görə ranqlaşdırmaq da mümkündür.

NƏTİCƏ

İstifadəçilərin veb-resurslara müraciətləri haqqında loq-fayllarda toplanmış verilənlərin intellektual analizi istifadəçilərin və istifadəçi qruplarının maraqları və seçimləri haqqında fikir yürütməyə, təşkilatda İnternet-resurslardan nə dərəcədə məqsədyönlü istifadə edilməsi barədə nəticə çıxarmağa imkan verir. İstifadəçi seçimlərinin klasterləri haqqında informasiya əməkdaşların işinin effektivliyini qiymətləndirməyə, istifadəçi qruplarının informasiya tələbatlarını müəyyən etməyə və İT-infrastrukturun inkişafına investisiyaları əsaslandırmağa imkan verir.

ƏDƏBİYYAT

[1] W. Block, “Cyberslacking, business ethics and managerial economics,” Journal of Business Ethics, Vol. 33, No. 3, pp. 225-231, 2001.

[2] AzScienceNet-in veb-saytı http://azsciencenet.az/az

[3] M. Anandarjan, C.A.Simmers. Managing Web Usage in the Workplace: A Social, Ethical and Legal Perspective. IGI Publishing Hershey, 2002.

[4] T. Pitichat, “Smartphones in the workplace: Changing organizational behaviour, transforming the future,” A Journal of Trans-disciplinary Writing and Research from Claremont Graduate University, Vol. 3, No. 1, pp.1-11, 2013.

[5] О. Слепов, А. Отт, “Контроль использования интернет-ресурсов,” Jet Info Информационный Бюллетень, № 2 (141), 2005.

[6] L. Qinyu, L. Xin, G. Anil, L. Long, “Workplace Management And Employee Misuse: Does Punishment Matter?” Journal of Computer Information Systems, Vol. 50, No. 2, pp. 49-59, 2009.

[7] J. C. Ugrin, J. M. Pearson, “Exploring Internet Abuse in the Workplace: How Can We Maximize Deterrence Efforts?” *Review of Business*, Vol. 28, No. 2, pp.29-40, 2008.

[8] H. Jiang, A. Tsohou, “Dual nature of personal web usage in workplace: Impacts, antecedents and regulating policies,” Proc. of the 22nd European Conference on Information Systems (ECIS), pp.1-16, 2014.

[9] P. Mahatanankoon, M. Anandarajan, M. Igbaria, “Development of a Measure of Personal Web Usage in the Workplace,” Cyberpsychology & Behavior, Vol 7, No 1, pp. 93-104, 2004.

[10] Internet abuse is top reason for workplace discipline. http://www.computerweekly.com/news/2240047296/Internet-abuse-is-top-reason-for-workplace-discipline

[11] B. L. Coker, “Freedom to surf: the positive effects of workplace Internet leisure browsing,” New Technology, Work & Employment, Vol. 26, No. 3, pp. 238-247, 2011.

[12] R.Ələkbərov, “AzScienceNet şəbəkəsinin informasiya təhlükəsizliyi siyasəti haqqında,” Azərbaycan xalqının ümummilli lideri Heydər Əliyevin 90 illik yubileyinə həsr olunmuş “İnformasiya təhlükəsizliyi problemləri üzrə I respublika elmi-praktiki konfransı” nın əsərləri, səh. 126-127, 2015.

[13] B.Nəbiyev, “AzScienceNet şəbəkəsində informasiya təhlükəsizliyinin monitorinqi sistemi,” Azərbaycan xalqının ümummilli lideri Heydər Əliyevin 90 illik yubileyinə həsr olunmuş “İnformasiya təhlükəsizliyi problemləri üzrə I respublika elmi-praktiki konfransı” nın əsərləri, səh. 128-130, 2015.

[14] Q. Zhang, R. Segall, “Web mining: a survey of current research, techniques and software,” International Journal of Information Technology & Decision Making, Vol. 7, No. 4, pp. 683-720 ,2008.

[15] K. Sharma, G. Shrivastava, V. Kumar, “Web mining: Today and tomorrow,” Proc. Of the 3rd International Conference on Electronics Computer Technology (ICECT), Vol. 1, pp.399-403, 2011.

[16] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, “Web usage mining: discovery and applications of usage patterns from Web data,” ACM SIGKDD Explorations Newsletter, Vol. 1, No. 2, pp. 12-23, 2000.

[17] J. Pei, J. Han, B. Mortazavi, H. Zhu, “Mining Access Patterns efficiently from Web Logs,” Proc. of the 4th Pacific-Asia Conf. Knowledge discovery and Data Mining (PAKDD’00), pp. 396-407, 2000.

[18] B. Berendt, M. Spiliopolou, “Analysis of navigation behaviour in web sites integrating multiple information systems,” VLDB Journal, Vol. 9,No. 1, pp. 56-75, 2000.

[19] M. Eirinaki, M. Vazirgiannis, “Web mining for web personalization,” ACM Transactions on Internet Technology (TOIT), Vol. 3, No. 1, pp. 1-27, 2003.

[20] D. Stevanovic, N. Vlajic, A. An, “Unsupervised Clustering of Web Sessions to Detect Malicious and Non-malicious Website Users,” Procedia Computer Science, Vol. 5, pp. 123-131, 2011.

[21] Р.М. Алыгулиев, Ф.Ф. Юсифов, Разработка модели для изучения активности пользователей Web с помощью анализа Web-трафика,” Автоматика и вычислительная техника, No. 4, С. 73-77, 2007.

[22] Р.М. Алгулиев, Р.М. Алыгулиев, Ф.Ф. Юсифов, “Автоматическая идентификация интересов Web-пользователя,” Автоматика и вычислительная техника, Vol.41, No.6, с.32-47, 2007.

[23] Р.М. Алгулиев, Ф.Ф. Юсифов, “Кластеризация Web-пользователей на основе анализа посещаемости Web-сайтов,” Телекоммуникации, No: 2, С. 13-17, 2008.

[24] Ф.Ф. Юсифов, “Извлечение знаний из Интернет с использованием лог-файлов,” Телекоммуникации, №8, с. 13–18, 2006.

[25] К.Г., Паутов, Ф.А. Попов, “Метод кластеризации тематических профилей пользователей и его применение для анализа Интернет-трафика,” Фундаментальные Исследования, № 7 (часть 4), с.765-769, 2015.

[26] J. Han, M. Kamber, Data Mining: Concepts and Techniques. 2nd edition. Morgan Kaufmann, 2006.

[27] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/

Uşaqların İnternet Təhlükəsizliyində Big Data Texnologiyalarından İstifadə Məsələləri

Sabirə Allahverdiyeva
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
allahverdiyevasabira@gmail.com

***Xülasə* — Məqalədə Big Data uşaqların həyatında oynadığı rol, rəqəmsal əşyaların adi əşyalarla müqayisədə fərqlilikləri araşdırılmışdır. Uşaqların rəqəmsal texnologiyalara olan münasibəti göstərilmişdir. Big Data-nın uşaqların şəxsi məlumatlarından istifadədə müsbət və mənfi cəhətləri qeyd olunmuşdur.**

Açar sözlər — Big Data, məlumat bazaları, rəqəmsal əşyalar, əşyaların İnterneti, rəqəmsal oyuncaqlar, uşaqların təhlükəsizliyi.

I. GİRİŞ

Son bir neçə il ərzində Big Data ən çox müzakirə olunan aktual mövzu kimi yeniliyə çevrilib və demək olar ki, İnternetə olan baxışları tamamilə dəyişib. Big Data müasir texnologiyalar haqqında daha böyük məlumatlar əldə etməyə və qərarlar qəbul etməyə kömək edir.

İnternet inkişaf etdikcə texnologiyalarla qarşılıqlı münasibətlərin yeni formaları meydana gəlməyə başlayır. İnformasiya texnologiyaları və vasitələrinin inkişafı insanların informasiya əlaqələrini sürətləndirir.

Big Data demək olar ki, bir cox sahələrə tətbiq edilir. Təbiidir ki, bu tətbiqlərin ailələrə və uşaqlara yönəlməsi gözlənilən nəticədir.

Hər zaman ailələr övladlarının qorunması, təhlükəsizliyi, sağlamlığı yaxud onlara daha xoş həyatın təmin edilməsi üçün ən yaxşı yolları axtarıb tapmağa cəhdlər edirlər. Müasir dövrdə onlara bu yolda Big Data onlayn bələdçi olaraq köməklik göstərir.

II. RƏQƏMSAL ƏŞYALAR, ƏŞYALARIN İNTERNETİ VƏ UŞAQLAR

Son zamanlar rəqəmsal əşyaların gündəlik həyatımıza nüfuz etməsi və müxtəlif tip verilənlərin məlumat bazalarında yerləşdirilməsi aktualdır. Rəqəmsal əşyalar Big Data texnologiyalarının inkişafına təsir edir. Məlumatlar bazaya daxil edilir və bununla da müxtəlif araşdırmaların aparılmasına gətirib çıxarır. Əşyaların İnterneti (Əİ) məhz rəqəmsal texnologiyaların təzahürüdür. Bir-biri ilə əlaqəli kompüterlərin eyni zamanda əşyalarla bağlılığını göstərir. Zaman keçdikcə indiki İnternet mühitindən fərqli olaraq, bizi əhatə edən bütün əşyaların İnternetə qoşulması ilə Əİ reallaşacaq [1].

Uşaqların həyatında əşyalar xüsusi əhəmiyyət kəsb edir. Xüsusilə daha azyaşlı uşaqlar üçün gündəlik fəaliyyət, adətən, oyuncaqlardan ibarətdir. Rəqəmsal oyuncaqlar mövcud olduğu müddətdə uşaqların işıq və ya siqnallarla qarşılıqlı münasibətinin rəqəmsal təsiri zaman və məkan çərçivəsində birbaşa manipulyasiyaya səbəb olur.

Uşaqların rəqəmsal texnologiyalara olan münasibətində fərqliliklər yaranır. Onların istifadə etdikləri maus və klaviatura qurğuları artıq öz təyinatını dəyişməyə başlamışdır. İnterfeys və qurğular sensor ekran və jestlərin tanınma qurğuları kimi müasir qarşılıqlı münasibət formaları ilə əvəz edilmişdir. Uşaqların münasibət və hərəkətlərini əhatə edən çox sayda qurğular mövcuddur. Məsələn, KidFit — uşaqların fiziki fəaliyyətini qeydə alan və ona cavab verən qola taxılan fitnes qurğusudur. HereO — valideynlərə uşaqlarının harada olması haqqında məlumat verən saatdır [2]. Spourting uşaq monitoru — uşaqların yuxu saatlarını qeydə alır və uşaq haqqında əks əlaqəni təmin edir [3].

Əİ-nin uşaqlar həyatında fəaliyyətini nəzər yetirmək üçün “Teddy the Guardian” oyuncağını nəzərdən keçirək [4]. Bu yumşaq, rəqəmsal ayı oyuncaqdır - uşaqların ürək döyüntüsü, oksigen çatışmazlığı, bədən temperaturu və stress səviyyəsi kimi vacib sağlamlıq parametrlərini qeydə alır və xəbər verir. Bu ayı, adətən ənənəvi testlərlə bağlı uşaqların narahatlığını əhəmiyyətli dərəcədə azaldan və bu məlumatların səhiyyə mütəxəssislərinə çatdırılması üçün xəstəxanalarda istifadə olunur.

Əİ uşaqların əşyalarla münasibəti haqqında məlumat əldə etməyə qadirdir. “Teddy the Guardian” topladığı biometrik məlumatlar səhiyyə mütəxəssisləri tərəfindən istifadə olunur. Lakin, həmin məlumatın səhiyyə şirkətlərindən başqa valideyn və digər şəxslər üçün nə dərəcədə dəyərli olduğunu təsəvvür etmək çətin deyil. Digər Əİ qurğuları daha az şəxsi məlumat toplamasına baxmayaraq uşaqların gündəlik fəaliyyətinə xüsusi təsir göstərə bilər. Nə qədər məlumatın toplandığı və istifadə edildiyini araşdırmaq çox vacibdir.

“Teddy the Guardian” uşaqlar üçün adi yumşaq oyuncaqdır. Uşaq oyuncağı qucaqlamaqla səhiyyə fəaliyyətinin interpretasiya olunmuş formasını həyata keçirir. Bu səbəbdən qarşılıqlı əlaqələrin yeni formaları uşaqların müxtəlif fəaliyyətini bir araya gətirə bilər. Oyuncağın səhiyyə xidməti ilə əlaqəsi haqqında uşağın biliyi həmin oyuncağın

qoruyucu kimi qavranılmasına səbəb ola bilər (oyuncağın adı da buradan götürülmüşdür: Guardian – qoruyucu, müdafiəçi).

Uşaqlar və onların valideynləri Əİ asılıdır və gündəlik əşya olaraq rəqəmsal texnologiyaların interfeysinə çevrilir. Texnologiyalar daha çox inkişaf etdiyinə görə uşaqlar öz hərəkətlərinin necə qeydə alındığı və istifadə edildiyi barədə xəbərdar deyildir. Uşaqlar, səbəbsiz şəkildə Teddyni qucaqlamaqla digər insanları şəxsi biometrik məlumatlarla təmin edir. Şübhəsiz ki, valideynlər öz uşaqları üçün tədbir görmək məqsədilə bu məlumatları əldə etməlidirlər. Lakin, Əİ-nin açıq olması həmin məlumatların mühafizəsi üçün böyük təhlükə yaradır.

III. BİG DATA-NIN UŞAQLARIN HƏYATINDA MƏNFİ VƏ MÜSBƏT CƏHƏTLƏRİ

Bu gün informasiya ayrı-ayrı şəxlərin, qrupların, təşkilatların özlərinin şəxsi maraqları üçün aşkar və gizli məqsədlərlə fərdi və sosial subyektlərin fəaliyyətinə ziyan vurmaq üçün həyata keçirdikləri psixoloji təsir vasitəsidir [5]. Uşaq və yeniyetmələr İnternet mühitində aktivdirlər. Yeniyetmələrin çoxu veb-gündəliklərə və ya bloqlara, sosial şəbəkələrə şəxsi məlumatlarını (adı, ailəsi haqqında, yaşayış yeri, oxuduğu məktəb, telefonları və s.) dolğun şəkildə yerləşdirir. Bu isə uşaqlara əlyetərliyə şərait yaradır, onların istismarına və cinayətkarların qurbanına çevirir.

Məsələ bundadır ki, Big Data vasitəsilə edilən kömək valideyinləri iki yol ayrıcında qoyur:

1) Big Data kömək edir ki, valideynlər öz uşaqları haqqında olan məlumatları (sağlamlıq, təhsil, rəftar, psixoloji vəziyyət və s.) əldə etməklə onlara nəzarət etsin;
2) Uşaqların şəxsi məlumatlarını Big Data vasitəsilə marketinq şirkətləri əldə edə bilər.

Qeyd etmək lazımdır ki, çox vaxt valideynləri narahat edən əsas cəhətlərdən biri məhz bu olur. Onlar düşünürlər: Big Data faydalıdırmı, faydasızdırmı və s.

Big Datanın müsbət cəhətlərini qeyd edək. Məsələn, yeni doğulmuş körpə haqqında ən kiçik məlumatlar məlumat bazasında toplanır və valideynin uşağı böyutməsində ona məsləhçi rolunu oynayır. Bunu Sproutling, a “FitBit for babies” proqramları vasitəsilə həyata keçirmək olar [6].

Uşaqların təhsil aldıqları dövrlərdə isə Big Datanın məlumatları müəllimlər tərəfindən həyata keçirilir. Bura uşaq və yeniyetmənin davranışı, rəftarı, fənn üzrə qavrama qabliyyəti, psixoloji vəziyyəti və s. aid etmək olar. Məlumat bazalarında yerləşən statistik göstəricilərə əsasən valideyn övladı haqqında qərar çıxarmaqda çətinlik çəkmir.

Məlumdur ki, öz uşaqları haqqında məlumat toplamaq bütün valideynlərin maraq dairəsidir. Çox vaxt uşaqlar haqqında toplanan bu statistik məlumatlar onların özlərinə qarşı müxtəlif məqsədlərlə yönələ bilər. Başqa bir tərəfi də ondadır ki, valideynlər uşaqları haqqında toplanmış məlumatlardan ciddi narahat ola bilərlər. Bəzən qurğunun, yaxud proqram təminatının nasazlığı da belə bir narahatçılığa səbəb ola bilər.

Texnologiyaların təsiri altında daha bir mənfi cəhət valideyn və uşaq arasında bağların itməsinə səbəb olmasıdır. Uşaqların hissləri nəzərə alınmır, məlumat bazalarında toplanmış məlumatlara görə valideyin tərəfindən qərar qəbul edilir. Sanki bir əşya, predmet kimi qiymətləndirilir. Lakin valideynlər unutmamalıdırlar ki, uşaqlar əşya deyil ki, ondan statistik məqsədlərlə istifadə olunsun.

Ümumiyyətlə, Big Data-nın mənfi və ya müsbət cəhətlərin hansının üstün olması barəsində siyasətçilər, ailələr, müəllimlər və s. əlaqədar insanlar arasında hər zaman fikir ayrılıqları olur. Müasir dünyada gedən prosesləri nəzərə alaraq məsələ bir qədər obyektiv yanaşsaq qeyd etmək lazımdır ki, Big Data-nın mənfi cəhətlərinə baxmayaraq, valideynə övladı haqqında məlumatların daha tez əldə etməsi üçün faydalıdır.

NƏTİCƏ

Uşaqların gündəlik fəaliyyətini əhatə edən qurğuların gündən-günə artması onların fəaliyyətini necə dəyişdirir, onların fəaliyyəti haqqında məlumat hansı məqsədlərlə istifadə olunur, uşaqlar və onların valideynləri həmin texnologiyalardan nə dərəcədə asılıdır? Bu qurğuların uşaqların davranışına nə dərəcədə təsir etməsi, həmin qurğular vasitəsilə əldə olunan məlumatın necə istifadə olunması, bu barədə uşaqların xəbərdar olub-olmaması, onların münasibətdə olduğu bu texnologiyalardan istifadəsi bütün bu problemlər haqqında araşdırmaların aparılmasını zəruri edir.

Zaman keçdikcə yeni texnologiyalar uşaqların həyatına inteqrasiya olunur. Bu qurğular rahatlıq, təhlükəsizlik və ya idman məqsədi üçün əlverişlidir. Buna baxmayaraq, həmin qurğuların meydana gətirdiyi bəzi problemlər vardır. Bu problemlərin daha dərindən tədqiq edilməsinin vacibliyi ortaya cıxır.

ƏDƏBİYYAT

[1] R.M.Əliquliyev, R.Ş Mahmudov. "Əşyaların interneti: mahiyyəti, imkanları və problemləri", İnformasiya cəmiyyəti problemləri, №2(4), 29-40, 2011.
[2] https://www.indiegogo.com/projects/hereo-the-first-gps-watch-designed-for-young-kids#/story
[3] A.Marches, P.Duncan, L.Plovman, S.Sabeti. " Three questions about the Internet of things and children", TechTrends №1. yanvar/fevral səh 76-77.
[4] http:// www.teddytheguardian.com
[5] R.Mahmudova, "İnformasiyanın neqativ təsirləri və onların aradan qaldırılması yolları", Beynəlxalq Telekommunikasiya İttifaqının 150 illiyinə həsr olunmuş “İnformasiya təhlükəsizliyinin multidissiplinar problemləri” üzrə II respublika elmi-praktiki konfransının əsərləri, Bakı, 14 may, 2015, s.74-77.
[6] http://telecareaware.com/tag/sproutling.

Crypto Usbtoken və Smartsd-də E-imzanın Tətbiqi

Həbib Abbasov
AMEA İnformasiya Texnologiyaları İnistutu, Bakı, Azərbaycan
hebib@rambler.ru

***Xülasə* — Müasir dövrdə informasiya texnalogiyalarının inkişafı informasiya təhlükəsizliyini daim vacib amil kimi xarakterizə edir. Bu baxımdan smart kart texnalogiyasından əlavə olaraq USBToken və SmartSD kart həllərində ümumi təhlükəsizliyi təmin edən mexanizimlərə çevrilmişdir. Elektron imzanın (e-imza) tətbiqi e-sənəd e-xidmətlərin təhlükəsizliyinin təmin edilməsində vacib yer tutur.**

E–imza yaratma və yoxlama məlumatlarının alqoritmlərinin standartlara uyğunluğu USBToken və SmartSD kart vasitəsi ilə elektron sənədlərin imzalanması və açarların generasiya vaxtı və mexanizmini təmin edilməsi tədqiqat əsasında aparılmalıdır. Məqaləd kriptoqrafik SmartSD kart və Crypto USBToken -lərin köməyi ilə Windows 7 əməliyyat sistemində müxtəlif ölçülü açarların generasiyası və elektron imzanın formalaşmasını təmin edən proqram təminatlarında aparılan ekspermentlərdə əldə edilmiş nəticələrin müqayisəsinə həsr olunmuşdur.

Açar sözlər — PKCS7, e-gov; e-imza; PKİ; e-sənəd; CSP, CA, RSA, SHA-1

I. GİRİŞ

İnformasiya təhlükəsizliyi mühitində məlumat mübadiləsində ötürülən məlumatların səhihliyinin təmin edilməsi və ötürənin identikləşdirilməsi vacib amildir. İnformasiya texnalogiyalarının inkişafı ilə cəmiyyətdə elektron həllərinin təhlükəsizliyinin təmin edilməsi vacibdir. Avropa Birliyinin bu sahədə qəbul etdiyi Direktivə[1] əsasən birliyə daxil olan ölkələr həmin standartlar əsasında işlərə başlamışdılar. Bir çox inkişaf etmiş ölkələrin təcrübəsinə nəzər saldıqda imza sahiblərinin istəklərinə və ya imza təqdim edən mərkəzlərin təklif etdiyi daşıyıcılarda (bura USBToken SmartSD kart daxildir) və s elektron imzaların formalaşmasını təmin edən açarlar və sertifikatlar daxil edilmişdir. Yuxarıda da qeyd edilən kimi e- sənədlərdə imzanın formalaşmasında e-imza daşıyıcıları və onların texniki xarakteristikaları vacib olan təhlükəsizlik meyarlarını özündə ehtiva edir.

II. E-İMZA HƏLLİNDƏ AÇARLARIN SAXLANMA VASİTƏLƏRİ

İmza yaratma (gizli açar) və imza yoxlama məlumatları (açıq açar) müxtəlif növ yaddaş vasitələrində saxlana bilər. Əsas şərt açarların saxlanması üçün seçilən həllin gizli açarın təhlükəsizliyini təmin edə bilməsi və CSP (Kriptoqrafiya Xidməti Provayderi) müxtəlif əməliyyat sistemləri üçün açarlardan və sertifikatdan istifadəni təmin edə bilməsidir. Aşağıdakı CSP-lərə əsasən açarların ölçülərinə görə generasiya vaxtı və MSOffice və P7Signer[2] –imzalayıcının imkanlarından istifadə edərək imzalanma müddətləri ölçülmüşdür.

1) Provider Name: Athena ASECard Crypto CSP
   Provider Type: 1 - PROV_RSA_FULL
2) Provider Name: Gemalto Classic Card CSP
   Provider Type: 1 - PROV_RSA_FULL
3) Provider Name: EnterSafe ET199SD CSP v1.0
   Provider Type: 1 - PROV_RSA_FULL
4) Provider Name: eToken Base Cryptographic Provider
   Provider Type: 1 - PROV_RSA_FULL
5) Provider Name: Oberthur Card Systems Cryptographic
   Provider Type: 12 - PROV_RSA_SCHANNEL
   Provider Name: Charismathics Smart Security Interface
6) *Provider Type 24 - PROV_RSA_AES*

Açarlar proqram təminatı və ya aparat təhlükəsizlik modullarının imkanları vasitəsilə qorunur. Proqram təminatı vasitəsi ilə təmin edilən açarların təhlükəsizliyik səviyyəsi coxd yüksək olmur və nəticə etibarı ilə tam təhlükəsiz sayılmır.

Gizli açarların qorunması üçün aşağıdakı vasitələr istifadə edilir:

- HSM (Hardware Security Module)
- Smart kart
- USBToken
- SmartSD kart

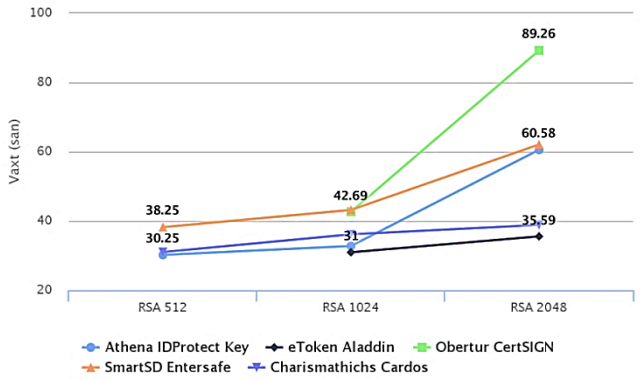
Aşağıdakı cədvəldə kriptopravayderlərə əsaslanaraq müxtəlif ölçülü açarların generasiya müddətləri təqdim edilir.

Cədvəl 1. Sertifikat sorğusunun və açarların generasiyası müddəti

№	Açarların saxlanma sistemlərində generasiya müddəti			
	Gizli açarların qorunma vasitələri	**RSA 512**	**RSA 1024**	**RSA 2048**
1	Athena İDProtect Key	30.25 san	32.81 san	60.58 san
2	eToken Aladdin	**Null**	31 san	35.59 san
3	Obertur CertSİGN	**Null**	42.69 san	89.26 san
4	SmartSD Entersafe	38.25 san	43.22 san	62.11 san
5	Charismathichs Cardos	31.06 san	36.18 san	38.90 san

Null: *–gizli açar qoruyucuları üçün parametrlər təyin edilməyib.*

Aşağıda qeyd olunan Şəkil 1-də Kriptoprovayderlərə əsasən RSA 512, 1024, 2048 bit ölçülü açarlarla sertifikat sorğusu yaranmasının qrafik təsviri öz əksini tapır.



Şəkil 1. Sertifikat sorğusunun yaranması müddəti

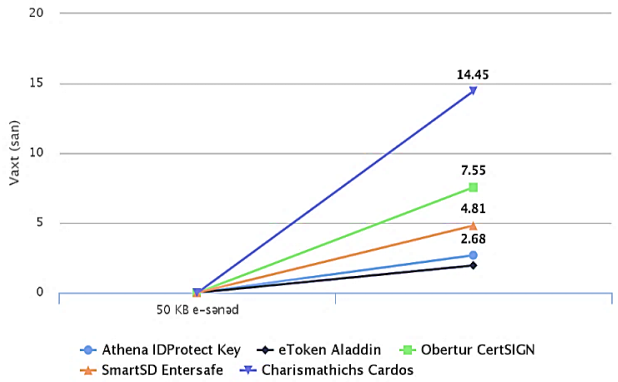
Sertifikat sorğusunun və açarların generasiyasının zamandan asıllığı təsvir edilib.

SmartSD[3] və USBToken açar daçıyıcıları həmçinin konteynerləri şifrələmə funksiyalarına malik xarakteristikalarla təmin edilir.

Növbəti ekspermentdə yuxarıda qeyd olunan CSP-lərə əsaslanaraq RSA 2048 bit SHA1 -ə əsaslanaraq MSOffice və E-imzalayıcının imkanlarından istifadə etməklə ölçmələrin nəticəsi qeyd edilir:

Cədvəl 2. MS Office –də e-sənədin imzalanması müddəti

	e-sənədin MSOffice-də imzalanması (e-sənədin ölçüsü 50 bayt)	
	Gizli açarların qorunma vasitələri	**RSA2048 SHA1**
1	Athena İDProtect Key	2.68 san
2	eToken Aladdin	1.95 san
3	Obertur CertSİGN	7.55 san
4	SmartSD Entersafe	4.81 san
5	Charismathichs Cardos	14.45 san



Şəkil 2. MS Office-də e-sənədin imzalanması müddəti

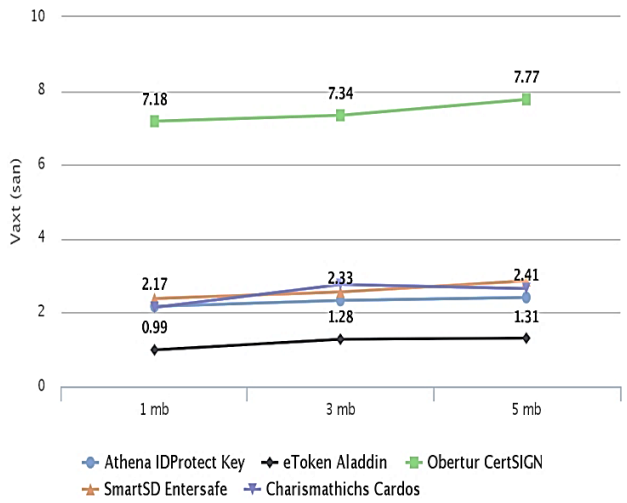
MS Office-də e-sənədin imzalanmasının zamandan asıllığı təsvir edilib.

E-sənədin digər e- imzalayıcı ilə əldə edilməsi zamanı aşağdakı qiymətləndirmələr əldə edilmişdir.

Cədvəl 3. E-imzalayıcıda müxtəlif ölçülü e-sənədlərinin imzalanması müddəti

№	**E-sənədin e-imzalayıcıda imzalanması (e-sənədin ölçüsü 1,3,5 Mb)**			
	Gizli açarların qorunma vasitələri	**1 Mb**	**3 Mb**	**5 Mb**
1	Athena İDProtect Key	2.17 san	2.631 san	2.41 san
2	eToken Aladdin	0.99san	1.28 san	1.31 san
3	Obertur CertSİGN	7.18 san	7.34 san	7.77 san
4	SmartSD Entersafe	2.38 san	2.56 san	2.86 san
5	Charismathichs Cardos	2.14 san	2.76 san	2.65 san

Ekspermentdə yuxarıda qeyd olunan CSP-lərə əsaslanaraq RSA 2048 bit SHA1 -ə əsaslanaraq və E-imzalayıcının imkanlarından istifadə etməklə ölçmələrin nəticəsini zamandan asılılığı müəyyən edilib.



Şəkil 3. PKCS7-də e-sənədin imzalanması müddəti

Müxtəlif tip CSP daşıyıcılarının PKCS7[4] imzalanma həllərinin zamana görə imzalama qrafiki təsvir edilib. Aparılan ölçmələr nəticəsində alınan qrafiklər və cədvəllərdəki qiymətləndirmər belə bir halı özündə əks etdirir ki, CSP –lər[5] özlərinin daxili qurluşu və iş pirinsipi və həmçinin sənəd imzalama sistemləri eyni olçülü məlumatları müxtəlif zaman kəsiklərində icra edirlər. Bu baxımdan e-gov həllərinə və yaxud istənilən informasiya sisteminə e-imza həllərini inteqrasiya etdikdə daşıyıcılar və avadalıqların texniki xarakteristikası işin təşkilinə hər zaman təsir imkanlarına malik ola bilər. CSP-lərə[6] nəzərən alınan nəticələrdən sürətli iş imkanlarına malik olanları müəyyən edib, yüksək sürətlə məlumatların imzalanması və şifrələnməsini təmin etmək mümkündür.

ƏDƏBİYYAT

[1] European Union: Directive 1999/93/EC of the European Parliament and of the Council of 13 December 1999 on a Community framework for electronic signatures. Official Journal of the European Communities. (1999).

[2] Eastlake, D, Reagle, J, Solo, D: XML Signature Syntax and Processing (Second Edition), W3C Recommendation, http://www.w3.org/TR/xmldsig-core

[3] https://www.sdcard.org/developers/overview/ASSD/smartsd/

[4] https://en.wikipedia.org/wiki/PKCS

[5] http://pki.escb.eu/epkweb/pdf/ESCB-PKI-Basic_operations_leaflet.pdf

[6] https://www.charismathics.com/fileadmin/files/pdf/manuals/CSSI_49_EN.pdf

Вопросы Применения Методов Машинного Обучения для Решения Проблем Информационной Безопасности

Ядигар Имамвердиев[1], Людмила Сухостат[2]
[1,2]Институт Информационных Технологий НАНА, Баку, Азербайджан
[1]*yadigar@lan.ab.az*, [2]*lsuhostat@hotmail.com*

***Аннотация* – в статье представлен обзор основных методов обнаружения аномалий на основе машинного обучения для решения проблем информационной безопасности. Рассмотрены методы классификации и кластеризации. Приведены проблемы машинного обучения во враждебной среде.**

Ключевые слова – машинное обучение, информационная безопасность, машинное обучение во враждебной среде, обнаружение аномалий.

I. ВВЕДЕНИЕ

Машинное обучение изучает алгоритмы, которые улучшают характеристики работы путем накопления опыта. Методы машинного обучения доказали свою большую практическую эффективность во многих областях применения. Они особенно полезны: (а) в плохо понимаемых предметных областях, где существует мало знаний, чтобы разработать эффективные алгоритмы; (б) в областях, где имеются большие базы данных, содержащие ценные неявные закономерности, которые могут быть обнаружены, или (в) в областях, где программы должны адаптироваться к меняющимся условиям. Неудивительно, что поле информационной безопасности оказывается благодатной почвой, где многие задачи могут быть сформулированы как проблемы обучения и изучены с точки зрения алгоритмов обучения.

По сути, машинное обучение изучает задачи применения математических методов к большому количеству данных для прогноза вероятностей, например, что электронное письмо является спамом. Эти системы работают эффективно благодаря поступлению большого количества данных, на основе которых они могут строить свои прогнозы. Более того, системы спроектированы таким образом, чтобы со временем улучшаться за счет отслеживания самых полезных сигналов и моделей по мере поступления новых данных.

Технологии машинного обучения позволяют максимально эффективно использовать накопленные в организации данные о безопасности и поэтому получают все большее распространение в сфере информационной безопасности. Но информационная безопасность имеет дополнительные трудности для успешного применения методов машинного обучения.

Данная работа посвящена анализу вопросов применения методов машинного обучения для решения проблем информационной безопасности.

II. ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ

Методы обнаружения вторжений могут быть разделены на два класса:

1) методы, основанные на обнаружении злоупотреблений (misuse detection);
2) методы, основанные на выявлении аномалий (anomaly detection, AD).

Сигнатурный подход занимает центральное место в большинстве современных антивирусов и обеспечивает на уже известных вторжениях 100-процентную точность обнаружения. Но сигнатурный анализ не работает на вирусах, не известных антивирусу, т.е. если не известны их сигнатуры. Так называемые вторжения нулевого дня (zero-day). Их обнаружить позволяет аномальный подход. Такие методы строят классификаторы или решающие правила. В последнее время наряду с этими методами [1, 2] большой интерес представляют методы, использующие машинное обучение.

Три основные задачи машинного обучения – классификация, кластеризация и регрессия – успешно применялись для обнаружения атак, идентификации и анализа угроз и вредоносных программ [3].

Методы классификации. Классификация (обучение с учителем) является классической задачей анализа данных, с корнями, уходящими в машинное обучение [4].

Процесс классификации включает в себя следующие этапы:

1) создание обучающего набора данных;
2) определение атрибута и классов;
3) определение полезных атрибутов для классификации (релевантный анализ);
4) обучение модели использованию примеров из обучающей выборки;
5) использование модели классификации неизвестных образцов данных.

Можно выделить следующие методы классификации.

а) Деревья принятия решений

Дерево решений – это древовидная структура, где каждый нетерминальный узел представляет испытание или решение о рассматриваемом элементе данных. Выбор определенной ветки зависит от исхода испытания. Решение принимается тогда, когда найден терминальный узел. Дерево решений – особая форма набора правил, характеризуемая их иерархической организацией.

б) Нейронные сети

Нейронные сети (НС) являются системами, смоделированными на основе работы мозга человека, состоящего из миллионов нейронов, соединенных между собой с помощью синапсов. Нейронная сеть – набор соединенных модулей входа/выхода, в котором каждое соединение имеет вес, связанный с ним. Обучение сети направлено на определение набора весов с целью минимизации ошибки классификации. Искусственная НС используется для классификации и предсказания латентных переменных, которые сложно измерить, и решения задачи нелинейной классификации. Результаты классификации НС трудно интерпретировать в отличие от других методов классификации, где имеется функциональная зависимость между данными (например ассоциативные правила).

в) Наивные байесовские классификаторы

Наивные байесовские классификаторы используют теорему Байеса для классификации новых экземпляров данных. Каждый экземпляр представляет собой набор значений атрибутов, описанных вектором $X=(x_1,x_2,\ldots,x_n)$. Учитывая M - классов, образец X присваивается классу C_i, тогда и только тогда, когда $P(X\mid C_i)P(C_i)>P(X\mid C_j)P(C_j)$ для всех $j=\overline{1,M}$ таких, что $i\neq j$. Образец принадлежит к классу с максимальной апостериорной вероятностью для образца. $P(X_k\mid C_i)$ рассчитывается как отношение значения частоты X_k для атрибута A_k к общему количеству образцов в обучающем наборе. В наивном байесовском подходе атрибуты, как предполагается, условно независимы. Несмотря на это предположение, наивные байесовские классификаторы дают удовлетворительные результаты, потому что акцент делается на идентификацию классов, а не точных вероятностей. Подход применяется для классификации спама и текста. Теоретически байесовские классификаторы менее склонны к ошибкам. С увеличением числа классов или атрибутов пространство и вычислительная сложность байесовских классификаторов экспоненциально возрастают.

г) Нечеткие множества

Нечеткие множества образуют ключевую методику представления и обработки неопределенности. Неопределенность возникает во многих формах в современных базах данных: неточность, неспецифичность, непоследовательность, расплывчатость и т.д. Нечеткие множества используют неуверенность в попытке сделать сложную систему управляемой. Таким образом, нечеткие множества представляют собой мощный подход не только к неполным, зашумленным или неточным данным, но также могут быть полезны при разработке неопределенных моделей данных, которые обеспечивают более высокую и точную производительность, чем традиционные системы.

Методы кластеризации

Кластеризация (обучение без учителя) может рассматриваться как обобщение классификации. Техника кластеризации может быть использована для обнаружения вторжений и неизвестных атак [4].

Известно много методов кластеризации. Выбор конкретного метода зависит от типа желаемого выхода. В общем, их можно разделить на две категории, основанные на структуре кластера, которые они образуют: иерархические и неиерархические.

1) Неиерархические методы

Неиерархические методы делят набор данных из N - объектов на M - кластеров, с перекрытием или без него. Каждый объект является членом кластера, к которому он наиболее близок. Однако порог сходства должен быть определен.

2) Иерархические методы

Иерархические методы образуют набор вложенных кластеров, каждая пара которых вкладывается в больший кластер. Процесс продолжается до тех пор, пока не останется только один кластер. Иерархические методы могут быть агломеративные и дивизимные.

В агломеративных методах [5] каждый объект соотносится с отдельным кластером. Кластеры объединяются до получения единственного кластера, включающего в себя все объекты.

В дивизимных методах [6] все рассматриваемые объекты определяются в один кластер на начальном этапе. На основании выбранной меры сходства выполняется разделение кластера до тех пор, пока каждый объект не будет определен в отдельный кластер. Примером является метод COBWEB.

а) Метод ближайшего соседа

Метод ближайшего соседа является одним из известных методов, относящихся к иерархическим агломеративным методам [7]. В данном методе информация о расстоянии между ближними, дальними и центральными объектами кластеров объединяется.

б) Минимальное покрывающее дерево

Данный метод выполняет кластеризацию объектов «сверху вниз» [8], в котором вначале все объекты определяются в единственный кластер. Кластеры разбиваются на два до тех пор, пока каждый объект не окажется в отдельном кластере. При этом происходит оценка расстояния между ними, которое должно быть максимальным.

в) Метод COBWEB

В этом методе реализуется вероятностное представление кластеров [5]. Объект, принадлежащий

кластеру, определяется с помощью вероятности появления конкретных значений параметра у объекта.

г) Метод k-means

Цель метода заключается в минимизации расстояний между объектами и центром кластера, к которому они относятся. Для метода k-means функцию минимизации расстояний можно представить [9]:

$$J = \sum_{k=1}^{M}\sum_{i=1}^{N} d^2(x_i, c_k),$$

где $x_i \in X$ – объект, а $c_j \in C$ – центр кластера, $|X| = N$, $|C| = M$.

д) Метод Fuzzy C-Means

Данный метод относится к нечетким методам, в которых делается предположение, что каждый элемент с определенной степенью принадлежит каждому кластеру [10]. Минимизация происходит на основе функции [10]:

$$J = \sum_{k=1}^{M}\sum_{i=1}^{N} u_{ij}^{m} d(x_i, c_k),$$

где $x_i \in X$ – объект, а $c_j \in C$ – центр кластера, m – мера нечеткости, $|X| = N$, $|C| = M$.

е) Метод EM

Данный метод основан на широко известном подходе максимизации ожиданий (Expectation Maximization), который предполагает, что кластеры формируются из результатов наблюдений, имеющих нормальное распределение [9]. Алгоритм состоит из двух основных шагов: E-шаг ожидания и M-шаг максимизации.

Ассоциативные правила

Типичным примером обучения без учителя являются ассоциативные правила. Это выражения вида $[X \to Y, c, s]$, где X, Y – подмножества множества значений на множестве атрибутов, $s = Support(X \bigcup Y)$, $c = \dfrac{Support(X \bigcup Y)}{Support(X)}$ – процент записей в базе, которые содержат $X \bigcup Y$ [3]. На этапе обучения добавляются новые правила при обнаружении аномалий. Обучение прекращается, пока по исходным данным уже нельзя будет получить новые данные.

Методы машинного обучения (Machine learning, ML) традиционно нашли применение в области информационной безопасности для обнаружения аномалий в работе информационной системы. Основная идея проста и применяется для различных видов информационных систем, таких как сети, операционные системы, прикладное программное обеспечение, и так далее [1]. Работа механизма обнаружения аномалии (AD) с использованием методов ML как правило, делится на две фазы: фаза обучения и фаза обнаружения. В фазе обучения механизм AD узнает нормальное поведение информационной системы. Результатом этапа обучения является множество профилей, характеризующих поведение системы. Профили используются для логического разделения поведения системы таким образом, чтобы поддерживать выделение аномалий на этапе обнаружения. Например, поведение операционной системы может быть смоделировано в виде набора профилей, где каждый профиль связан с действиями одного пользователя. Такая модель может быть использована для обнаружения аномальных действий пользователя. Кроме того, если количество пользователей больше и большинство пользователей демонстрируют аналогичное поведение, один профиль может быть связан с группой пользователей (например как роли), которые обладают подобными поведениями.

Информация в созданном профиле зависит от наблюдаемых системных событий, типа информации, собранной из этих событий, и от используемого алгоритма обучения. После того, как профили созданы, механизм AD работает в режиме обнаружения. В режиме обнаружения новое наблюдаемое событие сравнивается с созданными профилями для соответствия. Различные критерии, такие как статистические тесты отклонения, тесты обнаружения выбросов, апостериорные вероятности и т.д., могут быть использованы для определения соответствия события к созданным профилям [1]. Если событие не соответствует профилям (на основе некоторого порога), то оно классифицируется как аномальное. В противном случае информация от событий либо объединяется в профили, либо отбрасывается. Система AD имеет преимущество над системой на основе сигнатур в том, что она может обнаружить потенциально новые атаки, в то время как система на основе сигнатур может работать только для тех атак, для которых были созданы подписи нападений. Недостатком системы AD является число ложных тревог, которые она выдает, когда профили, созданные на этапе обучения, не полностью соответствуют нормальному поведению.

Выделяют ошибки следующих видов: а) нормальное поведение системы или пользователя ошибочно принимается за злоумышленное (false positives); б) попытка злоумышленного проникновения в систему принимается за нормальную активность (false negatives). Более опасной является вторая ситуация [11].

III. ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ К ПРОБЛЕМАМ ИБ

Одна из проблем связана с парадигмой обнаружения аномалий в сети. Некоторые авторы [12, 13] ставят под сомнение применение модели Деннинга для обнаружения аномалий в одной системе, предложенной в 1987 году [14] в неизменном виде к обнаружению аномалий в сети, и предлагают заменить его новой парадигмой в свете изменений в операционной среде.

К системам IDS (Intrusion detection system) можно предъявить различные требования. Важным требованием

является то, что система должна обнаружить значительный процент вторжений, при этом сохраняя частоту ложных тревог на приемлемом уровне. Работа демонстрирует, что для разумного набора предположений частота ложных тревог является ограничивающим фактором для производительности системы IDS. Это связано с явлением base-rate fallacy, которое применительно к системам IDS означает, что для достижения существенных значений уровня обнаружения P(Intrusion|Alarm), требуется обеспечить низкий уровень ложных тревог (возможно, в некоторых случаях недостижимо). Анализ отчетов производительности IDS показывает, что, по крайней мере для некоторых видов обнаружения вторжений, достичь таких низких частот ложных тревог является проблематичным [13].

Одной из больших проблем является то, что злонамеренная активность находится среди терабайтов нормальных данных как иголка в стоге сена, в результате чего методы из учебников по машинному обучению, часто не предназначенных для таких сценариев, производят массу ложных срабатываний. Еще одной проблемой является неизбежное отсутствие доступа к данным атак нулевого дня для тренировки методов машинного обучения.

Хотя применение машинного обучения для ИБ имеет аналоги использования в других областях, оно также имеет существенные различия. Главным различием применения машинного обучения в области ИБ является то, что в других областях никто не пытается обмануть систему. В случае применения машинного обучения для обнаружения атак или вредоносных программ противник постоянно развивает свои технологии обхода системы безопасности. Согласно Saxe [15], точность систем, которые вначале способны обнаружить более 80% атак, в течение двух лет может снизиться до 60% или меньше из-за эволюции вредоносных программ или тактики атак.

Проблемы машинного обучения во враждебной среде

Исследование методов машинного обучения во враждебной среде (adversarial machine learning) находится на пересечении машинного обучения и ИБ [16, 17]. Оно направлено на безопасное применение методов машинного обучения во враждебной среде, таких, как фильтрация спама, обнаружение вредоносных программ и биометрическое распознавание.

Проблема возникает из-за того, что методы машинного обучения были изначально разработаны для стационарных сред, в которых предполагается, что данные для обучения и тестирования генерируются из одного и того же распределения. В присутствии интеллектуальных и адаптивных противников, однако, эта гипотеза, вероятно, будет нарушена, по крайней мере в некоторой степени (в зависимости от противника). На самом деле злонамеренный противник может тщательно манипулировать входными данными, эксплуатируя конкретные уязвимости алгоритмов обучения для компрометации всей системы безопасности.

Чтобы понять свойства безопасности алгоритмов обучения во враждебной среде, следует рассмотреть следующие основные вопросы [18–20]:

- идентифицировать потенциальные уязвимости алгоритмов машинного обучения в процессе обучения и классификации;
- разработать соответствующие атаки, которые соответствуют идентифицированным угрозам, и оценить их влияние на целевую систему;
- предложить контрмеры для улучшения безопасности алгоритмов машинного обучения против рассматриваемых атак.

В литературе встречается описание следующих атак на методы машинного обучения.

Атаки уклонения

Атаки уклонения [21] являются наиболее распространенными типами атак на алгоритмы машинного обучения во враждебной среде. Ярким примером атак уклонения является спам на основе изображения, в котором содержание спама встраивается в прилагаемое изображение для уклонения от текстового анализа, выполняемого антиспамфильтром. Другим примером атак уклонения является спуфинг против биометрических систем распознавания [22, 23].

Атаки отравления

Алгоритмы машинного обучения часто повторно обучаются на данных, собранных во время работы. Например, системы обнаружения вторжений (IDS) часто переобучаются на множестве образцов, собранных во время работы в сети. В этом случае злоумышленник может отравить обучающие данные путем введения тщательно разработанных образцов для компрометации процесса обучения. Примеры атак отравления алгоритмов машинного обучения могут быть найдены в [18, 19, 23, 24].

Атаки на алгоритмы кластеризации

Алгоритмы кластеризации применяются в системах безопасности для обнаружения опасных или незаконных действий. Например, кластеризация вредоносных программ и компьютерных вирусов направлена на выявление и классификацию различных семей существующих вредоносных программ, и генерацию конкретных сигнатур для их обнаружения антивирусами или на основе сигнатур систем обнаружения вторжений как Snort. Алгоритмы кластеризации не были изначально разработаны для борьбы с попытками преднамеренных атак, которые предназначены для подрыва самого процесса кластеризации. Можно ли безопасно применять кластеризацию в таких условиях, остается пока невыясненным. Предварительные работы, которые сообщают о некоторых уязвимостях кластеризации, можно найти в [25, 26].

Отметим, что встречаются также примеры атак на методы распознавания, не применяемые непосредственно в системах информационной безопасности. Например, в [27] доказывается, что для каждой обученной нейронной

сети можно найти такие примеры входных данных, которые будут давать на выходе отрицательный ответ, в то время как эти примеры данных ничтожно отличаются от примеров, дающих положительный ответ.

В области машинного обучения во враждебной среде были предложены ряд защитных механизмов против атак уклонения, атак отравления и атак на конфиденциальности, в том числе:

- определение безопасных алгоритмов обучения [28];
- использование нескольких систем классификаторов [29, 30];
- использование рандомизации или дезинформации для заблуждения злоумышленника при приобретении знаний о системе [30];
- разработка методов обучения, сохраняющих конфиденциальность [31].

ЗАКЛЮЧЕНИЕ

Машинное обучение, визуализация данных и технологии масштабируемого хранения данных создают платформу для улучшения состояния ИБ. Машинное обучение позволяет обнаружить полезные корреляции и разработать эффективные правила обнаружения атак. Машинное обучение также поможет системе обнаружить не замеченные ранее признаки атак. Ранние прототипы показали отличные результаты и доказали, что системы машинного обучения прекрасно справляются с уточнением правил обнаружения атак.

ЛИТЕРАТУРА

[1] M.I.Jordan, T.M.Mitchell, Machine learning: Trends, perspectives, and prospects, Science, Vol. 349, No. 6245, pp. 255–260, 2015. DOI:10.1126/science.aaa8415

[2] N.Idika, A.P.Mathur, A Survey of Malware Detection Techniques, Technical report, Software Engineering Research Center, 2007.

[3] D.Barbara, S.Jajodia, Applications of Data Mining in Computer Security, Kluwer, 2002.

[4] S.Dua, X.Du, “Data Mining and Machine Learning in Cybersecurity, Auerbach Publications, 2011.

[5] D.H.Fisher, Knowledge acquisition via incremental conceptual clustering, Machine Learning, Vol. 2, No. 2, pp. 139–172, 1987.

[6] Д.А.Вятченин, “Нечеткие методы автоматической классификации: Монография,” Мн.: УП «Технопринт», 2004.

[7] M.A.Maloof, "Machine Learning and Data Mining for Computer Security: Methods and Applications, Springer-Verlag: New York, 2005.

[8] Н.Г.Загоруйко, “Прикладные методы анализа данных и знаний”, Новосибирск: ИМ СО РАН, 1999.

[9] A.K.Jain, M.N.Murty, and P.J.Flynn, Data Clustering, ACM Computing Surveys, Vol. 31, pp. 264–323, 1999.

[10] В.П.Боровиков, STATISTICA. Искусство анализа данных на компьютере: Для профессионалов”, СПб.: Питер, 2003.

[11] S.Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In ACM Conference on Computer and Communications Security, pp.1–7, 1999.

[12] C.Gates, C.Taylor, Challenging the Anomaly Detection Paradigm: A Provocative Discussion, Proceedings of the Workshop on New Security Paradigms, pp. 21–29, 2007.

[13] S.Axelsson, The base-rate fallacy and the difficulty of intrusion detection, ACM Transactions on Information and System Security, Vol. 3, No. 3, pp.186–205, Aug. 2000 [doi>10.1145/357830.357849]

[14] D.E.Denning, An Intrusion-Detection Model, IEEE Transactions on Software Engineering - Special issue on computer security and privacy, Vol. 13, Issue 2, February 1987, pp. 222–232.

[15] J.Saxe, Why Security Data Science Matters and How Its Different: Pitfalls and Promises of Data Science Based Breach Detection and Threat Intelligence. BlackHat USA 2015 Briefings. https://www.blackhat.com/us-15/speakers/Joshua-Saxe.html

[16] P.Laskov, R.Lippmann, Machine Learning in Adversarial Environments, Machine Learning Journal, 81, 2010.

[17] A.D.Joseph, P.Laskov, F.Roli, J. Doug Tygar, and B. Nelson (eds) Machine Learning Methods for Computer Security Report from Dagstuhl Perpectives Workshop 12371.

[18] B. Biggio, G. Fumera, and F. Roli. "Security evaluation of pattern classifiers under attack". IEEE Transactions on Knowledge and Data Engineering, 26(4):984–996, 2014.

[19] B.Biggio, I.Corona, B.Nelson, B. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli. Security evaluation of support vector machines in adversarial environments. In Y.Ma and G.Guo (eds), Support Vector Machines Applications, pp. 105–153. Springer, 2014.

[20] B.Biggio, G.Fumera, and F.Roli, “Pattern recognition systems under attack: Design issues and research challenges, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 28, No. 7, 1460002, 2014. DOI: 10.1142/S0218001414600027.

[21] B. Biggio, I.Corona, D.Maiorca, B.Nelson, N.Srndic, P.Laskov, G.Giacinto, and F.Roli. Evasion attacks against machine learning at test time. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Part III, LNCS Vol. 8190, pp. 387– 402. Springer Berlin Heidelberg, 2013.

[22] R.N.Rodrigues, L.L.Ling, and V.Govindaraju. Robustness of multimodal biometric fusion methods against spoof attacks, Journal of Visual Languages and Computing, Vol. 20, No. 3, pp. 169–179, 2009.

[23] B.Biggio, L.Didaci, G.Fumera, and F.Roli, Poisoning attacks to compromise face templates, In 6th IAPR Int’l Conf. on Biometrics (ICB 2013), pp. 1–7, 2013.

[24] B.Biggio, B.Nelson, and P.Laskov, Poisoning attacks against support vector machines. Proc. of the 29th Int’l Conf. on Machine Learning (ICML), 2012. arxiv.org/vc/arxiv/papers/1206/1206.6389v1.pdf

[25] B.Biggio, I.Pillai, S. R. Bulò, D.Ariu, M.Pelillo, and F.Roli. Is data clustering in adversarial settings secure? Proc. of the ACM Workshop on Artificial Intelligence and Security (AISec’13), pp. 87-98, 2013.

[26] D.B.Skillicorn. Adversarial knowledge discovery, IEEE Intelligent Systems, Vol. 24, pp. 54–61, 2009.

[27] C.Szegedy, W.Zaremba, I.Sutskever, J.Bruna, D.Erhan, I.Goodfellow, R.Fergus, Intriguing properties of neural networks, Proc. of the International Conference on Learning Representations (ICLR), 2014. http://arxiv.org/abs/1312.6199

[28] O.Dekel, O.Shamir, L.Xiao, Learning to classify with missing and corrupted features, Machine Learning, Vol.81, No.2, pp. 149–178, 2010.

[29] B.Biggio, G.Fumera, and F.Roli. Multiple classifier systems for robust classifier design in adversarial environments. International Journal of Machine Learning and Cybernetics, Vol. 1, No. 1, pp. 27–41, 2010.

[30] B.Biggio, G.Fumera, and F.Roli, Adversarial pattern classification using multiple classifiers and randomisation. In 12th Joint IAPR International Workshop on Structural and Syntactic Pattern Recognition (SSPR 2008), LNCS Vol. 5342, pp. 500–509, 2008.

[31] B.I.P. Rubinstein, P.L.Bartlett, L. Huang, and N.Taft. Learning in a large function space: Privacy- preserving mechanisms for svm learning. Journal of Privacy and Confidentiality, Vol. 4, No. 1, pp. 65–100, 2012.

Botnetlər Aşkarlanması Üsullarının İcmalı

Yadigar İmamverdiyev[1], Gülnarə Qarayeva[2]
[1]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[2]Azərbaycan Müəllimlər İnstitutunun Şəki filialı, Şəki, Azərbaycan
[1]*yadigar@lan.ab.az*, [2]*qarayevagulnare@mail.ru*

***Xülasə* — Botnetlər kiber-hücum infrastrukturunda mühüm yer tuturlar. Botnet yoluxmuş kompüterlərdən və onları idarə edən botmasterlərin istifadə etdikləri C&C serverlərdən ibarət şəbəkədir, bəzən bu şəbəkəyə milyonlarla kompüter cəlb edilir. Botnetlər daim inkişaf edir, strukturları, istifadə etdikləri protokollar, yoluxdurma üsulları, hücum məqsədləri daim dəyişir. Məqalədə botnetlərin arxitekturası, təsnifatı və aşkarlama üsulları araşdırılmışdır.**

Açar sözlər* — *botnet, C&C server, honeypot, DDoS hücum, şəbəkə əsaslı aşkarlama üsulları

I. GİRİŞ

Botnet – “robot” və “network” sözlərinin birləşməsindən yaranmışdır. Bədniyyətliyə istifadəçinin xəbəri olmadan yoluxmuş kompüteri məsafədən idarə etməyə imkan verən ziyankar proqramlarla (botlarla) yoluxmuş kompüterlərdən ibarət şəbəkədir. Bot istifadəçinin kompüterində gizli quraşdırılan və bədniyyətliyə yoluxmuş kompüterin resurslarından istifadə etməklə müəyyən əməlləri yerinə yetirməyə imkan verən proqramdır. Belə şəbəkələr çox vaxt verilmiş əmrləri avtomatik yerinə yetirən zombi-kompüter şəbəkələri də adlanır.

Botnetlər 2003-cü ildən başlayaraq İnternetdə bir çox təhlükəsizlik problemlərinin meydana çıxmasında əsas rol oynayır. Belə şəbəkələr vasitəsilə həyata keçirilən müxtəlif məqsədli hücumlar müasir ümumdünya şəbəkəsinin əsas və demək olar ki, ən irihəcmli təhlükəsi hesab olunur.

Hücumların həcmi oğurlanmış elektron (rəqəmsal) kimliklərin və yoluxmuş kompüterlərin sayı ilə xarakterizə olunur. Botnetlər bir çox hücum məqsədləri üçün istifadə olunur. Əsasən spamların göndərilməsi (spamming), fişinq, fərdi və konfidensial məlumatların oğurlanması, DDoS (Distributed Denial of Service, paylanmış xidmətdən imtina) hücumları, zərərli proqramların yüklənməsi və yayılması, klikləmə saxtakarlığı (ing. click fraud) və s. məqsədlər üçün istifadə olunur.

Bu gün müxtəlif ölkələrdə milyonlarla yoluxmuş kompüter vardır. Bu şəbəkələr artıq kifayət qədər təhlükəli vəziyyətdədir. Botnetlərin aktiv olaraq yaradıldığı və istifadə olunduğu ölkələr əsasən ABŞ, Cənubi Amerika ölkələri, Avropa ölkələri və bir neçə Asiya ölkəsidir. Botnetləri idarəetmə mərkəzlərinin yerləşdiyi ölkələr ABŞ, Çin, Böyük Britaniya, Rusiya və s.dir (Kasperski şirkətinin 2015-ci ilin I rübü üçün hesbatı).

Kibercinayətkarlar tərəfindən idarə olunan komanda-nəzarət (Command and Control, C&C) serverləri müxtəlif ölkələrdə yerləşir. Serverlərin fiziki olaraq yerləşdiyi ərazi çox vaxt kibercinayətkarların olduğu yer deyildir. Serverlərin yerləşdiyi ölkələr isə ABŞ, Çin, Böyük Britaniya, Rusiya, Cənubi Koreya və s. ölkələrdir.

Botnetlərin yaradılması üsulları, onların istifadəsi, yayılması internet vasitəsilə həyata keçirilir. Kibercinayətkarlığa qarşı görülən tədbirlərə baxmayaraq, botnet “bazar”ları aşkar fəaliyyət göstərməkdədir. Aparılan araşdırmalar və statistik nəticələr onu göstərir ki, botnetlərin istifadə olunduğu əsas məqsəd DDoS hücumlarıdır. Saytların çökdürülməsi nəticəsində vurulan maddi zərər isə milyonlarla ölçülür.

İstər DDoS hücumların sayı, istərsə də botnetlərin səbəb olduğu digər problemlər onların nə qədər təhlükəli olduğunu sübut edir. Botnetlərin qarşısının alınması, zamanında aşkarlanması və onlara qarşı ehtiyat tədbirlərinin həyata keçirilməsi vacibdir.

Botnetlərlə mübarizə iki mərhələdə aparıla bilər. Birincisi, botla yoluxmadan qorunmaq (bu birbaşa istifadəçilərin üzərinə düşür), ikincisi isə, yoluxmuş qurğuların aşkarlanması və botnetin fəaliyyətinin dayandırılması. Botnetlərin istifadə etdiyi yoluxma üsullarının müxtəlifliyi, onların arxitekturasının kifayət qədər öyrənilməməsi, istifadəçilərin qorunma yollarından məlumatsızlığı botnetlərin yaranmasının əsas səbəblərindəndir.

Hazırda botnetlər və onların istifadə olunduğu sahələr sürətlə artmaqdadır. Botnetlərin qurulma və idarəetmə üsulları, həmçinin hücumların növləri artdıqca onlarla yeni mübarizə üsullarının işlənməsi də aktual olaraq qalır.

II. BOTNETLƏRİN QISA XARAKTERİSTİKASI

Botnet zərərli proqramlarla yoluxmuş və bədniyyətli tərəfindən idarə edilə bilən kompüterlər şəbəkəsidir. Belə şəbəkədəki hər bir botu və bütövlükdə botneti idarə edən bədniyyətlilər “botmaster” adlanır. Botmasterlər botnetləri müxtəlif məqsədlər üçün istifadə edirlər. Botmasterlər işə bot toplamaqla, yəni yeni kompüterləri ələ keçirməklə başlayırlar. Bunun üçün müxtəlif üsullardan istifadə olunur. Məsələn; viruslar və ya digər zərərli proqramlar, e-poçt, spamlar, müxtəlif məqsədli cəlbedici şəkillər, reklamlar və s. vasitəsilə. Botmasterlər tərəfindən yayılan belə vasitələrin istifadəçi tərəfindən yüklənməsi nəticəsində həmin proqram kompüterdə qurulmuş olur. Kompüter botla bir dəfə yoluxduqda o, botnet idarəetmə mərkəzinə bağlanır və əmr gözləyir. Yoluxmuş kompüterin açıq vəziyyətdə olması onun botmaster tərəfindən

idarə edilməsi üçün kifayətdir. Belə istifadə istifadəçinin icazəsi olmadan yerinə yetirilir.

Botnetlər müxtəlif məqsədlər üçün istifadə olunurlar. Ən çox istifadə olunduqları məqsədlər aşağıdakılardır:

1) Spam göndərilməsi (Spamming). Bildiyimiz kimi internet trafikinin 90%-dən çoxunu spamlar təşkil edir. Belə spamların 95%-dən çoxu botnetlər tərəfindən göndərilir. Spamların botnetlər tərəfindən yayılmasının botmasterlər üçün əsas üstünlükləri spamları qəbul edən tərəflərin leqal yolla spamı göndərən tərəfə dönə bilməməsi və daha böyük ölçüdə spamların yayıla bilməsidir;

2) DDoS (Distributed Denial of Service, Paylanmış Xidmətdən İmtina) hücumlar. Botmasterlər botnetlərdən istifadə edərək böyük sayda istifadəçisi olan veb saytlara eyni anda müraciəti təşkil edirlər. Belə olduqda veb serverdə gecikmə və ya tamamilə xidmətdən imtina vəziyyəti yaranır. Ən çox istifadə olunan DDoS hücumlarının bir növü də Syn hücumlarıdır (Syn Flood);

3) Klikləmə saxtakarlığı (Click fraud). Veb saytların, şəxsi bloqların, sosial şəbəkə səhifələrinin istifadəçi sayını süni olaraq artırmaq məqsədilə, həmçinin elektron səsvermə sistemlərində klik sayı artırmaq (və ya azaltmaq) üçün botnetlərdən geniş istifadə olunur;

4) İnformasiya oğurluğu. Belə informasiyaya plastik kartlar haqqında məlumatlar (son istifadə tarixi, CVV2 və s.), ünvan, telefon nömrəsi, istifadəçi adı/parol, SSN (Social Security Number), ATM PIN və s. məlumatlar daxildir;

5) Fişinq hücumları üçün infrastruktur yaratmaq;

6) Digər zərərli proqramları yaymaq, məsələn, casus proqramların yayılması, reklam materiallarının yayılması və s.

III. BOTNETLƏRİN TƏSNİFATI

Botnetlərin təsnifatı onların arxitekturasına və botları idarə etmək üçün istifadə edilən protokollara əsaslanır. Botnetləri xarakterizə edən əsas kateqoriyalardan biri də C&C (Command & Control, Əmr və Nəzarət) serverlərdir. Botmasterlər botlarla əlaqə saxlamaq üçün belə xüsusi serverlərdən istifadə edirlər. C&C server olaraq çox vaxt leqal olaraq yerləşdiyi coğrafi ərazi məlum olmayan gizli serverlərdən istifadə olunur. Botmaster botneti idarə etmək üçün C&C server qurur və yoluxmuş kompüter hər hansı bir kanalla (məs; IRC) bu serverə bağlanır və ondan əmr gözləyir. Botnetləri arxitekturasına görə bir neçə formada təsnif etmək olar. [5]-də botnetlər 3 arxitekturada – mərkəzləşmiş, P2P (peer – to – peer) və hibrid (ağacvari) olaraq təsnif edilmişdir. Lakin, bəzi mənbələrdə, məs; [8]-də C&C sistemlər aşağıdakı kimi təsnif edilmişdir:

- *1) Mərkəzləşmiş C&C modeli (Centralized C&C Model).*

Mərkəzləşmiş C&C modeli botnetlərdə istifadə olunan əsas modellərdən biridir. Ən çox tanınan botnetlərdən AgoBot, SDBot, Rbot və s. mərkəzləşmiş modelə aiddir. Belə modeldə botmaster botlarla əlaqə saxlamaq üçün bir yüksəksürətli kanal seçir. Adətən, belə modeldə C&C server IRC (Internet Relay Chat), HTTP kimi şəbəkə protokollarını istifadə edərək kompüteri razı salır. Yeni kompüter botla yoluxduqda o, C&C serverlə əlaqə yaradaraq botnetə qoşulur. Müvafiq C&C serverə bir dəfə qoşulduqdan sonra botmasterin əmrlərini gözləyir.

Mərkəzləşmiş C&C modelin istifadə edilməsinin aşağıdakı kimi üstün cəhətləri var:

- daha asan əldə olunan proqramlar (məs; IRC skriptlər və IRC botlar) səbəbindən mərkəzləşmiş C&C modelində yoluxdurma və xüsusiləşdirmə daha sadədir. Belə modeli istifadə edərək botmaster eyni vaxtda minlərlə botu idarə edə bilər. Mənfəət məqsədli botnetlər yaradılarkən daha çox botu idarə etməyə imkan verən və maksimum qazanc gətirən mərkəzləşmiş C&C modeli seçilir;
- mərkəzləşmiş modeldə mesajların gecikmə vaxtı çox kiçikdir. Bu da botmasterə botları rahat idarə etməyə və hücumlar təşkil etməyə imkan verir.

Lakin, bu modelin də zəif cəhətləri vardır. Belə ki, mərkəzləşmiş modeldə bütün mübadilələr server üzərindən aparıldığından, C&C server zəif halqa təşkil edir. Əgər C&C server aşkarlanarsa, bütün botnet çökmüş olur.

2) P2P əsaslı (mərkəzi olmayan) C&C modeli (P2P-Based C&C Model).

Şəbəkədə aşkarlanmalara daha davamlı hesab olunan botnet modellərindən biri də P2P (nöqtə-nöqtə birləşmiş) əsaslı C&C modelidir. P2P əsaslı botnetlər olduqca az olmasına baxmayaraq, mərkəzləşmiş modellə müqayisədə aşkarlanması və dağıdılması daha çətindir. Mərkəzi olmayan botnetlərdə botlar idarəetmə mərkəzinə deyil, zombi-şəbəkənin bir neçə yoluxmuş maşınına qoşulur. Əmrlər botdan bota göndərilir, hər botda bir neçə “qonşusunun” ünvanı olan siyahı olur və onların hər hansı birindən əmr alınan zaman o, əmri digər qonşularına ötürür, bununla da əmri yayır. Bu halda, botneti idarə etmək üçün botmasterin botnetə daxil olan ən azı bir kompüterə birbaşa çıxışı olmalıdır.

Bu modellə qurulmuş botnetlərin getdikcə artmasına və daha çətin aşkarlanmasına baxmayaraq mənfi cəhətləri də vardır. Birincisi, P2P sistemlər çox kiçik saylı istifadəçi qrupları arasında əlaqəyə imkan verir (10 – 50 arası). Bu ölçü mərkəzləşmiş modellə müqayisədə olduqca kiçikdir. İkincisi, P2P sistemlər mesajların vaxtında çatmasına və yayılmanın gecikməməsinə zamanət verə bilmir. Aktiv olmayan bir və ya bir neçə bot şəbəkədə yayılmanı kifayət qədər zəiflədə bilər. Ona görə də belə botnetləri idarə etmək çətindir. Bu iki səbəb P2P modelin daha geniş yayılmasını və istifadəsini məhdudlaşdırır.

3) Təsadüfi C&C modeli (Random C&C Model).

Təsadüfi C&C modellər real botnetlər üçün çox da istifadə olunmur. Daha çox gələcəyə istiqamətlənmiş dözümlü botnetlər üçün düşünülmüşdür. Belə modeldə bot digər botlarla və ya botmasterlə birbaşa əlaqə saxlamır. Adətən, bot öz botmasterindən gələcək əlaqəni gözləyir. Botmaster lazım olduqda öz aktiv botlarını tapmaq üçün internetdə axtarış edir.

Bot tapıldıqda botmaster əmr verir. Belə modeldə hücumu həyata keçirmək asandır, botneti aşkarlama və dağıtma cəhdləri isə olduqca çətindir. Həmçinin böyük ölçülü hücumlarda istifadə etmək mümkün deyil.

Botnetlər üçün xarakteristik kateqoriyalardan biri də C&C serverlərin işləmə mexanizmidir. Bu mexanizmlər yeni botlar kəşf etmək və onları botmasterdən asılı salmaq üçün vacibdir. Ən çox istifadə olunan mexanizmlər aşağıdakılardır:

Sabit (statik) IP ünvanlar (Hard-coded IP address). Bu üsulda bot ilk yoluxduğu anda sabit IP ünvanlı C&C serverlə əlaqə saxlayır. Bu üsulun mənfi cəhəti odur ki, sabit IP ünvan istifadə edən C&C serverləri asanlıqla aşkarlamaq və əlaqəni bloklamaq olar. Bu vaxt əgər serverlə botlar arasında əlaqə kəsilirsə, botnet tamamilə məhv edilmiş olur. Ona görə də bu üsul müasir botnetlərdə demək olar ki, istifadə olunmur.

Dinamik DNS (DNS flux). Botnetlər DGA (Domain Generation Algorithm, Domen Yaratma Alqoritmi) istifadə edərək kriptoqrafik olaraq yaradılmış domen adlar istifadə edirlər. Bu texnologiya statik sistemlər üçün bütün mümkün C&C ünvanları tapmağı olduqca çətinləşdirir. Əgər C&C server sahibi tərəfindən bağlanarsa, idarəetmə asanlıqla yeni serverə ötürülür. Köhnə serverlə əlaqəni itirən bot DNS sorğu göndərir və yeni C&C serverə qoşulur. Dinamik DNS adlar istifadə etməklə botmaster C&C server funksiyasını itirdikdə onu bərpa edə bilir. Həmçinin çox vaxt aşkarlamanı çətinləşdirmək üçün C&C serverlər tez-tez bilərəkdən dəyişdirilir.

Dinamik IP (IP flux). Bu texnologiya FFSN (Fast Flux Service Networks, Xidmət şəbəkələrinin sürətli dəyişməsi) kimi də tanınır. Fast flux botnetlərin şəbəkədəki zərərli fəaliyyət göstərən serverlərini gizlətmək üçün istifadə olunan bir DNS texnologiyasıdır. Bu texnologiyanın əsasını bir domen adı üçün birdən çox IP ünvan saxlamaq və bu ünvanların DNS yaddaşlarının daimi olaraq dəyişdirilməsi təşkil edir. İki cür fast flux texnologiyası məlumdur: tək və double (ikiqat) flux.

Bot kompüterə botnet sahibinin əmrlərini ötürmək üçün bot və əmr göndərən kompüter arasında şəbəkə bağlantısı yaratmaq lazımdır. Bütün şəbəkə qarşılıqlı əlaqəsi şəbəkə daxilində kompüterlərin ünsiyyət qaydalarını müəyyən edən şəbəkə protokollarına əsaslanır. Buna görə də, botnetləri istifadə olunan ünsiyyət protokolu əsasında təsnif etmək olar. İstifadə edilən şəbəkə protokollarının növünə görə botnetlər IRC, IM (Instant Messaging), veb və digər (TCP, ICMP, UDP və s.) yönümlü qruplara bölünür.

Cədvəl 1-də köhnə və yeni botnetlərin istifadə etdiyi arxitektura və protokollara görə müqayisəsi verilmişdir:

CƏDVƏL 1 BOTNETLƏRİN MÜQAYİSƏSİ

Əlaqə	Köhnə	Yeni
Topologiya	mərkəzləşmiş	paylanmış və ya hibrid, bəziləri hələ də mərkəzləşmiş
Protokollar	IRC və ya HTTP	P2P
Qurulması	asan	çətin
Aşkarlanması	asan	çətin
Əlaqə	kiçik gecikməli	kiçik – orta gecikməli
Dözümlülük	pis	yaxşı
Anonimlik	pis	yaxşı

IV. BOTNETLƏRİN AŞKARLANMASI ÜSULLARI

Yoluxmuş kompüterləri bir çox əlamətlərinə görə digər kompüterlərdən fərqləndirmək olar. Məsələn; istifadə olunan IRC trafik, C&C serverlərə qoşulma cəhdləri, şəbəkədəki bir neçə maşının eyni DNS sorğu göndərməsi, yüksək SMTP (Simple Mail Transfer Protocol) trafiki (spam göndərilməsinin nəticəsi kimi), yavaş hesablama/yüksək CPU istifadəsi, istifadəçi tərəfindən göndərilməyən e-poçt, sosial media, ani mesajlar, internetə giriş problemləri və s. Bir çox şəbəkə müdaxilələrinin aşkarlanması sistemləri (Network Intrusion Detection System, NIDS) bu və ya digər xüsusiyyətlərə əsaslanır və 3 kateqoriyaya bölünürlər [7]:

- host-əsaslı (host -based);
- şəbəkə-əsaslı (network -based);
- hibrid (host- və şəbəkə-əsaslı birlikdə).

Host-əsaslı sistemlər yoluxmuş botları aşkarlamaq üçün adətən fərdi hostlarda siqnatura və ya davranış əsaslı üsullar istifadə edir. Belə ki, şəbəkə trafiki və ya sistem tarixçəsi, bot imzaları və ya davranış informasiyaları ilə öyrənilir. Host əsaslı IDS-lər tək botun aşkarlanmasına imkan verir, ona görə də daha çox digər üsullarla birlikdə tətbiq olunurlar.

Şəbəkə əsaslı üsullar şəbəkəni izləyərək bir neçə hostun oxşar davranışlarına əsasən yoluxmanı aşkarlamağa cəhd edir. Belə üsullar zamanı hostun əvvəlki siqnaturaları və ya davranış məlumatları lazım deyil, çünki, eyni botlar şəbəkədə özlərini yoluxmamış kompüterlərə görə fərqli aparırlar. Şəbəkə əsaslı aşkarlama zamanı eyni şəbəkədə bir neçə yoluxmuş kompüterin olması kifayətdir. Həmçinin belə üsullar şəbəkə administratoru ilə əməkdaşlıq tələb edir ki, bu zaman şəbəkə istifadəçilərinin təhlükəsizliyi təmin olunmalıdır.

Şəbəkə əsaslı üsullarda botnetlər şəbəkə trafikinin müxtəlif xarakteristikalarını istifadə etməklə aşkarlanır. Məsələn, istifadə olunan şəbəkə statistikası, əlaqə protokolları, şübhəli trafik davranışları, davranışların qrafik təsviri, honeypotların (bal küpü) qurulması və izlənməsi, davranış xüsusiyyətləri və s. Şəbəkə əsaslı botnet aşkarlama üsulları özü də 2 sinfə ayrılır: honeynetlər və passiv trafikə əsasən. Honeypotların qurulması və onların davranışlarının izlənməsi ilə aparılan müşahidələr daha aktiv xarakter daşıyır.

Aşkarlama texnologiyaları aşağıdakı siniflərə də bölünür:

- siqnatura əsasında;
- anomaliya əsasında;
- DNS əsasında (Domain Name System, DNS);
- intellektual analiz əsasında (mining-based).

Müdaxilələri aşkarlama üsulları olduqca müxtəlifdir. Bir çox müdaxilələri aşkarlama sistemləri botnet şəbəkəsinin normal şəbəkə ilə müqayisədə daha çox trafik istifadə etməsinə, bəziləri isə botları aşkarlamaq üçün anomaliyaların

aşkarlanmasına əsaslanır. Anomaliyanın aşkarlanması normal davranışlardan əhəmiyyətli dərəcədə sapmalara əsaslanır. Məsələn; şəbəkə davranışları, sistem səviyyəsində davranışlar, CPU istifadəsi və ya fayl sisteminin dəyişməsi kimi. Anomal aşkarlamanın əsas üstünlüyü odur ki, dəyişikliyin məqsədindən asılı olmayaraq yoluxmanı daha tez təyin etmək mümkündür.

BotMiner

Şəbəkə əsaslı müdaxiləni aşkarlama sistemlərindən biri BotMiner-dir [1]. Bütün üsulların üstün cəhətləri olsa da, bir sıra problemlər də vardır. Aşkarlamanı çətinləşdirən əsas problemlər aşağıdakılardır:

- Rutkitlərdən istifadə edərək sistem səviyyəsində analizlərin gizlədilməsi;
- Botnetlərin çox dəyişkən və fərqli C&C strukturlarının olması;
- Botlar və onların davranışlarının daim inkişaf etməsi;
- Tək strukturlu şəbəkəyə baxaraq ümumi bir aşkarlama modeli qurmağın çətinliyi və s.

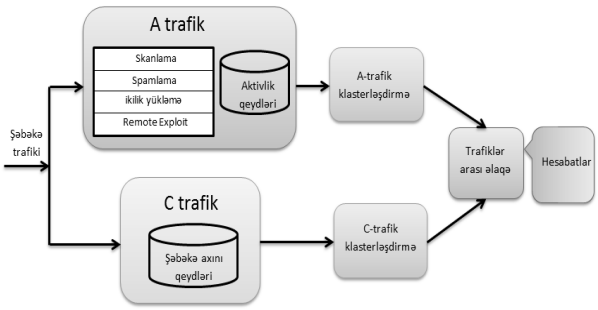
Botnetlər müxtəlif protokol və C&C infrastruktur istifadə etdiklərinə görə elə yanaşma təklif etmək lazımdır ki, protokol və strukturu müstəqil şəkildə aşkarlamağa imkan versin. BotMiner botnetlərin sabit qalan xassələrinə əsaslanır. Belə ki, botlar botmasterə uzunmüddətli istifadə üçün lazımdır. Ona görə də daimi müşahidə onu göstərir ki, eyni botnetdəki botlar adətən eyni C&C serverə bağlanır və botmasterlə eyni üsulla əlaqə saxlayırlar. Həmçinin botlar əmrləri oxşar koordinasiyalı şəkildə cavablandırırlar.

BotMiner şəbəkə aktivliyini iki hissədə araşdırır: əlaqə aktivliyi və zərərli aktivliklər. BotMinerin sistem arxitekturası aşağıdakı kimidir (Şəkil 1):

- C-trafik (C&C əlaqələr): "Kim kimlə danışır?"
- A-trafik (zərərli fəaliyyət): "Kim nə edir?".

BotMiner hostları oxşar əlaqə aktivlikləri və oxşar zərərli aktivliklərə görə klasterləşdirir. C-trafik aşağıdakıları qeyd edir:

- Şəbəkə axınını izləyir, bu iş nüvə səviyyəsində itən paketləri aşkar etmək üçün çox səmərəli şəkildə qurulmalıdır;
- Başlama vaxtı, müddəti, srcIP (çıxış ünvanı), srcPort (çıxış portu), dstIP (son ünvan), dstPort (son port), paketlərin sayı, hər iki istiqamətdə transfer olunan baytların sayı.



Şəkil 1. BotMinerin sistem arxitekturası

BotMiner üsulu ilə aşkarlama zamanı müşahidə olunan botnetlərdə (IRC, HTTP və P2P protokollarına uyğun) müşahidə olunan paketlərin ölçüsü, müddəti, miqdarı və s. xarakteristikalar qeyd edilir.

BotSniffer

BotSniffer anomaliyalara əsaslanan şəbəkə əsaslı aşkarlama üsuludur [3]. Əvvəlki siqnatura və C&C ünvanları bilmədən zərərli kanalları oxşar xüsusiyyətləri və zaman-məkan əlaqələrinə görə aşkarlamağa imkan verir. Bu üsul şəbəkədə bir neçə bot, hətta bir bot belə olduqda C&C kanalı aşkarlaya bilir. BotSniffer dünya miqyaslı real botnetlərin aşkarlanması zamanı tətbiq olunmuşdur və xəta faizi 0,18-dir.

Bildiyimiz kimi, botnetləri xarakterizə edən əsas kateqoriya C&C serverlərdir. Botnetin fəaliyyəti müddətində C&C kanalları nisbətən sabit olur və botmasterlər öz botlarını və idarəetmə variantlarını nadir hallarda dəyişirlər. Həmçinin bu əlaqə botmasterə öz bot "ordusuna" əmr vermək üçün lazımdır. Əgər C&C kanal açılarsa, həm C&C serveri, həm də şəbəkədəki botları aşkarlamaq olar. Bu gün botnetlərin əksəriyyəti IRC protokol və IRC PRIVMSG mesajlarını istifadə edirlər. Çox vaxt əmrlər şifrlənməmiş göndərilir, bəzi botnetlər isə sadə şifrləmə üsullarından (XOR, əvəzləmə və s.) istifadə edirlər.

Botnet C&C trafikini aşkarlamaq çox çətindir, çünki:

- normal protokol istifadə edir (IRC və HTTP) və normal trafikə oxşayır;
- aşağı həcmdə trafik istifadə edir və gizlətmək çətin deyil;
- izlənilən şəbəkədə çox az sayda bot ola bilər;
- şifrlənmiş əlaqə istifadə edilə bilər.

BotSniffer alqoritminin aşağıdakı kimi üstün cəhətləri var:

1. C&C serverlərin əvvəlki bilgiləri və ya istifadəçinin siqnaturaları tələb olunmur;
2. Şifrlənmiş kanalları belə aşkarlamağa imkan verir;
3. İzlənilən şəbəkədə çox sayda bot olması tələb edilmir;
4. Böyük həcmdə C&C əlaqə paketi tələb etmir və yanılma ehtimalı çox azdır.
5. Şəbəkələrdə əlaqələr sorğu-cavab texnologiyası əsasında həyata keçirilir. Normal şəbəkələrdə eyni vaxtda bir neçə müştərinin oxşar cavablar verməsi nadir hadisədir. Əgər şəbəkədə bir neçə belə davranış müşahidə olunursa, onların bot olub-olmamasının yoxlanması BotSniffer aşkarlama sisteminin əsasını təşkil edir.

BotSnifferin əsas komponentləri izləmə mexanizmi (Monitor Engine) və korrelyasiya mexanizmidir (Correlation Engine).

İzləmə mexanizmi şəbəkə trafikini araşdırır, şübhəli C&C protokollarla əlaqələri qeyd edir və aktiv cavab davranışlarını (məs., skanlama, spam və s.) və mesaj cavabları aşkarlayır. Daha sonra bu məlumatlar korrelyasiya mexanizmi tərəfindən məkan-zaman əlaqəsi və fəaliyyətlərin oxşarlığına görə qrup şəklində analiz olunur. Belə mexanizmlər bir neçə şəbəkədə qurulur və korrelyasiya analizi üçün məlumatlar toplanır.

Əvvəlcə lazım olmayan trafik ayrılır. C&C protokolun təyini zamanı C&C əlaqələrdə istifadə olunmayan ICMP, UDP kimi protokollar süzgəclənir. Google, Yahoo və s. kimi C&C

server olma ehtimalı az olan trafiklər isə ağ siyahıya salınır və vaxtaşırı olaraq dəyişdirilir.

Müşahidə edilən müştərilərdən hansıların C&C-yə bənzər protokol istifadə etdiyini müəyyənləşdirmək üçün əsasən IRC və HTTP protokollar izlənilir. Bildiyimiz kimi, bot əmrlərə iki cür cavab verir: aktiv və mesaj şəklində. Bunun üçün də mesaj cavablar kimi IRC PRIVMSG mesajları əlaqələr analizi üçün izlənilir. Burada 2 anomal modul hesablanır: anormal yüksək skan dərəcəsi və itən əlaqələrin həcmi.

Əlaqələr səviyyəsində əvvəlcə müştərilər IP və port cütünə görə qruplaşdırılır və sonra zaman-məkan əlaqəsi və oxşarlıq xüsusiyyətlərinə görə qrup analizi aparılır. Əgər hər hansı şübhəli C&C əlaqəsi aşkar edilərsə, bu bot siqnalı ola bilər.

BotHunter

Şəbəkə əsaslı aşkarlama üsullarından biri də BotHunterdir. BotHunter yoluxmanı və botlar arasında əlaqəli dialoqu ardıcıl yoluxma modelini istifadə edərək tapır [2]. Bu üsul Snort IDS-in dəyişdirilmiş variantını istifadə edərək şəbəkə çıxışında paketləri qeyd edir. Botun yoluxmasını müəyyən etmək üçün çarpaz analizdən istifadə olunur. IDS siqnalları zaman üzrə izlənilir və bütün hostlar üçün hər ana uyğun yoluxma ehtimalı hesablanır. Əgər aşağıdakı iki hal müşahidə olunarsa, onun haqqında bot olma qərarı alınır:

1) Lokal yoluxma haqqında sübut tapıldıqda və botun hər hansı hücum cəhdi sübut olunduqda;
2) Ən azı iki fərqli çıxış siqnaturası və ya bot əlaqəsi tapıldıqda.

NƏTİCƏ

Botnetlər ölkələrin kritik informasiya resurslarına genişmiqyaslı kiber-hücumların həyata keçirilməsində əsas vasitələrdən biridir. Bu işdə aparılmış analizlər göstərir ki, botnetlərin aşkarlanmasına yanaşmalar maşın təlimi və verilənlərin intellektual analizi metodlarına əsaslanır. Big Data texnologiyaları müxtəlif ölkələri əhatə edən geniş coğrafi əraziyə səpələnmiş botların şübhəli fəaliyyətini izləməyə, milyonlarla mənbədən informasiyanı toplayaraq korrelyasiya etməyə, ilk baxışda əlaqəsiz görünən, zamana görə bir-birindən uzaq hadisələr arasında səbəb-nəticə əlaqələri tapmağa imkan verməklə botnetlərin effektiv tapılmasına imkan verə bilər. Gələcək tədqiqatların bu istiqamətdə aparılması planlaşdırılır.

ƏDƏBİYYAT

[1] G. Gu, R. Perdisci, J. Zhang, W. Lee, “BotMiner: clustering analysis of network traffic for protocol- and structure-independent botnet detection,” Proc. of the 17th Conference on Security Symposium (SS’08), pp. 139–154, 2008.

[2] G. Gu, P. Porras, V. Yegneswaran, M. Fong, W. Lee, “Bothunter: detecting malware infection through IDS driven dialog correlation,” Proc.of 16th USENIX Security Symposium, pp. 1–16, 2007.

[3] G. Gu, J. Zhang, W. Lee, “BotSniffer: detecting botnet command and control channels in network traffic,” Proc. of the 15th Network and Distributed System Security Symposium (NDSS), 2008.

[4] C. Li, W. Jiang, X. Zou, “Botnet: survey and case study,”Proc. of the 4th International Conference on Innovative Computing Information and Control, Vol. 0, pp. 1184–1187, 2009.

[5] J. Liu, Y. Xiao, K. Ghaboosi, H. Deng, J. Zhang, “Botnet: classification, attacks, detection, tracing, and preventive measures,” EURASIP Journal on Wireless Communications and Networking, Vol. 2009, Article ID 692654, 2009.

[6] S. A. Yeshwantrao, V. J. Jadhav, “Threats of botnet to Internet security and respective defense strategies,” International Journal of Emerging Technology and Advanced Engineering, Vol. 4, No. 1, pp. 121-127, January 2014.

[7] S. García, A. Zunino, M. Campo, “Survey on network-based botnet detection methods,” Security and Communication Networks, Vol. 7, No. 5, pp. 878–903, January 2014.

[8] TrendMicro. Taxonomy of botnet threats. Technical Report, 2006. http://www.cs.ucsb.edu/ kemm/courses/cs595G/TM06.pdf

[9] D. Zhao, I. Traore, B. Sayed, W. Lu, S. Saad, Ali Ghorbani, D. Garant, “Botnet detection based on traffic behavior analysis and flow intervals,” Computers and Security, Vol. 39, pp. 2-16, November 2013.

[10] R. Sharifnya, M. Abadi, “DFBotKiller: Domain-flux botnet detection based on the history of group activities and failures in DNS traffic,” Digital Investigation, Vol. 12, pp. 15-26, March 2015.

[11] OpenDNS Security Whitepaper. The Role of DNS in Botnet Command & Control. http://info.opendns.com/rs/opendns/images/OpenDNS_SecurityWhitepaper-DNSRoleInBotnets.pdf

[12] R. A. Rodríguez-Gómez, G. Maciá-Fernández, P. García-Teodoro, M. Steiner, D. Balzarotti, “Resource monitoring for the detection of parasite P2P botnets,” Computer Networks, Vol. 70, pp. 302-311, 2014.

[13] K. Singh, Sh. Ch. Guntuku, A. Thakur, C. Hota, “Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests,” Information Sciences, Vol. 278, pp. 488-497, September 2014.

[14] S. S.C. Silva, R. M.P. Silva, R.C.G. Pinto, R. M. Salles, “Botnets: A survey,” Computer Networks, Vol. 57, No. 2, pp. 378-403, 2013.

[15] L. P. Song, Z. Jin, G.Q. Sun, “Modeling and analyzing of botnet interactions,” Physica A: Statistical Mechanics and its Applications, Vol. 390, Issue 2, pp. 347-358, January 2011.

[16] A. McKewan, “Botnets – zombies get smarter,” Network Security, Vol. 2006, Issue 6, pp. 18-20, June 2006.

[17] J. Jabez, B. Muthukumar, “Intrusion Detection System (IDS): Anomaly Detection Using Outlier Detection Approach,” Procedia Computer Science, Vol. 48, pp. 338-346, 2015.

[18] B. McCarty, “Botnets: big and bigger,” IEEE Security & Privacy, Vol. 1, No. 4, pp. 87–90, 2003.

[19] C. Wilson, Botnets, cybercrime, and cyberterrorism: vulnerabilities and policy issues for congress. Congressional Research Service Reports (CRS) and Issue Briefs. 2007.

[20] Y. Zhang, Y. Xiao, K. Ghaboosi, J. Zhang, H. Deng, “A survey of cyber crimes,” Security and Communication Networks, Vol. 5, No. 4, pp. 422–43, 2012.

[21] Ch. Day. Intrusion Prevention and Detection Systems. in Managing Information Security (Second Edition), pp. 119-142, 2013.

[22] A. Patel, M. Taghavi, K. Bakhtiyari, J. Celestino Júnior, An intrusion detection and prevention system in cloud computing: A systematic review,” Journal of Network and Computer Applications, Vol. 36, Issue 1, pp. 25-41, January 2013.

[23] A. Castiglione, R. De Prisco, A. De Santis, U. Fiore, F. Palmieri, “A botnet-based command and control approach relying on swarm intelligence,” Journal of Network and Computer Applications, Vol. 38, pp.22-33, February 2014.

[24] S. Lee, J. Kim, “Fluxing botnet command and control channels with URL shortening services,” Computer Communications, Vol. 36, Issue 3, pp. 320-332, February 2013.

[25] Y. Boshmaf, I. Muslukhov, K. Beznosov, M. Ripeanu, “Design and analysis of a social botnet,” Computer Networks, Vol. 57, Issue 2, pp. 556-578, February 2013.

[26] A. K. Seewald, W. N. Gansterer “On the detection and identification of botnets,” Computers & Security, Vol. 29, Issue 1, pp. 45-58, February 2010.

[27] C. Schiller, J. Binkley, G.Evron, C. Willems, T. Bradley, D. Harley, M. Cross. Botnets: the killer web app. Syngress. 2007.

“Big Data” Təhlükəsizliyi və Həlli yolları

Tural Yunusov
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
turaly@mail.ru

***Xülasə* — Məqalə “Big Data” təhlükəsizliyi və həlli yollarına həsr edilmişdir. Məqalədə “Big Data” fenomeni, “Big Data” texnologiyası, verilənlərin təhlükəsizlik məsələləri, o cümlədən verilənlərin idarəedilməsi, icazələrin idarəedilməsi, məlumatın mühafizəsi və konfidensiallığı, şəbəkə təhlükəsizliyi analiz edilmiş və mövcud problemlər müəyyən edilmişdir.**

Açar sözləri— “Big Data”; “Big Data” Təhlükəsizliyi; “Big Data” təhlükəsizlik meyarları; “Big Data”-nın emal edilməsi; konfidensiallıq; DLP sistemlər

I. GİRİŞ

Keçmişdən günümüzə qədər gələn müddət ərzində məlumat gündən-günə artmış, hətta son illərdə İnformasiya Texnologiyalarının sürətli inkişafı ilə əlaqədar dəfələrlə çoxalmışdır. Bunun nəticəsində "informasiya çirkliliyi" adlanan termin yaranmışdır. Bir çox şirkətlər bu mövzu ilə əlaqədar araşdırma apardıqdan sonra “Big Data” (Böyük Verilənlər) termini yaranmışdır.

“Big Data” dedikdə müxtəlif yollarla əldə edilən informasiyaların (şəbəkə, bloq, şəkil, video, log faylları və s.) yığılaraq yaratdıqları böyük həcmdə və normal alətlərlə emal oluna bilməyən verilənlər toplusu nəzərdə tutulur. Bu yeni termin həcm və mürəkkəblik baxımından mövcud idarəetmə metodları və intellektual analiz vasitələri ilə emal oluna bilməyən verilənləri təyin etmək üçün istifadə edilir.

İnkişaf etməkdə olan İnformasiya Cəmiyyəti böyük elmi-texniki inqilab çağındadır və bu inqilab biznesdən başlayaraq elmə qədər həyatımızın bütün sahələrinə nüfuz etməkdədir. Bu eyni zamanda sosial-iqtisadi, təhlükəsizlik elmi baxımdan mürəkkəb bir problemdir. Problem böyük verilənlər ideyası ilə daha da dərinləşməkdədir. Belə ki, “Big Data” saxlanması, idarə edilməsi və onlardan informasiya əldə edilməsi ciddi problem yaratmışdır. Problem müxtəlif mənbələrdən avtomatik və fasiləsiz olaraq generasiya olunan verilənlərin real-vaxt ərzində emalı və analizində mövcud İT (İnformasiya Texnologiyaları) həllərin səmərəsiz olmasındadır. Mövcud vəziyyət “verilənləri yaratmaq son dərəcə asan, emal etmək isə son dərəcə çətin olur” fikrini söyləməyə əsas verir [1].

“Big Data” texnologiyalarının təqdim etdiyi imkanlardan informasiya təhlükəsizliyi problemlərinin həlli üçün geniş istifadə etmək mümkündür. İnformasiya təhlükəsizliyi və fərdi məlumatların qorunması “Big Data” üçün həlli vacib məsələlərdəndir. Bu problemlərə böyük həcmdə verilənlərin idarə edilməsi infrastrukturunda informasiya təhlükəsizliyi, verilənlərin konfidensiallığı və fərdi məlumatların qorunması, girişə nəzarət, audit və informasiya təhlükəsizliyinin real zaman rejimində monitorinqi və s. daxildir [2].

“Big Data” texnologiyaları informasiya təhlükəsizliyi insidentlərinin avtomatik cavablandırılması, onların nəticələrinin qısa müddətdə, minimal xərclərlə aradan qaldırılması və daha keyfiyyətli tədqiqatın aparılmasında xüsusi rol oynayır [3].

II. BİG DATA TEXNOLOGİYALARI

“Big Data” platformasının meydana gəlməsində beş komponent iştirak edir. Bunlar həcm (volume), sürət (velocity), müxtəliflik (variety), həqiqilik (veracity) və dəyər (value) –dən ibarətdir. İngilis dilli mənbələrdə bunu «5V» də adlandırırlar [4].

Müxtəliflik (variety): Emal olunan informasiyanın 80% -i strukturlaşmış (strukturlaşmış - verilənlər bazaları ilə əlaqəli olan) informasiya deyildir və hər yeni texnologiya müxtəlif tiplərdə informasiya emal edir. Telefonlardan, planşetlərdən, birləşmiş şəbəkələrdən müxtəlif tip informasiya emal edilir və bu informasiyaların fərqli dillərdə ola biləcəyi nəzərə alınmalıdır, həmçinin bu verilənlərin bir-birlərinin tipinə çevrilmələri də vacibdir.

Sürət (Velocity): Böyük verilənlərin emal sürəti çox yüksəkdir və ildən ilə bu sürət artmaqda davam edir. Sürətlə artan verilənlər, əməliyyatların sayının və müxtəlifliyinin də eyni sürətlə artmasına gətirib çıxarır.

Həcm (Volume): IDC (International Data Corporation) hesabatına əsasən 2020-ci ildə informasiyanın həcmi 2009-cu ildəkindən 44 dəfə çox olacaq. Hal-hazırda istifadə edilən, "böyük" adlandırdığımız məlumat saxlama yaddaşları və "böyük sistemlər" gələcəkdə sürətlə artan verilənlər üçün kifayət etməyəcək [5, 6].

Həqiqilik (Veracity): İnformasiya axını zamanı verilənlərin “etibarlı” şəkildə təşkil olunması vacib məsələdir. Verilənlərin təhlükəsizlik səviyyəsinin monitorinqi, səlahiyyətli şəxslər tərəfindən istifadəsi və ya gizli qalması öz vacibliyini qoruyur.

Dəyər (Value): “Big Data” platformasının ən əhəmiyyətli komponenti isə dəyər yaratmasıdır. Bütün yuxarıda sadalananlar “Big Data” komponentlərinin emalından sonra təşkilat üçün bir müsbət dəyər yaratmalıdır. Məsələn:

Müəyyən qərarlar verərək yoluxucu xəstəliklərin qarşısını almaq üçün səhiyyə mövzusunda strateji qərarlar verən bir dövlət təşkilatının bölgə, rayon, xəstəlik, dərman, həkim məlumatlarını əldə etməlidir. Digər bir misal, istənilən bank kredit verəcəyi adamın yalnız demoqrafik məlumatlarını deyil, onun tətil etmə vərdişlərini belə izləyə bilməli, lazım gələrsə sosial şəbəkələrdə nə etdiyini görə bilməlidir.

“Big Data”-nın emal edilməsi

Hal-hazırda istifadə olunan verilənlər bazaları inkişaf etsə də, böyük verilənləri emal etmək üçün bizə lazım olan imkanları vermir. Çünki, əldə edilən məlumatlar bir-birləri ilə əlaqəli olmur, bu səbəbdən də bazalarla əlaqə yaratmaq üçün tələb olunan SQL sorğularını yazmaq və böyük məlumatları analiz etmək çətinlik yaradır. Bununla birlikdə əldə etdiyimiz məlumatlar strukturlaşmış olmur, yəni strukturlaşdırılması üçün yenidən baxılır. Bizə lazım olan məlumatları əldə etməyimiz üçün klassik üsullardan daha fərqli həllər lazımdır [7]. Məsələn, Google özü yaratdığı texnologiya əsasında məlumatları yığır (Google File System), emal edir (MapReduce) və saxlayır (Big Table). Eyni şəkildə Amazon da özü yaratdığı texnologiya (DynamoDB) əsasında məlumatları yığır. Bu şirkətlər yaratdıqları texnologiyalar haqqında yalnız elmi yazılar dərc etdirir və bir çox proqramçılar bu texnologiyaları inkişaf etdirir, məsələn, Hadoop, HBase, Lucene kimi texnologiyaları [8]. Facebook, Twitter, LinkedIn kimi şirkətlər isə daha inkişaf etmiş texnologiyalar təklif edir, məsələn,. Cassandra, Hive, Pig, Voldemort, Storm kimi texnologiyalar.

III. BİG DATA TƏHLÜKƏSİZLİYİ MEYARLARI

Big Data Təhlükəsizliyin meyarları aşağıdakıladan ibarətdir [9].

1. Verilənlərin idarə edilməsi
2. İcazələrin idarə edilməsi
3. Məlumatın mühafizəsi və konfidensiallıq
4. Şəbəkə təhlükəsizliyi

“Big Data” təhlükəsizliyi kriterilərinə yuxarıda sadalanan 5 əsas təhlükəsizlik riski və təhlükəsizlik riskinin azaldılması üçün zəruri olan 21 alt komponent daxildir. Ümumi təhlükəsizlik kriteriləri aşağıdakılardır [10]:

– *Verilənlərin idarə edilməsi.* Verilənlərin idarə edilməsi üç əsas qrupa bölünür: verilənlərin təsnifatı, verilənlərin aşkarlanması və verilənlərin nişanlanması.

Verilənlərin təsnifatı “Big Data” platformasında effektiv təhlükəsizlik nəzarətinin həyata keçirilməsində əhəmiyyətli rol oynayır. Çox vaxt təşkilatlarda böyük həcmdə verilənlər yığılır. Verilənlərin hansı mövzularda olduğunu aydınlaşdırmaq, onlar arasında hansı səviyyədə kriptoqrafiq mühafizənin təmin edilməsi lazım gəlir.

Müəyyən situasiyalarda həsas verilənlərin aşkarlanması təşkilatı bir çox risklərdən azad edir. Verilənlərin aşkarlanması iki formada yerinə yetirilir: strukturlaşmış verilənlər üçün və strukturlaşmamış verilənlər üçün.

Strukturlaşmış verilənlər – verilənlər bazaları ilə əlaqəli olan verilənlərdir. Buna misal olaraq Hadoop (Böyük verilənlərin de-fakto standartı hesab olunan Apache Software Fondation-un layihəsi *Hadoop* paylanmış hesablama mühitində böyük verilənlərin emalı və analizi üçün əsas platformadır) üçün verilənlərin yerini və təsnifatını təyin edən, verilənlər bazası ilə məlumatların əlaqələndirilməsi kimi *comma-separated values* (CSV) və ya *JavaScript Object Notation* (JSON) fayl tiplərini göstərmək olar. Bu səbəbdən sütunların və xanaların təhlükəsizliyi proqramlar vasitəsi ilə həll edilir. Məsələn mütəxəssis tərəfindən sütun və sətirlər müəyyən təhlükəsizlik səviyyəsində gizlədilə və ya göstərilə bilər [11].

Strukturlaşmamış verilənlər – bir-biri ilə və verilənlər bazaları ilə əlaqəli olmayan verilənlərdir**.** Strukturlaşmamış verilənlər üçün təsnifatın aparılması və emal edilərək həsas verilənlərin çıxardılması daha çətindir [12].

– *İcazələrin idarə edilməsi*. Verilənlər bazaları və ona olan icazələr mütləq şəkildə idarə olunmalıdır. Əgər istifadəçilər daha çox səlahiyyətlərə malik olarsa, bu verilənlərin dəyişdirilməsi və ya silinməsinə gətirib çıxarar. İcazələrin idarə edilməsi üçün istifadəçi avtorizasiyası qaydaları yaradılmalı, mərkəzləşdirilmiş icazələrin idarə edilməsi siyasəti ilə istifadəçilərə məlumatlara icazələr təyin edilməlidir.

– *Məlumatın mühafizəsi və konfidensiallığı.* Verilənlərin mühafizəsi və konfidensiallığı böyük verilənlərin təhlükəsizliyində çox böyük əhəmiyyətli rol oynayır. Məlumatın mühafizəsi və konfidensiallığı iki səviyyədə həyata keçirilir: Tətbiq səviyyəsində şifrlənmə – verilənlər proqram təminatı səviyyəsində şifrlənir; disk səviyyəsində şifrləmə – burada verilənlərin yerləşdiyi yaddaş qurğuları bütövlüklə şifrlənir.

– *Şəbəkə təhlükəsizliyi.* Böyük verilənlərin yığılması, analizi və emalında şəbəkədən istifadə olunur. Bu səbəbdən verilənlərin şəbəkə vasitəsi ilə ötürülməsi zamanı təhlükəsizlik təmin edilməlidir.

IV. VERİİLƏNLƏRİN TƏHLÜKƏSİZLİK HƏLLƏRİ

Böyük verilənlərin təhlükəsizlik problemlərinin həlli üçün şirkətlər müxtəlif praktiki üsullardan istifadə edir. Bu həllərdən biri də (DLP – Data Loss Prevention) verilənlərin itkisinin qarşının alınması sistemlərinin istifadəsidir [13]. DLP sistemlər həssas verilənlərin şəbəkə daxilində icazəsiz istifadəsini və monitorinqini apararaq təhlükəsizliyini təmin edir. DLP sisteminin şəbəkəyə qoşulma sxemi aşağıdakı kimidir (şəkil.1).

Şəkil 1. DLP sisteminin şəbəkəyə qoşulma sxemi

DLP sistemlərində verilənlərin tipləri

İnformasiya təhlükəsizliyində “verilənlərin təhlükəsizliyi” kateqoriyasında qiymətləndirilən verilənlərin itkisinin qarşısının alınması texnologiyasının məqsədi verilənlərin arxivdən istifadəçiyə təhlükəsiz şəkildə çatdırılmasını təmin etməkdir. Bu məqsədlə DLP sistemlər üç səviyyədə təhlükəsizliyi təmin edir [14]:

– *Hərəkətli verilənlər (Data in Motion).* Şəbəkə daxilində hərəkət edən, yəni e-poçt, ismariclər, veb və P2P (Peer-to-peer) ötürmə kanalları üzərindən daimi hərəkət edən verilənlər tipidir.

– *İstifadə olunmayan verilənlər (Data at Rest).* Verilənlər bazası, qovluq sistemləri və digər xüsusi arxiv bölmələrində saxlanılan, əsasən ilk addımda təhlükəsizliyinin təmin olunması vacib olmayan verilənlər tipidir.

– *İstifadə olunan verilənlər (Data in Use):* Son istifadəçinin daimi olaraq istifadə etdiyi və verilənlər bazası ilə əlaqəli olan aktiv verilənlər tipidir.

NƏTİCƏ

“Big Data” böyük informasiya massivlərindən istifadə etməyə imkan verən yeni nəsil texnologiyadır. Bu texnologiyadan düzgün istifadə müxtəlif fəaliyyət sahələrində iş prossesinin sürətlənməsinə gətirib çıxarır. Böyük həcm, sürət və müxtəliflik kimi xüsusiyyətlərlə xarakterizə olunan verilənlərin emalı və analizi məqsədilə MapReduce, Hadoop, HDFS, NoSQL və s. proqram-aparat platformaları yaradılmışdır.

“Big Data” texnologiyalarının təhlükəsizlik sahəsində istifadəsi təhdidlərin və risklərin erkən aşkar edilməsinə inkan verər. Bunla yanaşı “Big Data”-nın emal prosesi zamanı təhlükəsizliyinin təmin edilməsi çox vacib məsələdir və məsələnin həlli üçün yeni metodların işlənməsi aktualdır. Böyük həcmli verilənlərin emalı sahəsində superkompüterlərin, qrid və bulud texnologiyaların istifadəsi ilə müəyyən problemlər aradan qaldırılsa da, bu sahədə hələ də ciddi texniki və elmi-nəzəri problemlər möcuddur. Problem ancaq böyük həcmdə verilənlərin saxlanılması və idarə olunmasında deyil, həm də strukturlaşdırılmamış verilənlərin analizi və nəticələrin interpretasiyasındadır.

ƏDƏBİYYAT

[1] R. Əliquliyev, M. Hacırəhmanova, “Big Data” Fenomeni: problemlər və imkanlar // İnformasiya texnologiyaları problemləri jurnalı, 2014, №2, 3-16

[2] Big Data, Big Impact: New Possibilities for International Development, world economic forum, 2012, *www.weforum.org*

[3] EMC Big Data 2020 Projects, http://www.emc.com/leadership/digital-universe/iview/big-data-2020.htm

[4] Oracle and Big Data: Big Data for the Enterprise, 2013, http://www.oracle.com/

[5] Big Data Solutions, 2013, http://www8.hp.com/

[6] Big Data, 2013, http://www.microsoft.com/.

[7] D. Agrawal, S. Das, A. Abbadi, Big Data and Cloud Computing: Current State and Future Opportunities, EDBT, Uppsala, Sweden, march 22–24, 2011.

[8] Hadoop Distributed File System, http://hadoop.apache.org/docs/

[9] Securing Big Data Security issues with Hadoop environments, Securosis, https://securosis.com/blog/securing-big-data-security-issues-with-hadoop-environments

[10] Ways to Build a Better Big Data Security Strategy, http://www8.hp.com/h20195/V2/GetPDF.aspx/4AA5-0863ENW.pdf

[11] Big Data Security Assessment, http://www.intel.ru/content/dam/www/public/us/en/documents/datasheets/big-data-security-assessment-datasheet.pdf

[12] Big Data Analytics for Security Intelligence, https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf

[13] T. Smith, Big Data Security: The Evolution of Hadoop’s Security Model, http://www.infoq.com/articles/HadoopSecurityModel

[14] Data loss prevention - Ernst & Young, http://www.ey.com/Publication/vwLUAssets/EY_Data_Loss_Prevention/$FILE/EY_Data_Loss_Prevention.pdf

[15] Symantec Data Loss Prevention Solution, https://www.symantec.com/content/en/us/enterprise/fact_sheets/data-loss-prevention-solution-ds-21350666.pdf

Роль Big Data в Обеспечении Безопасности Детей в Интернете

Назакет Меликова
Институт Информационных Технологий НАНА, Баку, Азербайджан
naranara_68@mail.ru

Аннотация — **В данной статье мы рассматриваем виды рисков, ожидающих ребенка в Интернете, определение термина Big Data, а также роль Big Data в предотвращении рисков и обеспечении безопасности детей в Интернете.**

Ключевые слова — безопасность; Big Data; методика анализа больших данных; кибер-риски.

I. ВВЕДЕНИЕ

Много лет назад, когда Интернет только начал входить в жизнь простых людей, даже взрослые люди боялись начать осваивать эту, прежде неисследованную, территорию. Как оказалось, прогресс обогнал взрослых, и все чаще новое поколение помогает сориентироваться в Интернете своим родителям. Конечно, такое положение дел не может не радовать, так как естественно, наше будущее зависит от наших детей и поэтому поступающая к детям информация должна отфильтрована как по возрастным категориям, так и по моральным принципам. В связи с этим, в нынешнюю информационную эру крайне остро поставлен вопрос обеспечения безопасности детей в Интернете.

Число пользователей Интернета в Азербайджане стремительно растет, и стоит отметить, что доля лиц, в возрасте до 18 лет достаточна велика. Не удивительно, что Интернет является информационной средой для подростков и детей, без которой многие юные пользователи не представляют себе жизнь. Ведь в киберпространстве можно найти единомышленников и обсудить с ними насущные вопросы, отыскать нужную информацию для учебных целей, послушать любимого исполнителя или приобрести редкую книгу. Таким образом, Интернет помогает детям как по учебе, так и в организации досуга. Но, кроме этого существуют многочисленные социальные сети, которые сами по себе затягивают ребенка. Онлайн-игры, а также огромное количество «вредных» сайтов, из которых дети получают информацию, совершенно для них не предназначенную, а иногда и опасную для психического, или даже физического здоровья. Выявить из такого большого объема не структурированных данных безопасную для детей информацию поможет технология Big Data.

Во всем мире технология Big Data успешно применяется. Внедрение Big Data во многих областях деятельности обещает увеличение производительности, сокращение расходов и т.д. Однако, чаще всего из виду упускается то, что сам по себе сбор большого количества исходных данных не является полезным или выгодным. Перспектива возможности использования данных есть статистические зависимости, которые можно почерпнуть из их анализа. Это действительно имеет значение [1]. Перспектива того, как вы можете использовать данные — и статистические зависимости, которые можно почерпнуть из их анализа, — вот что действительно имеет значение [1]. В Азербайджане разработка и применения технологий анализа данных началась недавно, причем данные технологии направлены в основном на бизнес стратегии, а не на обеспечение безопасности пользователей, и тем более детей. Однако, мы считаем, что применение Big Data в сфере обеспечения безопасности детей принесет большую пользу как детям, так и нашему государству, и это именно то направление, в котором использование Big Data будет рентабельно, целесообразно и продуктивно.

II. ВИДЫ ИНТЕРНЕТ- РИСКОВ

Статистика показывает, что каждую минуту в Интернете появляется много новых сайтов, на онлайн-шоппинг пользователи тратят $272 070, сервисы электронной почты отправляют 204 млн писем и т.д. [2]. Цифры действительно впечатляют. И таким образом наши дети ежеминутно подвергаются опасности в киберпространстве.

Итак, как же нам защитить своего ребенка в этой полезной, но, не редко опасной среде? Сперва разберем, какие виды риска поджидают нас в Интернет-среде.
Среди предполагаемых опасностей киберпространства можно выделить 4 основные группы: [3]

- ***контентные риски*** (сведения, включающие в себя элементы эротики, насилия, агрессии, ненормативной лексики; данные, зарождающие межрасовую вражду, пропагандирующие наркотики, суицид и прочие материалы, способные спровоцировать причинение вреда здоровью);
- ***коммуникационные риски*** (уголовно наказуемые контакты, а именно, груминг- общение взрослых и установление дружеских отношений с целью изнасилования, киберпреследования (или кибербуллинг)- домогательство, оскорбления по средством личных сообщений и др.);

- ***электронные или кибер-риски*** (хищение персональных данных, спам и вирусная атака (при помощи таких вредоносных или шпионских программ, как черви, вирусы и «троянские кони»), кибер-мошенничество, и т.д.);
- ***потребительские риски*** (приобретение недоброкачественного товара и незапланированные финансовые потери, злоупотребление правами потребителя, различные поделки продукции, продаваемой посредством Интернета, фальсификация изделий, хищение персональной информации с целью онлайн-мошенничества, агрессивная реклама и др.).

Столкновение с любым из этих рисков может тяжело повлиять на психическое или физическое состояние ребенка.

III. МЕТОДИКИ АНАЛИЗА BIG DATA

Как известно, Big Data- это совокупность технологий, которые должны быстро и результативно обрабатывать большие объемы структурированных и плохо структурированных данных параллельно в разных аспектах, причём, этих данных не просто много, их объемы постоянно растут.

Методики анализа Big Data довольно разнообразны, причем исследователи продолжают трудиться над разработкой современных методик и усовершенствованием имеющихся.

Наиболее широко распространены следующие виды методики: [4]

1) Поочередное сравнение контрольной выборки с другими, также широко применяется при работе с данными меньших объемов (A/B testing);
2) выявление взаимосвязей (Association rule learning);
3) классификация на основе выявления ранее не заданных признаков (Cluster analysis);
4) сбор данных из большого количества источников (Crowdsourcing);
5) обнаружение данных в ранее не рассматриваемых для данной цели знаниях, используемых для принятия решения (Data mining, в нем применяются следующие методы: association rule learning, classification, cluster analysis, regression);
6) использование множества предикативных моделей для повышения точности прогнозов (Ensemble learning);
7) представление решений в виде так называемых ‘хромосом`, которые аналогично могут комбинироваться и мутировать; за основу принимается решение, которое наиболее «приспособлено» (Genetic algorithms);
8) генерирование self-learning алгоритмов на основе анализа и оценки эмпирических данных (Machine learning);
9) совокупность присущих информатики и лингвистики методик, направленных на распознавание естественного человеческого языка (Natural language processing);
10) анализ и выявление взаимосвязи узлов в сетях (Network analysis);
11) Набор численных методов для оптимизации систем и улучшения результата, стремящегося к идеальному (Optimization);
12) совокупность заданных шаблонов с элементами обучения «без учителя» для предсказания поведенческой модели пользователей (Pattern recognition);
13) математическая модель предсказания событий (Predictive modeling);
14) статистические методы для анализа влияния одной или нескольких независимых переменных на зависимую, причем соответствующая зависимость математическая (Regression);
15) метод выявления нужной информации из общего потока, основанный на распознавании естественного человеческого языка (Sentiment analysis);
16) отчасти заимствованный набор методик анализа пространственных данных (Spatial analysis);
17) наука о сборе, организации и интерпретации данных, включая разработку опросников и проведение экспериментов (Statistics);
18) выявление функциональной взаимосвязи на основе машинного обучения (Supervised learning);
19) моделирование поведения сложных систем часто используется для прогнозирования, предсказания и проработки различных сценариев при планировании (Simulation);
20) анализ повторяющихся с течением времени последовательностей данных (Time series analysis);
21) совокупность методик, схожих с Cluster Analysis и основанных на Machine Learning, позволяющих выявить скрытые функциональные взаимосвязи (Unsupervised learning).

IV. BIG DATA В БОРЬБЕ С ИНТЕРНЕТ- РИСКАМИ

A. Контентные риски

Как мы видим, технология Big Data –это революционное решение в области анализа и обработки данных, и эффективно применяется во многих областях, например, в сфере Интернет-рекламы. С таким же успехом

мы можем также использовать данную технологию для обеспечения безопасности детей.

Анализ запросов поисковой системы и правильно построенный алгоритм могут помочь в этом деле. Существующие ныне детские браузеры слишком ограничены, так как сортировка ссылок для поисковой системы происходит вручную. Это - своего рода "сокращенный вариант" Интернета, куда вошли лишь те ресурсы, которые составители проекта посчитали безопасными для детей. Например, в браузере «Гогуль», направленном на русскоязычную аудиторию, по умолчанию доступ разрешён более чем к 7000 развивающих и обучающих ресурсов на русском языке, отобранных и одобренных детскими психологами и педагогами [5].

Естественно, в Интернет-пространстве в разы больше полезной для детей информации, и применение технологии BigData значительно разнообразило бы досуг детей, пользующихся детским браузером, и повысило безопасность тех несовершеннолетних пользователей, которые игнорируют детские поисковые системы.

B. Коммуникационные риски

Коммуникационные риски также можно предотвратить, проанализировав в социальных сетях, онлайн-мессенджерах, чатах, форумах входящие или отправляемые ребенком в личной переписке данные на предмет угрозы со стороны взрослых или сверстников. Big Data аналитика позволяет вычислить из потока данных те самые доли процента информации, необходимых для принятия решение об опасности данной переписки.

Если с помощью технологий Big Data возможно проконтролировать социальные настроения, выявить информацию о готовящихся терактах, волнениях, то почему же не направить анализ данных на обеспечение безопасности детей? На основе уже известных алгоритмов, с помощью машинного обучения можно отследить тематику разговора и своевременно определить попытки кибер-буллинга или груминга и при помощи правоохранительных органов предотвратить возможное преступление и уголовно наказать виновных.

Также, анализ поисковых запросов помогает сформировать понимание о направлении работы, круге интересов. Например, Google имеет возможность сохранять поисковые запросы в привязке к аккаунту пользователя [6]. Предположительно, это помогает выявить социально- опасного человека. И близкое общение ребенка с ним может насторожить.

C. Электронные риски

Возможность столкновения с хищением, незаконным использованием личной информации, заражением компьютера вирусом, спам-атакой называется кибер-риском. Произойти это может при помощи распространения и проникновения вредоносного программного обеспечения, не только посредством электронных носителей, но и через Интернет, при принятии спама или закачки вредоносных файлов.

Справится с этим также поможет технология анализа Big Data. Сегодня ни одна крупная компания не обходится без технологии Big Data для защиты компьютеров при разработке программного обеспечения. Но, с каждым днем появляются новые, способные проникнуть в сеть или компьютер с большим успехом вредоносные программы, количество, скорость и сложность производимых атак возрастает, а необходимый для их обнаружения объем вычислений уже невозможно выполнять на устройствах пользователя своевременно.

Облачные сервисы, применяемые в большинстве антивирусов, способны объединить данные от разных пользователей, и затем, проанализировав, сопоставить их между собой. Таким образом можно выявить источники распространения вредоносных программ, потенциально опасные сайты. В связи с этим, для защиты от электронных рисков, огромные усилия сосредоточены на разработке и усовершенствовании облачных решений (таких как Kaspersky Security Network, Dr.WebCloud и т.д.), основанных на технологии анализа BigData.

D. Потребительские риски

В киберпростанстве широко распространено мошенничество, являющиеся одним из самых распространённых видов потребительских рисков. Чаще всего кибермошенничество осуществляется с помощью различных технических средств и разнообразных программ, следовательно, некоторые его виды могут быть отнесены к группе электронных рисков. Также потребительские риски плотно связаны с коммуникационными, поскольку установление более близкого контакта с жертвой с помощью личных сообщений- один из залогов успеха реализации преступной схемы. По словам руководителя аналитическим подразделением Microsoft по борьбе с преступлениями в сфере высоких технологий DigitalCrimesUnit (DCU) Брайана Хёрда, единственный способ противостоять росту киберпреступности — развивать технологии анализа Big Data [7].

Мошенничество посредством Интернета принимает чудовищные масштабы, зачастую обкрадывают целые страны, не говоря уже об отдельных пользователях, и тем более доверчивых детях. Противостоять кибермошенничеству могут только столь же масштабные системы, основой которых является технология Big Data, т.к. используя очередную схему, мошенники оставляют цифровые следы. И если взятые по отдельности эти изменения ничтожны, то в совокупности, при помощи Big Data, можно выявить характерную структуру преступной деятельности.

ЗАКЛЮЧЕНИЕ

Таким образом, полное обеспечение безопасности детей в Интернете посредством ныне существующих методик анализа Big Data вполне реально. Эффективным

решением для реализации является внедрение известных на сегодняшний день алгоритмов в данный процесс и адаптация их для борьбы с Интернет - рисками. Для каждой группы рисков следуют по отдельности разработать самостоятельную схему защиты, применить ее исключительно на уровне определенной группы рисков, и только после успешного завершения данного этапа, использовать Big Data комплексно для защиты от всех рисков.

ЛИТЕРАТУРА

[1] Д. Райдер, Революционный потенциал больших данных, 2015, http://22century.ru/docs/big-data.

[2] İngate, 75 фактов об интернет-маркетинге, 2012, http://www.slideshare.net/zhomart/75-ingate, стр.7.

[3] Г. Солдатова, В. Серегина, П. Волкова, «Неотложка» в киберпространстве // Журнал Дети в информационном обществе, 2011, стр.56-62.

[4] М.А. Назаренко Технологии и методы анализа больших данных // Международный журнал экспериментального образования, 2015, № 11, стр. 40-41.

[5] Новое Поколение, Федеральная программа безопасного детского интернета «Гогуль», 2009, http://gogul.tv/press/5/.

[6] Google, Политика конфиденциальности, 2015, https://www.google.com/intl/ru/policies/privacy/

[7] А. Васильков. Microsoft :”большие данные “ключ к борьбе с преступностью, 2014, http://www.computerra.ru/96630/big-data-is-key-technology-for-digital-crime-investigations/.

Seksiya # 3
Big Data Tətbiqləri

Böyük Verilənlərin İnsan Resurslarının İdarə Olunması Məsələlərinin Həllində Tətbiqi, İmkanları, Problemləri

Məsumə Məmmədova[1], Zərifə Cəbrayılova[2]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1,2]*depart15@iit.ab.az*

***Xülasə* — Məqalə böyük verilənlərin (BV) insan resurslarının idarə olunması (İRİO) sferasına inteqrasiyası məsələlərinə həsr edilmişdir. İRİO məsələlərinin həllində, düzgün, daha obyektiv qərarlar qəbulunda BV-nin istifadəsinin aktuallığı göstərilmişdir. BV və İRİO təmasında olan bir sıra tətbiqlər, bu təmasın yaratdığı imkanlar, bu imkanlardan yararlanmaq istiqamətində vəziyyət və problemlər şərh edilmişdir.**

Açar sözlər — böyük verilənlər; insan resurslarının idarə olunması; qərarların qəbulu; işə cəlb etmə məsələsi.

I. GİRİŞ

Hazırda biz 100% verilənlər əsrində yaşayırıq. 1980-ci ildən başlayaraq rəqəmsal saxlanılan informasiya təqribən hər 40 aydan bir iki dəfə artır [1]. 2012-ci ildən bəri hər gündə 2.5 ekzobayt (2.5×10^{18}) informasiya hasil olunur [2]. Lakin həndəsi silsilə üzrə artan verilənlərin 90% nəzərə alınmır, ya da nəzərə alınsa da təhlil olunmur [3].

Cəmiyyətin bütün səviyyələrində sürətli verilənlər massivinin formalaşması prosesi gedir və bu “qorxulu böyük verilənlər” (BV) əsrinin gəldiyini bildirir [4]. BV elə verilənlər toplusudur ki, onların emalı ənənəvi verilənlər bazasının toplanması, saxlanması, idarəedilməsi və analizinin instrunental vasitələrinin imkanlarından kənardadır.

Bu anlamda nəhəng verilənləri ənənəvi sistemlərin vasitəsi ilə emal etmək çətindir və odur ki, bu verilənlər arasında korrelyasiya qurmaqla əhəmiyyətli informasiyalar əldə etmək üçün onlar analiz olunur və işlənilir.

[5]-də BV-nin istifadəsinin əhəmiyyətini qeyd edən müəlliflər BV analitikasının (BDA) qərar qəbulunu yaxşılaşdırdığını önə çəkir, BDA-dan istifadə etməklə, demək olar ki, bütün sahələrdə gələcəyin proqnozlaşdırılmasını reallaşdırmaq mümkün olduğunu iddia edirlər. Bu məqsədlə artıq *IBM, HP, Cape Gemini, SAP, Accenture, Oracle Corporation, Microsoft* kimi nəhəng kompaniyalar BV-nin realizasiyasına başladıqlarını bildirirlərər.

Odur ki, BV-nin tətbiqi tendensiyası bütün təşkilatlarda idarəetmə məsələlərinə, o cümlədən İRİO-ya da inteqrasiya olunur. Belə ki, İRİO-nun planlaşdırma, təşkil etmə, nəzarət, yönləndirmə, işə cəlb etmə, seçilmə və yerləşdirmə, əməyin qiymətləndirilməsi, təhsilləndirmə (ixtisasartırma) və s. kimi müxtəlif məsələləri özündə birləşdirdiyini nəzərə alsaq, İRİO sistemində toplanan məlumatın həcmi haqqında təsəvvür yaratmaq olar. Lakin bu qədər informasiyaya malik olduğu bir məqamda belə ənənəvi İRİO üzrə idarəetmə sistemləri insan resurslarının gələcək inkişafı, işçilərin karyera yüksəlişi haqqında informasiya verməyə çətinlik çəkir. Bu baxımdan BV fəlsəfəsinin İRİO-da istifadəsi müvafiq idarəetmə məsələlərinin həllində daha dəqiq və obyektiv qərarlar qəbuluna dəstək ola bilərlər [6].

II. BV-NİN İRİO MƏSƏLƏLƏRİNİN HƏLLİNƏ İNTEQRASİYASI

Əməkdaşlar haqqında informasiyanın həcmi zaman keçdikcə və müəssisə böyüdükcə kəskin artır. Digər tərəfdən əvvəlki əməkdaşlar haqqında köhnə məlumatlar müxtəlif səbəblər üçün lazım ola bilər. Hər bir işçinin fərdi və təşkilat səviyyəsində olan bacarıqları, davranışı və kompetensiyaları onun haqqında informasiyanın həcmini günbəgün artırır. Bütün bu verilənlərin saxlanılması, onlara əlyetərliyin təmin olunması və onlardan rahat və tez bir zamanda istifadənin təmin olunması məsələləri düzgün təşkil edilməzsə bu verilənlər yararsız, əhəmiyyətsiz ola bilərlər. Digər tərəfdən, hazırda insan resursları ilə əlaqəli olan və müxtəlif mənbələrdən əldə olunan böyük həcmli verilənlər istifadə olunmaq üçün əlyetər olmuşdur. Belə mənbəələrə ilk növbədə əməkdaşla bağlı müxtəlif səpkili rəyləri formalaşdıran, verilənləri əldə etməyə imkan verən veb səhifələri; insanın özü haqqında sosial şəbəkələrdə yazdığı məlumatları; insanların yaratdıqları elektron məktublar, sənədlər, icmallar, qeydləri; əməkdaşların davranışını izləyən sensorlar, video kameralar, texniki qurğuların göstəricilərini və s. aid etmək olar. Hazırda kadr məsələləninin həllində qərarların qəbulu təcrübə və proqnozlardan daha çox, məhz bu verilənlərin analizinə əsaslanır. Odur ki, İRİO-nun bütün mərhələlərində böyük verilənlərin analitikası yüksək keyfiyyətin və effektiv qənaətcil nəticənin əldə olunması üçün vacibdir. İRİO-da BVA qismən strukturlaşmış və strukturlaşdırılmamış verilənləri analiz etməklə ixtiyari təşkilatda qərarların qəbulu və təşkilatın inkişafı üçün lazım olan verilənlər, əhəmiyyətli informasiyalar arasında korrelyasiya əlaqələrini müəyyənləşdirir. Bu verilənlər hər biri ayrılqda heç bir məna verməyən simvollar, ədədlər və s. toplusu ola bilər, lakin onların interpretasiyası gələcəyin proqnozlaşdırılmasında böyük əhəmiyyətə malik ola bilər, İRİO ilə bağlı müdrik qərarlar verilməsində istifadə oluna bilər. Belə ki, İRİO-da BV-nin istifadəsi təşkilatda inkişafı, rəqabətə

dözümlüyü, sürəti, məmnuniyyəti (əməkdaş və rəhbərlik arasında şəffaflıq) təmin edir [3, 6].

İRİO və BV-in təması aktualdır, bu təmas təşkilatın yüksəlişinə və gizli potensial imkanlarını reallaşdırmağa imkan verən strategiya qurması üçün lazımdır. Lakin heç də bütün təşkilatlarda İRİO sistemləri belə təmasa hazır deyillər, baxmayaraq ki, onlarda bir sıra idarəetmə proseslərinin (biznes proseslərin) avtomatlaşdırılması artıq neçə onilliklərdir ki, həyata keçirilir. Bu onunla bağlıdır ki, çox az qismdə təşkilatlarda lazımi verilənlərin toplanmasını təmin edən *ERP (enterprises resource planning)* sistemləri istismar olunur. Bu sistemlər, yarandığı ilk dövrlərdə yalnız kadrların uçotu və əmək haqqının hesablanması üçün nəzərdə tutulurdularsa, hazırda kadr menecmentinin bütün elementlərini: kadrların seçilməsi (yığılması), planlaşdırılması, təlimi, qiymətləndirilməsi, yerləşdirilməsi və yenidən yerləşdirilməsi, motivasiyası və s. dəstəkləyir (məsələn SAP (Almaniyada 1972-ci ildə yaradılıb, illik dövriyə 14 mlyd-dan çox olub), Qartner (ABŞ), BAAN (Niderlandda 2003-cü ildə yaradılıb), Scala (İsveçdə 1978-ci ildə yaradılıb), Navision (1983-cü ildə Danimarkada yaradılıb, hazırda Norveç, Almaniya, İspaniya, İngiltərə, ABŞ-da istifadə olunur) və s.). İRİO-da BV-yə gedən yol məhz bu sistemlərin reallaşdırılmasından keçir [7].

Bu gün İRİO-da kadrların planlaşdırılması, yerləşdirilməsi, yenidən yerləşdirilməsi və s. üzrə Oracle, SAP, Workday, ADP, Ultimate, Infor (Lawson) kimi provayderlər (core H*R-provayderlər*) və işə cəlb etmədə (*talantların idəarə olunması (TM)) Cornerstone On Demand* (son dövrdə 61% artmışdır), *Silkroad, SumTotal, Lumesse, Halogen, PeopleFluent, Cабa, Kenexa* (IBM tərəfindən yaradılmışdır) kimi proayderlər buludlarda (claudlarda) öz həllərini təklif edirlər [8]. Bu həllər korporasiyanın İRİO üzrə menecerlərinə bir-iki ildən bir modernləşmə sistemindəki problemlərin qarşısını almağa imkan verir. Aparılan tədqiqatlara əsasən müəyyənləşdirilmişdir ki, 13% təşkilatlar vahid HR (*core HR)* sistemindən istifadə edirlər, orta hesabla kompaniyalar 3-4 İRİO sistem əlavələrinə (məsələn, təlim (istisasartırma) sistemi, işə cəlb etmə sistemi (TM), əmək resurslarının idarə olunması sisteminə (WFM)) malikdir [8]. *Aberdeen Group* tərəfindən 2011-ci ilin dekabrında İRİO-da BV-nin tətbiqini önəmli sayan 126-dan çox təşkilatda araşdırma aparılmışdı. Araşdırma əsasında belə qənaətə gəlmişlər ki, core *HR, TM, WFM* verilənlər sistemlərindən istifadə edən şirkətlər öz nəticələrindən daha razıdırlar [3].

Hazırda İRİO-da BV əsasən iki istiqamətdə – işə cəlb etmə və personalın seçilməsi məsələlərinin həllində istifadə olunur.

III. BV-NİN NAMİZƏDLƏRİN İŞƏ CƏLB EDİLMƏSİNDƏ İSTİFADƏSİ

Müasir dünyada hər bir təşkilatın rəqabətqabiliyyəti onun malik olduğu insan resurslarının biliyi, bacarıqları ilə təyin olunur və odur ki, daha bacarıqlı insanların təşkilata cəlb edilməsi insan resursları şöbəsinin əsas məsələsidir. Bu məsələnin ənənəvi yolla həlli aşağıdakı mərhələlərlə müşayiət olunur: birincisi, rəhbər vakansiyanın olması haqqında məlumat verir; ikincisi, bu məlumat korporativ portalda yayımlanır; üçüncüsü, maraqlananlar özləri barədə məlumatı göndərirlər; dördüncüsü, kadrlar üzrə mütəxəssis onlara uyğun olan namizədi *müsahibəyə* dəvət edir və müvafiq namizəd seçilənə kimi bu iş davam etdirilir. Seçim zamanı müsahibə verənin *təhsili, cinsi, yaşı, ixtisası, iş təcrübəsi* önəmli rol oynayır. Lakin reallıq göstərir ki, bəzən nəticələr aldadıcı olur. Belə ki, qısa müsahibə ərzində namizəd haqqında tam əhatəli məlumat almaq mümkün olmur. Onlar yalnız namizədin verdiyi məlumatlara istinad edirlər. Belə birtərəfli və bəzən də yanlış informasiyanın verilməsi nəticədən yayınmaya səbəb olur.

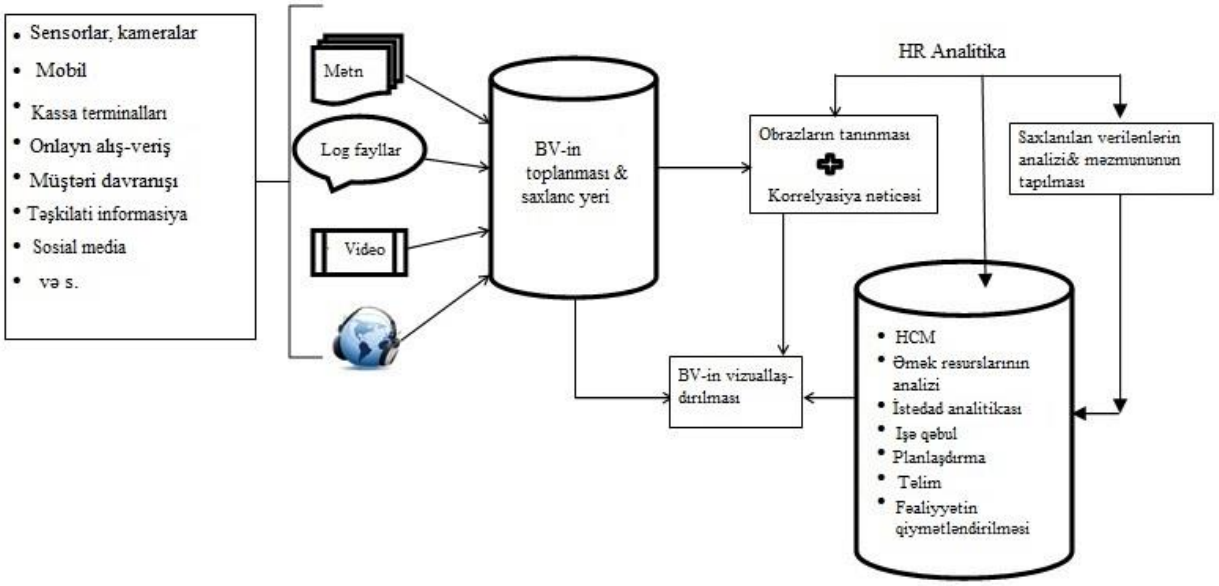
Google şirkəti müsahibə mərhələsinin uğurlu nəticəsinin gələcək effektiv fəaliyyətlə bağlılığının olmadığını iddia edərək, bu mərhələdən imtina etmişdir [7]. Bu sahədə BV-nin istifadəsi bu məqamları aradan qaldırmağa imkan verir. İlk növbədə BV müəssisələrə işçilərin cəlb edilməsi üçün daha geniş platformanı təmin edir ki, bu da İnternetdir. Statisikaya görə Çin Xalq Respublikasında müəssisələrin üçdə iki hissəsindən çoxu işçilərin işə cəlb olunmasında *onlayn* işə cəlb etmədən istifadə edir [9]. Kompaniya işə cəlb etməni sosial şəbəkələrə inteqrasiya etdirir, işə cəlb etmədə BVA üçün önəmli olan ilkin məlumatları (məsələn, namizədlərin *CV* məlumatları) toplayır, bununla da cəlb edilənlərin "böyük verilənlərin"in analizi üçün bünövrəni formalaşdırır. Bundan əlavə, hətta vakansiya olmadığı halda belə kompaniyalar namizədlər haqqında məlumatları mütəmadi olaraq toplayır. Sosial şəbəkələrdən toplanmış informasiyalar namizədlər haqqında *daha ətraflı məlumatın əldə olunmasına* yardım edir, məsələn, onun *videgörüntülərini, həyat tərzini, sosial münasibətlərini, bacarıqlarını* və s. əldə etməklə namizədin daha aydın və dəqiq "insan-portretini" formalaşdırmağa imkan verir [10]. BVA-nın işə cəlb etmədə tətbiqinin uğurlu nəticələrindən biri e**Q*quest-x*** BVA proqramı ilə reallaşmışdır. Belə ki, dünyanın 1000-dən çox karyera saytı, 1,5 milyard işçi qrupu, 250 milyondan çox vakansiya təqdim edən sosial media mənbəyindən ibarət olan BV və ***eQquest-x*** böyük verilənlərin analizi proqramı kompaniyalara iş yerlərini planlaşdırmağa və onun gələcək inkişafı (uğuru) ilə bağlı proqnozlar verməyə imkan vermişdir. İş yerini təqdim etməklə və BVA-ya əsaslanmaqla əməkdaşın işini xarakterizə edən müxtəlif verilənləri analiz etmək olar, nəticədə kompaniya hər pozisiyaya (iş yerinə) uyğun namizədin tapılmasını daha tez və az xərclə yerinə yetirə bilir [11]. Kompaniyaların 60%-dən çoxu BV və İRİO şöbəsində toplanmış verilənlər (HR) əsaslı analitik instumentariyaları tədqiq edir və bu da ənənəvi HR fəaliyyətini inqilabi sürətdə dəyişir (şəkil 1).

Jupiter Networks İRİO departamentində verilənlərin intellektual əldə olunması üçün BV-yə əsaslanan şəbəkə infrastrukturunu inkişaf etdirib. Kompaniya potensial işçini araşdırmaq üçün *Linkedin* sosial şəbəkəsindən istifadə edir [3]. Beləliklə, kompaniya hər zaman potensial perspektivli effektiv magistr və doktoranları, mühəndisləri işə cəlb etməyə nail olur.

IV. BV-NİN ƏMƏKDAŞLARIN TƏHSİLLƏNMƏSİNDƏ (İXTİSASARTIRMA) TƏTBİQİ

Müasir İRİO konsepsiaysına əsasən personal hər bir təşkilatın dayanıqlı inkişafının əsas hissəsidir, işçilərin bilik və bacarıq səviyyələrinin daima yüksəldilməsi onların iş məhsuldarlığının artmasına imkan verir [12]. Ənənəvi

ixtisasartırma adətən kompaniya tərəfindən təşkil olunur. Kompaniya professionalları cəlb edir, işdən sonra və ya evdə işçilərin hazırlığını planlaşdırır. Hansı yolun seçilməsindən asılı olmayaraq, kompaniya bu işə çoxlu maddi, maliyyə və insan resursu ayrılmalı olur. Bu tipdə hazırlıq ənənəvi sinif tədrisini xatırladır ki, bu da öyrənənlərin müxtəlif tələblərinin



Şəkil 1. BV-nin İRİO məsələlərinin həllinə inteqrasiyası

effektiv ödənilməsini təmin etməyə bilər. BV erasının başlaması bu problemin də həllinə müsbət təsir göstərir.

BV kontekstində əlyetər və birgə istifadə üçün nəzərdə tutulmuş informasiyalardan yararlanmaq daha rahatdır və hər bir kəs ixtiyari vaxtda və istənilən anda şəbəkədən ona lazım olan informasiyanı asanlıqla əldə edə bilər. Bundan əlavə hazırda bir çox təşkilatlar *peşəkar şəbəkə ixtisasartırma kurslarının* təşkilini inkişaf etdirir. Hər bir kompaniya konkret vəziyyətinə uyğun gələn belə şəbəkəni özündə sazlaya bilər və bu təcrübə artıq əksər kompaniyalarda formalaşmaqdadır. Belə proqram təminatı hər bir əməkdaşın davranışına uyğun verilənləri qeyd edə bilər, onlar da ixtisas artırmaq üçün öz tələblərinə uyğun olaraq sistemdən *onlayn* istifadə etməklə yanaşı, həm də özünə uyğun gələn tədris formasını seçə bilər. Belə ki, işçi öz təhsillənməsini daha məqsədyönlü təşkil edə bilər və tədrisin keyfiyyətini artıra bilər. İşçi hər anda *onlayn* test ala bilər, bu da effektiv təhsil almağa imkan yaradar və təhsilin effektivliyini artırar. Belə yanaşma bir qədərdən sonra tədrisin göstəricilərinə görə fərdin təhsil səviyyəsini xüsusi proqram vasitəsilə proqnozlaşdırmağa imkan verir [13]. Bundan əlavə belə yanaşma menecerlərə personalın kurslarda yeni bacarıqlara yiyələnməsini idarə etməsinə və işçilərin təhsil prosesinə nəzarət etməsinə imkan verir.

V. BV-NİN ƏMƏKDAŞLARIN FƏALİYYƏTİNİN QİYMƏTLƏNDİRİLMƏSİNDƏ TƏTBİQİ

İRİO-nun müasir konsepsiyasına əsasən əməkdaşların qiymətləndirməsi çox önəmli bir məqamdır. Hal-hazırda personalın qiymətləndirilməsində ekspert qiymətləndirilməsi, test yanaşması və kompleks qiymətləndirmə formalarına böyük yer verilir. Lakin bu metodlar çox subyektivdir. Odur ki, ənənəvi təcrübəni öyrənərək tədqiqatçılar BV texnologiyasından istifadə etməklə personalın fəaliyyətinin daha effektiv qiymətləndirilməsi və onların klassifikasiyası kimi bir sıra problemləri araşdırırlar [14]. Bu tədqiqatların kontekstindən çıxış edərək qiymətləndirmə metodlarında BV-nin istifadəsi analiz olunur. Hazırda BV-nin tətbiqi ilə kompaniyalarda nəhəng sistemlər qurmaqla hər hansı bir işçinin cari vəziyyəti ilə onun kompetensiya “portreti” arasındakı fərqi təyin etmək olar [15]. Bu fərlilik texniki təcrübəyə, şəxsi göstəricilərə, hətta psixoloji amillərə əsaslana bilər. Belə yanaşma kompetensiya modelinin qurulmasında bir yenilik olmaqla yanaşı işçilərin işə qəbulunda standart kimi qəbul oluna bilər. Bu əsasda nəzərdə tutulur ki, BV İRİO-da əməkdaşların qiymətləndirilməsində müdrik qərarlar qəbulunu təmin edəcəkdir.

VI. BV-NİN İŞÇİLƏRİN MOTİVASİYASINDA TƏTBİQİ

Hər bir kompaniyanın əmək haqqı sistemi onun əməkdaşları cəlb etməsində ən mühüm göstəricidir, lakin müəssisə rəhbərləri üçün də işçinin fəaliyyətinə görə məvacibin ödənilməsi onun motivasiyasını şərtləndirən vacib məsələdir. Əslində reallıq belədir ki, əmək haqqı sistemində bütün kompaniyalar problem qarşısında qalır [16].

Ənənəvi fəaliyyət sistemi əksər vaxt daha çox keyfiyyət və az qismdə kəmiyyət ifadələrində əks olunur, ödənişin səviyyəsi heç də fəaliyyətin nəticəsi ilə üst-üstə düşmür. Əmək haqqı əməkdaşların fəaliyyətinin (məhsuldarlığının) yüksək və aşağı səviyyədə olduğunu əks etdirmir və bu da əkdaşların məsuliyyət hissini azaldır. İRİO üzrə menecerlər üçün də gündəlik fəaliyyətin qiymətləndirilməsini aparmaq çox çətin məsələdir. Lakin BV-nin təsiri ilə kompaniya hər bir işçinin gündəlik iş yükünü, onun gördüyü işin konkret mahiyyətini və

işçinin məqsədinə nə qədər nail olduğunu qeyd edə bilər, sonra isə bu verilənlərin hesablama buludlarında emalını, analizini istifadə edə bilər [17]. Nəhayət, keyfiyyət standartlarına uyğun olaraq əmək haqqı avtomatik ödənilə bilər. Bu əməliyyatların reallaşması işin effektivliyini artırmaqla yanaşı, həm də insan kapitalına qoyulan biznes-investisiyanı azalda bilər. Bu məqsədlə bir sıra kompaniyalarda iş yerlərində əməkdaşların davranışını izləyən video qurğular quraşdırılmasına başlanılmışdır [18].

VII. BV-NİN İRİO MƏSƏLƏLƏRİNİN HƏLLİNDƏ TƏTBİQİ PROBLEMLƏRİ

Kompaniyaların İRİO məsələlərinin həllində BV-nin yaratdığı imkanlardan istifadə etməsi üçün aşağıdakı problemlər həll edilməlidir [19, 20]:

- **Kadr problemi** – kompaniyaların təkcə BV-yə malik olması kifayət deyil, kompaniyalar həmin verilənlərdən necə, nə vaxt və harada daha yaxşı üsulla, metodla istifadə etməyi bacaran liderlər qrupuna, kadrlara malik olmalıdır. Bu kadrlar müxtəli texniki bacarıqlara malik olan kateqoriyalara (*BV yaradıcısı, BV lahihələndiricisi, BV analitikləri, BV administratoru, BV layihə meneceri, BV dizayneri, elmi verilənlər üzrə mütəxəssis*) bölünürlər [21, 22].

SAS-ın hesabatına əsasən 2008-ci ildən 2013-cü ilə qədər BV üzrə işçilərin sayı orta hesabla 10 dəfə artmışdır. BV işçiləri ildə 55000 funt əmək haqqı qazanırlar, bu da İT-mütəxəssislərin maaşından 2% çoxdur. *SAS*-ın proqnozuna görə 2013-2020-ci ilə qədər BV mütəxəssislərinə tələbat 160% artacaq, bunun nəticəsində BV-də iş yerlərinin sayı 346000 artacaq [21, 22].

- **Elmi verilənlərin qıtlığı** bəzi sektorlarda bir sıra məhdudiyyətlərə səbəb olur. Rəqəmsal sferadakı yüksəliş strukturlaşdırılmamış verilənlərin strukturlaşdırılmasına və analizinin aparılmasına imkan yaradır, onlar zəngin verilənlər mənbələrini müəyyən edir, bu mənbələr arasında əlaqə yaradır. Verilənlərin daima artdığı və problemlərin dəyişdiyi rəqabət mühitində elmi verilənlər (data science) qərar qəbul edən şəxsə artıq analizlərdən əl çəkərək konkret həllə keçməyə imkan yaradırlar.

- **BV-nin strukturlaşdırılmış verilənlərlə birgə istifadəsi.** Əslində bu gün BV onunla bağlı anlayış və metodların işlənilmə dövrünü yaşayır, bu metodlar hələ mükəmməllikdən uzaqdır [7]. Odur ki, BV-nin hələ də ənənəvi strukturlaşmış verilənlərlə birgə istifadəsini qeyri-mümkün hesab edənlər var. Çünki hələ tam olaraq belə imkanı yaradan mükəmməl metod yoxdur, müəyyən mənada münasib metodlar işlənilir.

- **Əməkdaşların fərdi məlumatlarının təhlükəsizliyinin təmini.** BV-nin işlənilməsi və tətbiqi insanlara və müəssisələrə uğurlar gətirdiyi bir vaxtda şəxsi həyat və komersiya sirlərinin toxunulmazlığı ilə bağlı təhlükələr yaradır. *IDC* statistikasına görə 2010-cu ildə verilənlərin yalnız üçdə bir hissəsindən az qismi mühafizə olunurdusa, 2020-ci ildə bu beşdə iki hissəyə çatacaqdır [23, 24].

NƏTİCƏ

Bu gün BVA – idarəetmədə qərarların qəbulu və proqnozlaşdırmanın daha qabaqcıl və effektiv üsul hesab edilir. Onun İRİO sistemində toplanmış verilənlərlə birgə istifadəsindən əməkdaş haqqında bir çox yenilikləri bilmək olar. Əldə olunmuş nəticələr əsasında isə İRİO üzrə mütəxəssislər bu və ya digər biznes transformasiyaları daha yaxşı hiss edə bilərlər. Bu isə öz növbəsində düzgün qərarlar qəbul edilməsinə kömək edər [25]. BVA HR-menecerlərə:

- maliyyə xərclərini və zaman itkilərini minimallaşdırmağa;
- personalın idarə olunması və inkişafında yeni təşəbbüslərin həyata keçirilməsi risklərini azaltmağa;
- təşkilatın işinin vahid texnoloji platformada inkişaf etdirmək və dəstəkləməklə bütün təşkilatın işini daha yaxşı pronozlaşdırmağa və planlaşdırmağa imkan verir.

Bu baxımdan İRİO və BV-nin təmasının təmin olunması üçün verilənlərin hazırlanması aşağıdakı mərhələləri tələb edir:

- **dağınıq, pərakəndə və uzlaşmayan verilənlərdən uzaq olmaq lazımdır.** Bunun üçün ilk addım identifikasiyadır, yəni İRİO üzrə konkret məsələnin həlli üçün hansı verilənlərin lazım olması müəyyənləşdirilməlidir. Əksər vaxt İRİO şöbələri ənənəvi təcrübəyə uyğun olaraq verilənləri müxtəlif formatlarda saxlayırlar. BV-yə doğru miqrasiya İRİO-ya aid olan verilənləri müxtəlif informasiya sistemlərində identifikasiya etməyə və onların inteqrasiyası üçün strategiyanı müəyyənləşdirməyə imkan yaradır;
- **Vahid interfeysin istifadəsi**. Təşkilat daxilində verilənlərin necə toplandığını bilmək – BV-nin emalının uğurlu strategiyasının qurulmasında fundamental vacib məqamdır. Sadəcə olaraq verilənlərin avtomatik olaraq yığılması və bir mərkəzdən idarə olunsı təmin olunmalıdır;
- **Sadədən başlamalı**, yəni lazım olan verilənlərin tipini və onların toplanması üsulunu bilən kimi hansı analiz aparılacağını həll etmək lazımdır. Effektivliyin və sadəliyin dəstəklənməsi üçün sadə qarşılıqlı əlaqələrdən başlamaq lazımdır ki, bu da BV-nin analizinin necə aparılacağı və onun köməyilə hansı nəticələrin əldə oluna biləcəyini başa düşməyə kömək edəcəkdir. Daha sonra isə analizin mürəkkəbliyini artırmaq, İRİO-da proqnoz analitikanı dərinləşdirmək olar.

Belə yanaşma kompaniyanın məqsədlərinə cavab verən BV əsaslı İRİO-strategiyanı reallaşdırmağa imkan verəcəkdir.

ƏDƏBİYYAT

[1] Ingram Microdata center report on “4 Big Data Use Cases in the Manufacturing Industry”, www.ingrammicroadvisor.com/big-data/4-big-data-use-cases-in-the-manufacturingindustry

[2] www.cio.com/article/2386531/healthcare/healthcare-6-big-data-analytics-use-cases-forhealthcare-it.html

[3] Big Data in Human Resource Management - Developing Research Context file:///C:/Users/HP/Downloads/Big%20Data%20in%20Human%20Resource%20Management%20.pdf

[4] L. S. Valle, E. Lesser, R.Shockley, M. S. Hopkins and N. Kruschwitz, “Big Data, Analytics and the Path from Insights to Value”, *MIT Sloan Management Review*, 2011, vol. **52**, no. 2, pp. 74-82.

[5] C. W. Scott, T. A. H. Will , “Three Business and Societal Cases for Big Data: Which of the Three Is True?” *IEEE Engineering Management Review*, 2014, vol. 42, no. 3, pp.65-72.

[6] S. Zang, M. Ye, “Human Resource Management in the Era of Big Data” *Journal of Human Resource and Sustainability Studies*, 2015, 3, pp. 41-45. file:///C:/Users/HP/Downloads/JHRSS_2015031714271747%20(1).pdf

[7] “Большие данные BigData для HR. Как увидеть личность за цифрой?” *Журнал Компетенции* http://hr-media.ru/bolshie-dannye-bigdata-dlya-hr-kak-uvidet-lichnost-za-tsifroj/-dekabr-yanvar2015

[8] “7 Reasons HR Technology Is So Hot Today” www.forbes.com/sites/joshbersin/2013/05/31/7-reasons-hr-technology-is-so-hot-today/

[9] S. L. Huang and J. Xiang, () “Big Data” Light up “Wisdom” of Human Resources Management System. *Science & Technology for Chinese Mass Media*, 2013, **12**, pp.76–78.

[10] X. H. Dong, A. Ying. and J. G. Guo, “Research on the Application of the Big Data Technology in the Network Recruitment”. Human Resource Development of China, 2014, v.18, pp. 37-41.

[11] www.equest.com/solutions/big-data-for-hr/

[12] H. Tian, “The Concept, Thinking of Human Resources Management Based on the Era of Big Data” . *Human Resources*, 2014, no. **20**, pp.162-163.

[13] Y. He, Management Innovation in “the Era of Big Data”. *Human Resources*, 2013, no.**10**, pp.62-63.

[14] G. H. Zhou, “Thinking on Human Resources Management Based on the Value of Big Data”, *Science Mosaic*, 2013, no.**12**, pp.104-107.

[15] D. H. Wang, “Evidence-Based Human Resource Management in the Era of Big Data”, *Chinese and Foreign Entrepreneurs*, 2013, no.**8**, pp.159-168.

[16] Z. Y. Wu, “Research on Model of Human Resources Management System Based on Cloud Computing”, *China Computer & Communication*, 2012, n.**8**, pp.142-143.

[17] B. Wei, “The Recruitment Management System Based on Cloud Computing Model”, *Science & Technology Information*, 2012, no. 19, pp.108-109.

[18] “The Big Data Opportunity for HR and Finance. Harvard Business Review” https://hbr.org/resources/pdfs/comm/workday/workday_report_oct.pdf

[19] “Capgemini Consulting: Big Data trends & challenges” www.consultancy.uk/news/1516/capgemini-consulting-big-data-trends-challenges

[20] “BIG DATA: A Management Revolution” https://hbr.org/2012/10/big-datathe-management-revolution/ar

[21] SAS report on “Big Data Analytics Anassessment of demand for labor and skills, 2012-2017”, 2013.

[22] SAS report on “Big Data Analytics Assessment of Demand for Labour and Skills 2013–2020, 2014.

[23] J. Zhao, “How to Make the Enterprise Human Resources Management Adapt to Conditions during the Age of the Big Data”, *Human Resource Development of China*, 2012, no.**4**, pp.13–16.

[24] X. F. Meng and X. .Ci, “Big Data Management: Concepts, Techniques and Challenges”, *Journal of Computer Research and Development*, 2013,n. **50**, pp.146–169.

[25] https://blogs.oracle.com/applications/entry/three_steps_to_get_big

Superkompüter Texnologiyaları: Mövcud Vəziyyəti və İnkişaf Perspektivləri

[1]Rəşid Ələkbərov, [2]Tural Mustafayev, [3]Məmmədrəsul Yaqubov,
[1,2,3]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*rashid@iit.ab.az,* [2]*tural.mustafayev@iit.ab.az,* [3]*mrasul.yagub@iit.ab.az*

***Xülasə* - Məqalədə süperkompüter texnologiyalarının hal-hazırkı vəziyyəti və onlarin müxtəlif göstəricilər üzrə paylanma dinamikası analiz edilmişdir. Yaxın perespektivdə superkompüterlərin inkişaf dinamikası araşdırılmışdır.**

Açar sözlər - süperkompüter, real məhsuldarlıq, nəzəri məhsuldarlıq, green computing, big data, flop/s, klaster, top500 superkompüter .

I. GİRİŞ

Ölkəmizdə İnformasiya Cəmiyyətinin qurulması dövlət siyasətinin əsas prioritetlərindən biri kimi qəbul edilmişdir. İnformasiya Cəmiyyətinin qurulmasının əsas vəzifələrinə bu cəmiyyətin hüquqi əsaslarının yaradılması, ölkənin iqtisadi, sosial və intellektual potensialının möhkəmləndirilməsi, müasir informasiya-kommunikasiya infrastrukturunun formalaşdırılması, informasiya təhlükəsizliyinin təmin edilməsi, qlobal informasiya fəzasına inteqrasiya və digər vacib məsələlər daxildir. Göstərilən məsələlərin həlli zamanı meydana çıxan mürəkkəb hesablamalar və böyük həcmli yaddaş tələb edən məlumatları daha sürətlə emal edib, istifadəçilərə çatdırmaq üçün yüksək hesablama məhsuldarlığına və böyük yaddaşa malik olan superkompüterlərdən geniş istifadə edirlər. Eyni zamanda müasir dövrdə elmin müxtəlif sahələrində: fiziki-kimyəvi proseslərin, nüvə reaksiyalarının, qlobal atmosfer proseslərinin, iqtisadiyyatın inkişafının real zaman kəsiyində modelləşdirilməsində, kriptoqrafiyada, geologiyada, yeni dərman növlərinin yaradılmasında və s. meydana çıxan böyük hesablama və yaddaş resursları tələb edən mürəkkəb məsələlərin həllində fərdi kompüterlərin hesablama gücü kifayət etmir [1]. Buna görə də həmin mürəkkəb məsələlərin həllində superkompüterlərdən geniş istifadə olunur.

II. SUPERKOMPÜTERLƏRDƏ HƏLL OLUNMASI NƏZƏRDƏ TUTULAN MƏSƏLƏLƏR

Elmin müxtəlif sahələrində superkompüterlərdə həll edilən mürəkkəb məsələlər aşağıdakılardır[2]:

* nəzəri və eksperimental fizika, yüksək enerjilər fizikası, kvant fizikası və s.
* kvant və molekulyar kimya sahəsində;
* böyük sistemlərin riyazi modelləşdirilməsi;
* kimya mühəndisliyi, yeni materialların yaradılması;
* maşınqayırma sənayesində mürəkkəb məmulatların layihələndirilməsi;
* təhlükəsizlik sisteminin işlənib hazirlanması üçün qəzanın nəticələrinin modelləşdirilməsi;
* ekologiya sahəsində;
* genetikada;
* yerdə, okeanda, atmosferdə baş verən proseslərin modelləşdirilməsi;
* astronomiya sahəsində;
* nəqliyyat sektorunda;
* yarımkeçiriçi gurğuların yaradılmasında;
* hərbi sahədə;
* və s.

Superkompüterlərin reytinq cədvəli tərtib olunarkən, onların hesablama məhsuldarlığı əsas amil kimi götürülür. Superkompüterin hesablama məhsuldarlığı dedikdə onun bir saniyədə nə qədər əməliyyat yerinə yetirdiyi başa düşülür. Superkompüterlərin məhsuldarlığı nəzəri və real məhsuldarlıqla qiymətləndirilir. Nəzəri məhsuldarlığı təyin etmək üçün sistemdə iştirak edən eyni tip prosessorlardan birinin məhsuldarlığını götürüb sistemdəki prosessorların sayına vurmaq lazımdır. Real məhsuldarlıq isə xüsusi test proqramların köməyi ilə təyin olunur [3]. Real məhsuldarlıq təxminən nəzəri məhsuldarlığın 70-90 %-ni təşkil edir. Superkompüterdə məhsuldarlığı təyin etmək üçün Flop/s-dan (Floating point operations per second) – bir saniyə müddətində sürüşkən vergüllü ədədlər üzərində yerinə yetirdiyi əməliyyatların sayından istifadə olunur.

Superkompüterlərin real məhsuldarlığını təyin etmək üçün xüsüsü test proqramlardan istifadə olunur. Test proqramların ən geniş yayılanı LİNPACK proqramıdır. Bu, N (1000) dəyişənli xətti cəbri tənliklər sisteminin həll edilməsi üçün tərtib edilmiş test proqramdır. Dünya üzrə superkompüterlərin reytinq cədvəlinin (TOP 500) tərtib edilməsində bu cür test proqramlardan geniş istifadə olunur.

III. SUPERKOMPÜTER TEXNOLOGİYALARININ MÖVCUD VƏZİYYƏTİNİN ANALİZİ

1976-cı ildə, amerikalı alim Seymur Krey tərəfindən, ilk superkompüter (“CRAY-1”) yaradılmışdır. Həmin superkompüter saniyədə 160 milyon əməliyyat yerinə yetirib, əməli yaddaşının həcmi 8 Mbayta bərabər olmaqla, qiyməti 8,8 milyon ABŞ dolları idi [4]. 1993-cü ildən başlayaraq Amerikanın Lourens Berkli Milli Laboratoriyasının kompüter sahəsindəki ekspertləri tərəfindən hər il (son illər ildə iki dəfə - iyun, noyabr aylarında) dünya üzrə ən güclü 500 superkompüterin reytinq cədvəli çap olunur.

Superkompüterlərin 2015-ci ilin iyun ayında çap olunan 45-ci reytinq cədvəli cədvəl 1-də göstərilmişdir [5].

Cədvəl

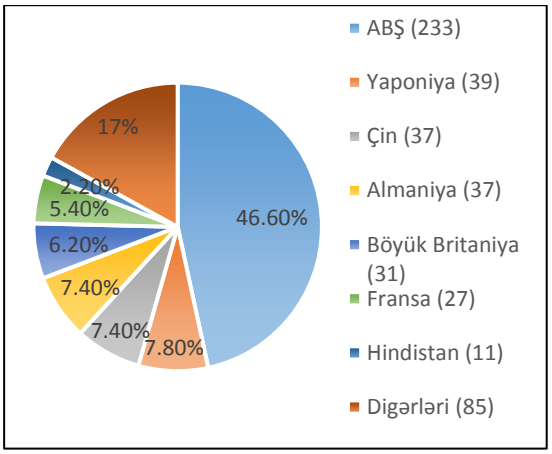
Sıra Nömrəsi	Təşkilat və dövlətin adı	Kompüterlərin adları / İstehsalçı	Nüvələrin sayı	Real məhsuldarlıq (TFlop/s)
1	National Super Computer Center (Guangzhou, Çin)	Tianhe-2 (MilkyWay-2) / Milli Müdafiə Texnologiyasl Universiteti	3 120 000	33 862.7
2	Oak Ridge National Laboratory (Oak Ridge, ABŞ)	Titan / Cray inc.	560 640	17 590
3	Department of Energy’s National Laboratory (California, ABŞ)	Sequoia / IBM	1 572 864	17 173.2
.....	.....	.....	.....	.....
100	Swiss National Supercomputing Centre (Isveçrə)	EPFL Blue Brain IV / İBM	65 536	715.6
.....	.....	.....	.....	.....
300	Amazon Web Services (ABŞ)	Amazon EC2 Cluster Compute Instances / noname	17 024	240.1
.....	.....	.....	.....	.....
500	E-Commerce (ABŞ)	xSeries x3650M3 / İBM	29 244	164.8

2015-ci ilin iyun ayında çap olunmuş statistikaya görə 1-ci yerdə Çin Xalq Respublikasının Guangjou şəhərindəki Milli Super Kompüter Mərkəzində yerləşən Tianhe-2 (Süd Yolu-2) superkompüteri dayanır. 3,120,000 nüvədən ibarət olan Tianhe-2 33,800 TFlop/s (10^{12} Flop/s) hesablama məhsuldarlığına malikdir. Qiyməti 290 milyyon ABŞ dollarıdır

On il müddətdə superkompüterlərin reytinq cədvəlini müqayisə etsək görərik ki, 2005-ci ildə birinci yerdə olan superkompüter cəmi 136 TFlop/s hesablama gücünə malik idi. 500-cü yerdə dayananın isə 1.2 TFlop/s məhsuldarlığı var idi. 2015-ci ilin reytinq cədvəlində birinci yerdə olan superkompüterin məhsuldarlığı (33 862 TFlop/s) 2005-ci ilin reytin cədvəlinin uyğun yerində yerləşən superkompüterin məhsuldarlığından (136 Tflops) 250 dəfə çoxdur. Qeyd etmiək lazımdır ki, on il əvvəl birinci yerdə yerləşən superkompüterin məhsuldarlığı cari reytinq cədvəlinin 500-cü yerində yerləşən superkompüterin məhsuldarlığından (164 TFlop/s) azdır.Bu isə superkompüterlərin məhsuldarlığlnın son illərdə sürətlə artmasının göstəricisidir.

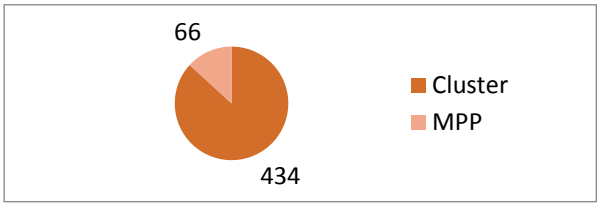
Son iki ildə ilk yeddi yerin sahibi isə ümumiyyətlə dəyişilməyib. Bu da onu göstərir ki, artıq superkompüterlərin hesablama gücünü artırmaqdan daha çox, onlardan daha səmərəli istifadə ön plana çəkilir. Hal-hazırda 68 superkompüterin məhsuldarlığı 1 Pflop/s-dan yuxarıdır.

Superkompüterlərin ölkələr üzrə paylanmasının analizi göstərir ki, reytinq cədvəlində iştirak edən superkompüterlərin 47% ABŞ dövlətinə məxsusdur (şəkil 1).



Şəkil 1 Superkompüterlərin ölkələr üzrə paylanması

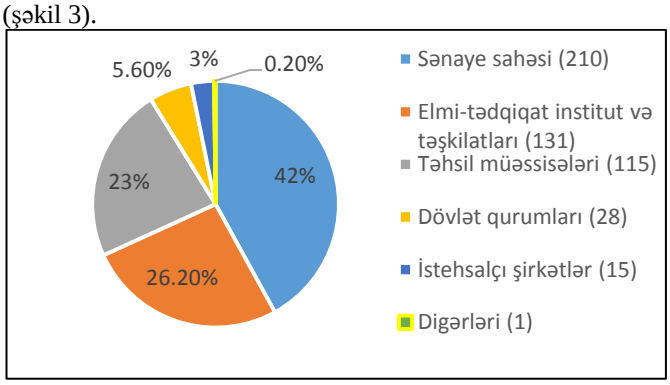
Son illərdə istehsal olunan superkompüterlərin arxitekturasında da geniş dəyişiklik baş vermişdir. Əvvəllər istehsal olunan superkompüterlər MPP (Massively Parallel Processors – kütləvi paralel prosessorlu) və SMP (Simmetric Multiprocessors – ümümi yaddaşlı simmetrik multiprosessorlu) arxitektura əsasında hazırlanırdı. Hal-hazırda istehsal olunan superkompüterlər daha şox klaster tipli arxitekturaya malikdir (şəkil 2). Klaster arxitekturalı superkompüterlər modul prinsipi əsasında hazırlanır, bu isə onlardan daha səmərəli istifadə etməyə imkan verir. Klaster arxitekturalı superkompüterlərin digər arxitekturalı superkompüterlərə nəzərən ucuz başa gəlməsi superkompüterlərin kutləvi istehsalına təkan vermişdir.



Şəkil 2. Superkompüterlərin arxitekturaya görə paylanması.

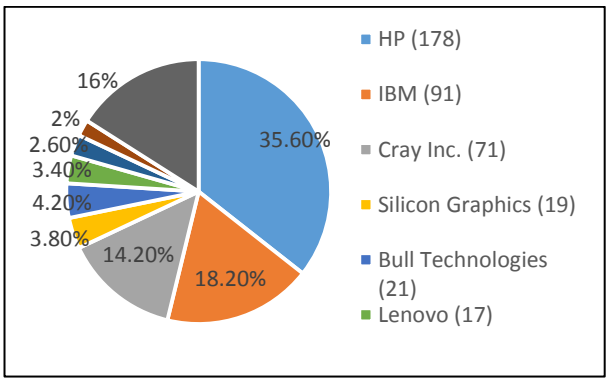
Aparılan tədqiqatlar göstərir ki, superkompüterlərin çox hissəsi (42%) sənaye sahəsində istifadə olunur. Eyni zamanda

26.2% elmi-tədqiqatlarda, 23% tədris işlərində, 5.6% dövlət layihələrinin, 3% isə istehsal müəssisələrində istifadə olunurn (şəkil 3).



Şəkil 3. Superkompüterlərin istifadə sahələrinə görə paylanması.

Reytin cədvəlində göstərilən 500 superkompüterdən 179-u HP, 90-nı İBM, 71-i Cray inc. şirkətlərinin avadanlıqlarından istifadə edərək qurulmuşdur. Sonrakı yerlərdə SGİ (29), Bull (22), Fujitsu (14), Dell (10) sirkətlərinin avadanlıqları əsasında qurulmuş superkompüterlər gəlir. Birinci yerdə olan Tianhe-2 superkompüteri NUDT (National University of Defence Technology) şirkəti tərəfindən qurulmuşdur. Superkompüterlərin istehsalçı şirkətlər üzrə paylanması (şəkil 4)-də göstərilmişdir.



Şəkil 4. Superkompüterlərin istehsalçı şirkətlər üzrə paylanması.

Qeyd edilən 500 superkompüterdən 488-inin əməliyyat sistemi Linux bazasındadır. Bu superkompüterlərin 246-sı okeanın o tayındadır (Şimali Amerika-240, Cənubi Amerika-6). Avropa regionunun payına 139 superkompüter düşür. Asiya 109 superkompüterlə 3-cü yerdədir. Avstraliyanın aktivində 6 superkompüter var.

Superkomputerlərin hesablama qovşaqları arasında əlaqəni təmin edən komunikasiya şəbəkəsi 51.8% (259) İnfiniband, 16.6% (83) 10G, 14% (70) İndividual, 12.8% (64) Gigabit Ethernet şəbəkə texnologiyaları əsasında quruşdırılmışdır. 500 superkompüterdən 433-ü (86.6%) İntel, 36 (7.2%) Power, AMD 23 (4.6%) AMD, 7 (1.4%) Spark firmalarının mikroprosessorları bazasında yığılmışdır.

IV. SUPERKOMPÜTER TEXNOLOGİYALARININ PERSPEKTİVLƏRİ

Böyük həcmli verilənlərin (Big Data) emal olunması və yadda saxlanmasında superkompüter texnologiyalarından geniş istifadə olunur. İBM şirkətinin ekspertlərinin apardığı araşdırmalara görə hər gün 15 Pbayt yeni verilənlər (elmi məqalələr, şəkillər, audio-video fayllar, sosial şəbəkə hesabatları və s.) hasil olunur. Dünyada yüksək məhsuldarlıga malik olan Tianhe – 2 (Çin) 12.4 Pbayt, Titan (ABŞ) isə 40 Pbayt yaddaş tutumuna malikdir. Utah Milli Təhlükəsizlik Aqentliyində (Utah, NASA-ABŞ) yerləşən superkompüterin yaddaş tutumu 1 Ybayt (yottabayt) – 10^{24} bayta bərabərdir. Qeyd edilən yaddaş tutumları həddən artıq böyükdür. Müqayisə üçün qeyd edək ki, bu günə kimi yazılmış bütün kitabları 400 Tbayt yaddaş tutumunda yerləşdirmək olar. Elmin müxtəlif sahələrindəki inkişaf bu sahədə emal olunacaq verilənlərin həcmini eksponensial şəkildə artırır. Bu isə gələcəkdə daha yüksək hesablama məhsuldarlığına və yaddaş tutumuna malik superkompüterlərdən istifadə edilməsinə gətirib çıxaracaq.

2020-ci ilədək istehsal olunacaq (Top 500-də 1-ci yerdə duracaq) superkompüterin məhsuldarlığının inkişaf dinamikası aşağıdakı kimi proqnozlaşdırılır (Cədvəl 2) [6].

Cədvəl 2

İllər	Real Məhsuldarlıq (PFlop/s)	Nəzəri Məhsuldarlıq (PFlop/s)
15.11.2015	84.7	137.1
15.11.2016	158.6	256.7
15.11.2017	296.8	480.6
15.11.2018	555.6	899.6
15.11.2019	1.040	1.684
15.11.2020	1.947	3.152

Yaxın illərdə yaradılacaq superkompüterlərin texniki imkanlarını nəzərdən keçirək. Hal-hazırda reytinq cədvəlində 2-ci yerdə dayanan “Titan” superkompüterinin yerləşdiyi “Oak Ridge National Laboratory – Yaponiya” elmi mərkəzində “Summit” adlanan yeni superkompüterin yaradılması istiqamətində işlər aparılır və 2017-ci ildə istifadəyə verilməsi nəzərdə tutulur. Yeni yaradılan superkompüterin hesablama gücü “Titan” superkompüterin hesablama gücündən 5 dəfə çox olmaqla 150 PFlop/s olması nəzərdə tutulur [7].

ABŞ-ın “Argonna National Laboratory United States” laboratoriyasında yerləşən, reytinq cədvəlində 5-ci yerdə dayanan 10 PFlop/s hesablama məhsuldarlığına malik olan “Mira” superkompüterinin əsasında yeni “Avrora” superkompüterinin yaradılması istiqamətində işlər aparılır. Yaradılan superkompüterin hesablama gücü 180 PFlop/s olacaq və 2019-cu ildə işə salınacaq.

Superkompüterlərin enerjidən istifadəsi həddindən artıq böyükdür. Məsələn: “Tianhe-2” -nin enerjiyə tələbatı 17.8 MW, “Mira” -nın enerji tələbatı 8.6 MW və s. : “Tianhe-2” superkompüterinin illik enerjiyə tələbatına çəkilən xərclər 24 milyyon dollardır. Qeyd edilən problemi aradan qaldırmaq üçün “Green Computing” texnologiyalarından son dövrlərdə geniş istifadə olunur [8,9]. Green Computing – ekoloji yönümlü informasiya-telekommunikasiya texnologiyasıdır. Bu texnologiya əsasında yaradılan kompüter məhsullarında təhlükəli materiallardan az istifadə edilməsi, enerji sərfiyyatının az olması, avadanlıqların istifadə müddətinin çox olması, ətraf mühitə ziyan vurmadan iqtisadi cəhətdən ucuz başa gələn utilizasiyanın edilməsi və s. tələblər ödənilməlidir. Bu texnologiyaların tətbiqi ilə əldə olunan müsbət imkanları nəzərdən keçirək. Bu texnologiyanın köməyi ilə “Oak Ridge National Laboratory”-də yerləşən “Titan” superkomputerinin əsasında yaradılan növbəti nəsil “Summit” superkompüteri ondan 1.21 dəfə artıq elektrik enerji sərf etməklə 5 dəfə çox nəzəri məhsuldarlığa sahib olacaq. Eyni vəziyyət “Argonne National Laboratory United States” laboratoriyasında yerləşən “Mira” və onun növbəti nəsli “Aurora”-ya da aiddir. Belə ki, “Mira”-ın nəzəri məhsuldarlığı 3 945.00 kW enerji sərf etməklə 10 PFlop/s-dirsə, “Aurora” təxminən 2.7 dəfə çox enerji sərf etməklə bu məhsuldarlığı 180 PFlop/s-yə (18 dəfə çox) qaldıracaq. Verilənlərin emalı mərkəzində superkompüterlərin yerləşdirilməsi və effektiv işləməsi üçün lazım olan maliyyənin böyük hissəsini bu superkompüterlərin soyutma sistemləri təşkil edir. Yaxın perspektivdə bu xərcləri azaltmaq üçün az xərc tələb edən (Free Cooling Systems) soyutma sistemlərinin tətbiqi nəzərdə tutulur. Qeyd edilən soyutma sistemləri əsasında yaradılan verilənlərin emalı mərkəzləri əsasən soyuq iqlim mühitində yerləşən regionlar və ölkələrdə qurulacaq. Az xərc tələb edən soyutma sistemləri verilənlərin emalı mərkəzində istifadə olunan enerji sərfiyyatını 40 %-a qədər azaldacaq [10].

İslandiyanın gələcəkdə data mərkəz sənayesində vacib rol oynayacağı gözlənilir. Artıq bir neçə böyük şirkət öz növbəti data mərkəzlərini İslandiyada qurmağı prespektivdə görür. İslandiyanın özü də bu sektoru inkişaf etdirmək üçün vacib addımlar atmağı planlaşdırır [11].

Azərbaycan Milli Elmlər Akademiyası (AMEA) İnformasiya Texnologiyaları İnstitutunun Data Mərkəzində yerləşən superkompüterin hesablama məhsuldarlığı 14.1 TFlop/s və yaddaşının həcmi 200 TBaytdır. Superkompüter AMEA-nın institut və təşkilatlarına elektron poçt, internet və hostinq xidməti göstərir. Bundan əlavə superkompüterdən akademiyanın institut və təşkilatlarında meydana çıxan böyük hesablama və yaddaş resursları tələb edən mürəkkəb məsələlərin həllində də istifadə edirlər. Azərbaycan Respublikasının Rabitə və Yüksək Texnologiyalar Nazirliyində Superkompüter mərkəzinin yaradılması sahəsində məqsədəyönlü işlər aparılır və 2015-ci ilin sonlarında işə salınması nəzərdə tutulur.

NƏTİCƏ

Məqalədə superkompüter texnologiyalarının hazırkı vəziyyəti təhlil olunmuş və onlarin müxtəlif göstəricilər üzrə paylanma dinamikası analiz edilmişdir. Yaxın perespektivdə superkompüterlərin inkişaf dinamikası araşdırılmışdır. Eyni zamanda, superkompüterlərin böyük həcmli verilənlərin (Big Data) emalı mərkəzlərində istifadə olunması təhlil edilmişdir. Superkompüterlərin enerji sərfiyyatını azaltmaq üçün yeni texlogiyalardan istifadə məsələlərinə baxılmışdır.

TƏŞƏKKÜRNAMƏ

Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişafı Fondunun maliyyə yardımı ilə yerinə yetirilmişdir – **Qrant № EİF-2014-9(24)-KETPL-14/02/1.**

ƏDƏBİYYAT

[1] Воеводин В.В., Воеводин Вл.В., “Параллельные вычисления”, Санкт-Петербург: БХВ - Петербург, 2002, 608 с.
[2] “Задачи для суперкомпьютеров”, http://parallel.ru /research/ apps.html
[3] http://parallel.ru
[4] Корнев В.В., “Вычислительные системы”, Москва: Гелиос АРВ, 2004, 512 с.
[5] http:// www.top.500.org
[6] С.М. Абрамов, Е.П. Лилитко, “Состояние и перспективы развития вычислительных систем сверхвысокой производительности”, http://paco2012.ipu.ru/procdngs/ P103.pdf
[7] “Ranking of supercomputers according to the LINPACK benchmark”, www.top.500.org
[8] http://www.theplatform.net/2015/08/04/future-systems-pitting-fewer-fat-nodes-against-many-skinny-ones/
[9] “Что такое зеленые вычисления или зеленые информационные технологии”, http://nature-time.ru/2014/07/zelenyie-vyichisleniya-ili-zelenyie-informatsionnyie-tehnologii/
[10] www.green500.org
[11] http://searchdatacenter.techtarget.com/video/Icelands-blueprints-for-a-data-center-industry

Elektron Elmin Elmi Verilənlər Problemləri Haqqında

Təhmasib Fətəliyev
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
depart3@iit.ab.az

Xülasə **– Məqalə elektron elm konsepsiyasının rellaşması çərçivəsində elmi verilənlərin emalı proseslərinin aktual məsələlərinə həsr olunmuşdur. Elmi fəaliyyət zamanı toplanan, saxlanılan və ötürülən elmi verilənlərin həcminin əhəmiyyətli dərəcədə artması göstərilmişdir. Elm sferasında mühüm Big data təşəbbüsləri tədqiq olunmuş və həmin sahədə tətbiq olunan texnologiyalar haqqında məlumat verilmişdir.**

Açar sözlər **–** ***e-elm, elmi verilənlər problemləri, AzScienceNet, Data Mərkəz, Big data təşəbbüsləri, qrid, cloud computing.***

I. GİRİŞ

Son illər ərzində informasiya texnologiyaları sahəsində əldə olunan nailiyyətlər elmin inkişafında əsas tendensiyalardan biri kimi elmi verilənlərin həcminin əhəmiyyətli dərəcədə artırmış və onların saxlanması və emalı problemini meydana gətirmişdir. Elmi fəaliyyət nəticəsində böyük həcmdə eksperimental verilənlər yaranmış, müxtəlif elmi nəşrlərin, görülən işlər haqqında hesabatların, layihələrin nəticələrinin və s. sayı artmışdır. Hazırda elmi fəaliyyətin təşkili üçün virtual elmi kollektivlərin formalaşdırılması məqsədilə ərazicə paylanmış mühitdə tərəfdaşların axtarışı, həmçinin yaradılmış texnologiyaların və elmi fəaliyyətin nəticələrinin istifadəsi və yayılması ayrıca bir problemə çevrilmişdir. Belə nəticəyə gəlmək olar ki, elmin gələcək inkişafı yalnız böyük həcmli informasiyanın emalı nəticəsində mümkün olacaqdır.

Verilənlərin elmdə rolu həmişə yüksək qiymətləndirilmişdir, lakin elmi verilənlərin qeydiyyatı, emalı və analizi “hesablama elminin” *(“computational science”)* – məsələlərin həlli üçün informatika və hesablama texnikasının ən son nailiyyətlərindən geniş istifadə edilən yanaşmanın inkişafı ilə xüsusi aktuallıq qazanmışdır. Nəticə olaraq, bu sahədə tıtbiq edilən elmi metodlar da nəzərə çarpacaq qədər dəyişir. Belə ki, son zamanlar təkcə əldə olunan nəticələr deyil, həmçinin sonrakı emal üçün nəzərdə tutulan və informasiya massivləri şəklində toplanan ilkin verilənlər də təqdim edilir. Nəticədə, bir tərəfdən çox böyük həcmli informasiyanın arxivləşdirilməsi problemi, digər tərəfdən isə informasiyanın axtarışı üzrə xidmətin təşkili, yayılması və istifadəçinin tələb olunan informasiya ilə təmin edilməsi zərurəti yaranır. Bunlarla yanaşı, toplanmış son dərəcədə böyük həcmli elmi verilənlərdən gizli biliklərin aşkarlanması da aktual və əhəmiyyətlidir.

Beləliklə, böyük həcmli verilənlər və onun mühüm tərkib hissələrindən olan elmi verilənlərin yuxarıda qeyd olunan problemlərinin həlli yeni bir elm sahəsinin – verilənlər haqqında elmin (*Data science*), həmçinin qısa bir zaman ərzində “*Big Data”, “Big Data analytics”, “Data science”, “Datalogy”, “Data scientist”, “Data-intensiv science”* kimi artıq populyar olan, yeni terminlərin yaranmasına gətirmişdir.

Elmi verilənlər problemlərinin həlli respublikada Azərbaycan Respublikası Rabitə və Yüksək Texnologiyalar Nazirliyinin dəstəyi ilə “Elektron Azərbaycan”-nın tərkib hissəsi kimi həyata keçirilən “e-elm” layihəsi çərçivəsində nəzərdə tutulmuşdur. E-elmin şəbəkə platforması olan AzScienceNet şəbəkəsinin və böyük yaddaş və hesablama resurslarına malik (yaddaş-200 Terabayt, hesablama məhsuldarlığı - 14 Tflops) Data Mərkəzin imkanları buna texniki dəstək verir [1].

Qeyd olunanlar araşdırılan məsələnin aktullağını bir daha təsdiq edir.

II. E-ELM BÖYÜK ELMİ VERİLƏNLƏRİN GENERATORU KİMİ

E-elm mürəkkəb bir sistem olub infrastruktur, verilənlərin generasiyası, toplanması, saxlanması, emalı, axtarışı, analizi, ötürülməsi, təqdim olunması və s. kimi tərkib hissələrə malikdir. E-elmə vahid bir sistem kimi baxıldığı halda isə, onun həll etdiyi məsələlərə uyğun müxtəlif alt sistemlərdən təşkil olunduğunu görmək olar. Elmin informasiya təminatı; elmmetrik (elmmetriya, bibliometriya, vebometriya) təhlillər; intellektual analiz və Big data alt sistemləri böyük həcmli elmi verilənlərin yaranmasına gətirməklə yanaşı, eyni zamanda onun aktual problemlərinin həllini həyata keçirir [2]. E-elmin müxtəlif elm sahələrində tətbiqi böyük həcmli elmi verilənlərin sürətlə artmasında generator rolunu oynayır.

Böyük verilənlər (*big data)* dedikdə elə böyük həcmli və mürəkkəb verilənlər toplusu başa düşülür ki, ənənəvi verilənlər bazasında nəzərdə tutulmuş standart alətlər onların real vaxtda saxlanması, idarə edilməsi və emalının öhdəsindən gələ bilmir. Misal olaraq dünyadakı ən böyük eksperimental qurğu olan Böyük Adron Kollayderi – yüklü hissəciklərin sürətləndiricisini göstərmək olar. Kollayder Avropa Nüvə Tədqiqatları Mərkəzində (frans. *Conseil Europėen pour la Recherce Nucėaire - CERN)* tikilmişdir. Hissəciklərin toqquşması kollayderin detektorlarında milyonlarla sensorlar vasitəsilə qeyd olunur və nəticədə böyük həcmdə verilənlər yaranır. 2010-cu ildə təcrübələr zamanı toplanmış verilənlər cəmi 13 Petabayt olmuşdur. İndi isə informasiyanın həcmi saniyədə 1 Gbaytı aşır. Aydındır ki, bu qədər böyük həcmli informasiyanın toplanması və emalını adi hesablama - yaddaş vasitələri ilə həyata keçirmək mümkün deyildir. Beləki, ilkin filtrasiyadan sonra *CERN*-də ildə 25 Petabayta qədər verilənlər

saxlanılır. Informasiyanın belə nəhəng həcminin analizi üçün paylanmış kompüter şəbəkə texnologiyası – *GRID*-dən istifadə olunur [3]. Bundan başqa astrofizikada məşhur olan *SETI (Search for Extraterrestrial Intelligence)* layihəsi radioteleskopdan alınan verilənlərin paylanmış emalı vasitəsilə yerdənkənar sivilizasiyaları axtarır. Layihədə böyük həcmdə radiosiqnalların emalında istifadə olunan kompüterlərin ümumi gücü Top 500-ə daxil olan superkompüterlərin gücündən dəfələrlə çoxdur [4]. Eyni zamanda 2000-ci ildə başlanmış *SDSS (Sloan Digital Sky Survey)* layihəsi çərçivəsində səmanın rəqəmsal astronomik müşahidələrində (sutkada 200 Gb-a yaxın yığım sürəti ilə) toplanmış verilənlərin həcmi hazırda 140 Terabaytı keçmişdir [4]. Bunlarla yanaşı asrofizika sahəsində tanınan *LSST (Large Synoptic Survey Telescope)* - Böyük müşahidə teleskopu göyün əlçatan hissəsinin çəkilməsi üçün nəzərdə tutulmuşdur. Müşahidə zamanı toplanan verilənlərin həcminin (sutkada 15 Terabayt olmaqla) 10 il ərzində 60 Petabayta çatması gözlənilir. Hesablama gücünün isə 150 Teraflops-dan başlayaraq sonda 950 Teraflops–a çatdırılması planlaşdırılır [6]. Təcrübi eksperimentlərlə bol olan tibb və biologiyada aparılan tədqiqatlar müasir informasiya texnologiyaları sahəsində əldə olunan ən yeni nailiyyətlərlə sıx bağlıdır və əsasən böyük həcmli rəqəmsal informasiyanın emalına əsaslanır. Verilənlərin böyük həcminin analizinin zəruriliyi problemi özünü ən çox bioinformatikada, DNK üzərində işləyərkən göstərir. Fərdiləşdirilmiş tibbdə xəstələr haqqında müxtəlif mənbələrdən toplanmış məlumatları özündə saxlayan elektron tibbi kartların emalı mühüm rol oynayır. Böyük həcmli tibbi verilənlərin analiz edilmə imkanı yeni müalicə üsulları tapmağa və xəstəliklərin təbiətini daha yaxşı başa düşməyə kömək edir. Riyazi modelləşdirməyə və böyük həcmli verilənlər üzərində yüksəksürətli hesablamalara əsaslanan tibbi tədqiqatlar tibbi xidmətin səviyyəsini qaldırır.

III. ELMDƏ BİG DATA TƏŞƏBBÜSLƏRİ

Elm üzrə Beynəlxalq Şuranın *(International Council for Science-ICSU)* məqsədi alimlərin beynəlxalq elmi fəaliyyətinin yayılmasına, beynəlxalq elmi birliklərin fəaliyyətinin əlaqələndirilməsinə, sahələrarası qlobal xarakterli elmi proqramların işlənilməsi və tədqiqatların aparılmasına yardım etməkdir [7]. *ICSU*-nun üzvləri 142 ölkəni təmsil edən 122 elmi qurum və 31 beynəlxalq elmi birlikdir. O, *UNESCO*-nun yanında məsləhətçi statusuna malikdir. *ICSU* fəaliyyəti ilə *Big Data* problemlərinin həllində mühüm rol oynayır. Onun elmi verilənlərə universal və bərabərhüquqlu çıxışın təmin edilməsi, onların daha yaxşı idarə edilməsi üçün yaradılmış aşağıdakı təşəbbüsləri misal göstərmək olar:

- Elmi informasiya və Verilənlərin sahə üzrə qiymətləndirilməsi Paneli;
- İnformasiya və Verilənlər üzrə Strateji Komitə;
- İnformasiya və Verilənlər üzrə Strateji Əlaqələndirmə Komitəsi.

ICSU-nun bu təşəbbüsləri Dünya Verilənlər Sisteminin (*World Data System - WDS*) inkişafında mühüm rol oynamışdır [8]. Bu sistemin məqsədləri aşagıdakılardır:

- Elmi verilənlərə universal və bərabərhüquqlu çıxışı təmin etmək;
- Verilənlərin standarlarına və konvensiyalarına əməl olunmasını həvəsləndirmək;
- Verilənlər və informasiyanın təminatlı keyfiyyətini təmin etmək;
- Verilənlərin daha yaxşı idarə edilməsi üçün şərait yaratmaq.

CODATA (*Committee on Data for Science and Technology*)- Elm və Texnologiya üçün Məlumat Komitəsi *ICSU*-nun sahələrarası komitəsi olub, onun fəaliyyətinin əlaqələndirilməsı, qlobal xarakterli problemlərin işlənilməsi və tədqiqatların aparılmasına kömək göstərir [9]. O, elm və texnikanın bütün sahələri üzrə verilənlərin idarə olunması və əlyetərliyinin yaxşılaşdırılmasına səy gösrərir. *CODATA* qarşıya qoyulan məqsədə çatmaq üçün məsələ qrupları, milli üzvlərin fəaliyyəti, konfranslar, simpoziumlar, nəşrlər, ümumi maraq dairəsində digər təşkilatlarla əməkdaşlıq kimi mexanizmlərdən istifadə edir. Onun 2002-ci ildən *“Data Science Journal*”ı nəşr olunur. *CODATA*-nın *ICSU*-nun maliyyələşdirdiyi beynəlxalq əməkdaşlıq elmi proqramlarına dəstək üçün 2014-cü ildə keçirdiyi “Beynəlxalq Elmi Proqramlar üçün *Big Data*: problemlər və perspektivlər” beynəlxalq seminarı *Big Data* problemlərinin və imkanlarının daha yaxşı başa düşülməsini təmin etmək məqsədi daşımışdır. Seminar “Beynəlxalq Elmi Proqramlar üçün *Big Data:* tövsiyələr və tədbirlər haqqında” adlı yekun bəyannamə qəbul etmişdir [10]. Bəyannamədə aşağıdakılar nəzərdə tutulmuşdur:

1.Beynəlxalq elmi proqramların sponsorları üçün tövsiyələr:

- Beynəlxalq elmi proqramların *Big Data* tələblərinə cavab verməsi;
- Cəmiyyət üçün *Big Data*-nın faydalarından istifadə edilməsi;
- Beynəlxalq əməkdaşlıq vasitəsilə *Big Data*-nın anlaşılmasının təkmilləşdirilməsi;
- Qlobal tədqiqat infrastrukturları vasitəsilə *Big Data*-ya universal çıxışın təbliği;
- *Big Data*-nın idarə olunması problemlərinin tədqiqi;
- *Big Data* sahəsində elmi potensialın və bacarığın artırılmasının təbliği;
- *Big Data*-dan istifadə siyasətinin inkişaf etdirilməsi.

2. CODATA işçi qrupuna təklif olunan tədbirlər:

- Beynəlxalq elmi proqramlar üçün Big Data sahəsində tematik tədqiqatların artırılması;
- Elm sahələri üzrə *Big Data* həllərinin mübadiləsinin təbliği;

- *Big Data* üçün siyasi, etik və hüquqi məsələlərin tədqiqi;
- *Big Data* üçün idarəçilik və sabitlik problemlərinin tədqiqi.

Maraq doğuran digər təşəbbüs Avropa Birliyinin yeni tədqiqat və innovasiya çərçivə proqramı "Horizon 2020" ilə əlaqədardır. Proqram çərçivəsində İqlim, Enerji, Qida, Səhiyyə, Nəqliyyat, Təhlükəsizlik və İctimai Elmlər sahələrində ən mühüm ictimai problemlərin həlli nəzərdə tutulmuşdur. Bu sahələrin hər birində böyük həcmdə verilənlərin toplanması, inteqrasiyası, emalı, təhlili və vizuallaşdırılması mühüm rol oynayır. Proqramın maliyyələşdirdiyi *Big Data Europe* layihəsinin məqsədi bu məsələlərin həlli üçün açıq *Hadoop, Cassandra* və *Storm* kimi *Big Data* texnologiyaları əsasında yeni platformanın işlənməsidir [11].

IV. BÖYÜK VERİLƏNLƏRİN EMALI ÜÇÜN QABAQCIL TEXNOLOGİYALAR

Superkompüter və klasterlər. Superkompüter çox prosessorlu xüsusi, yüksək məhsuldarlıqlı hesablama sistemidir. Klasterlər isə yüksəksürətli əlaqə kanalları ilə əlaqələndirilmiş və vahid hesablama vasitəsi kimi istifadə olunan kompüterlərdən (və ya superkompüterdən) təşkil olunur. Hesablama klasterləri verilənləri bölərək onları qovşaqlarda paralel emal etməklə, tək kompüterlə müqayisədə, hesablama vaxtını əhəmiyyətli dərəcədə azaltmağa imkan verir. Açıq proqram təminatı, sadə şəbəkə texnologiyaları və adi kompüterlər əsasında yüksək məhsuldarlıqlı belə hesablama komplekslərini yaratmaq mümkündür.

Qrid. Bu texnologiya elmi verilənlərlə intensiv əməliyyatlar aparılması üçün elmi təşkilatların öz hesablama güclərini birgə istifadə etməsi konsepsiyasıdır. Mahiyyətcə böyük resurslar tələb edən məsələlərin həlli üçün müxtəlif tipli hesablama resurslarının vahid infrastrukturda birləşdiyi paylanmış hesablamaların bir növüdür. Paylanmış hesablamaların üstünlüyü ondan ibarətdir ki, sistemin qovşaqları hətta adi kompüterlərdən təşkil oluna bilər. Beləliklə, nəzəri olaraq superkompüterlərdə olan hesablama güclərini daha aşağı qiymətlə almaq olar. Təəssüf ki, böyük həcmli informasiyanın emal üçün uzaq qovşağa göndərilməsi zərurəti yarandıqda, mövcud kompüter şəbəkələrində verilənlərin ötürülmə sürətinin aşağı olması səbəbindən, bu texnologiya yararlı olmur. *Qrid* layihələrə ABŞ-ın 4 iri superkompüter mərkəzinin resurslarını birləşdirən *TeraGrid*, böyük hesablama gücü tələb edən elmi tədqiqatlar *üçün Open Science Grid*, Avropada 18 kompüter mərkəzini birləşdirən *DEISA (Distributed European Infrastructure for Supercomputing Applications*), Avropa İttifaqında yüksək enerjilər, ekologiya və bioinformatika sahəsində verilənlərin emalı üçün qridin qurulması üzrə *DataGrid*, Yaponiyada elmi-tədqiqat fəaliyyəti üçün qrid infrastrukturunun qurulması üzrə *NAREGI (National Research Grid Initiative)* və s. göstərmək olar. Hazırda qrid texnologiyasından nüvə fizikasında, ətraf mühitin qorunmasında, hava proqnozunda və iqlim dəyişikliklərinin modelləşdirilməsində, maşınqayırmanın modelləşdirilməsində, bioloji modelləşdirilmədə və s. geniş istifadə olunur.

Hesablama buludları (Cloud Computing). İstifadəçilərin internet vasitəsilə xidmətlərə, hesablama resurslarına və proqramlara (əməliyyat sistemləri və infrastruktur daxil olmaqla) məsafədən girişini təmin edir. Böyük həcmli verilənlərlə iş zamanı onların əvvəlcədən eyni bir yerə toplanması və sonra isə analiz edilməsi mümkün deyildir. Doğrudan da, həcmi Petabaytlarla ölçülən verilənlərin analizi zamanı onların şəbəkə vasitəsilə, misal üçün, bir tədqiqat mərkəzindən digərinə ötürülməsi hətta yüksək məhsuldarlıqlı şəbəkələr vasitəsilə də effektiv həyata keçirilə bilməz. Ona görə verilənlər öz yerlərində qalır, bulud texnologiyaları isə onların müxtəlif yerlərdən seçilməsini və emalı üçün zəruri hesablama gücünü təqdim etməklə analizinin aparılmasını təmin edir. Beləliklə, bulud texnologiyaları bir tərəfdən böyük həcmli informasiyanın toplanması və saxlanmasını mümkün etməli, digər tərəfdən isə bu verilənlərin emalı üçün zəruri prosessor gücünü təqdim etməlidir.

NƏTİCƏ

Big Data texnologiyaları elmdə çoxlu problemlər yaratmaqla yanaşı, eyni zamanda onun inkişafı üçün də yeni imkanlar açır. Böyük həcmli elmi verilənlərin professional idarə edilməsi və analizi nəticəsində yeni biliklərin əldə olunması və elmin sürətli inkişafı üçün yeni imkanlar yaranır. Böyük həcmli elmi verilənlər probleminin həllinin perspektiv istiqaməti hesablama buludları texnologiyasından istifadədir. Bu yolla istifadəçilər böyük həcmli informasiyaya uzaqdan birgə giriş imkanı əldə edir və onların emalı üçün paylanmış hesablama resurslarından daha səmərəli istifadə edir.

TƏŞƏKKÜRNAMƏ

Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişafı Fondunun maliyyə yardımı ilə yerinə yetirilmişdir – **Qrant № EİF-2014-9(24)-KETPL-14/02/1**

ƏDƏBİYYAT

[1] R.M.Əliquliyev, R.Q.Ələkbərov, T.X.Fətəliyev. Elektron elm: cari vəziyyəti, problemləri və perspektivləri // İnformasiya Texnologiyaları Problemləri, 2015, №2, səh.4-15.

[2] T.X.Fətəliyev. Elektron-elmin intormasiya təhlükəsizliyi haqqında // Beynəlxalq Telekommunikasiya İttifaqının 150 illiyinə həsr olunmuş “İnformasiya təhlükəsizliyinin multidissiplinar problemləri” üzrə II respublika elmi-praktiki konfransı, Bakı, 14 may 2015, səh.145-147.

[3] Search for Extraterrestrial Intelligence. http://setiathome.berkeley.edu

[4] Brumfiel G. High-energy physics: Down the petabyte highway // Nature. 2011. V. 469. № 7330. pp. 282–283.

[5] The Sloan Digital Sky Survey. www.sdss.org

[6] The Large Synoptic Survey Telescope. www.lsst.org/lsst

[7] International Council for Science. www.icsu.org

[8] World Data System. www.icsu-wds.org

[9] CODATA-Committee on Data for Science and Technology. www.codata.org

[10] Big Data for International Scientific Programmes: Challenges and Opportunities. http://codata.org/blog/wp-content/uploads/2014/06 /CODATA-Big-Data-Workshop-STATEMENT-v07-FINAL.pdf

[11] Big Data Europe. www.big-data-europe.eu

Texnoparkların İnnovasiya Fəaliyyətinin Qiymətləndirilməsində Big Data Texnologiyalarının Tətbiqi

Ələvsət Əliyev[1], Roza Şahverdiyeva[2], Vüsalə Abbasova[3]

[1,2,3]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*alovsat_qaraca@mail.ru*, [2]*roza02@rambler.ru*, [3]*vusaleabbaszade@gmail.com*

***Xülasə* – Məqalə texnoparkların innovativ fəaliyyətinin təşkili və idarə olunması proseslərinin qiymətləndirilməsində Big Data texnologiyalarının tətbiqi problemlərinə həsr olunmuşdur. İnnovasiyaların işlənilməsində, innovativ məhsulların istehsalında və yüksək elmtutumlu xidmətlərin göstərilməsində xüsusi rola malik olan texnoparkların xüsusiyyətləri izah olunmuşdur. Texnoparklarda idarəetmə proseslərinin informasiya təminatının tərkib elementləri və arxitekturu təsvir edilmişdir. Texnoparkların innovasiya potensialı və fəaliyyət nəticələrinin qiymətləndirilməsi üçün zəruri göstəricilərin mürəkkəbliyi və struklaşdırılma səviyyəsi təhlil olunmuşdur. Texnoparklarda böyük həcmli verilənlərin operativ analitik təhlili üçün müasir Big data texnologiyalarının istifadə məqsədləri və üstünlükləri əsaslandırılmışdır. Big data texnologiyalarının texnoparklarda tətbiqi istiqamətləri, xüsusiyyətləri, vizuallaşdırma metodları və perspektivləri göstərilmişdir.**

Açar sözlər – innovativ iqtisadiyyat, texnopark, big data, verilənlərin analitik təhlili, verilənlər bazası, innovasiya potensialı, innovasiya fəaliyyəti, informasiya təminatı sistemi

I. GİRİŞ

Hazırkı dövrdə inkişaf etmiş dövlətlərdə iqtisadiyyat informasiyaya, biliyə və innovasiyaya əsaslanır. Bu isə elm-təhsil-istehsalat sahələri arasında qarşılıqlı əlaqələrinin gücləndirilməsini, bu istiqamətdə yeni idarəetmə mexanizmlərinin işlənilməsini, innovasiya mərkəzlərinin, texnoloji komplekslərin, texnoparkların, biznes-inkubatorların yaradılmasını və fəaliyyətlərinin təşkili məsələlərinin həll edilməsini tələb edir [1]. Müasir qloballaşma və inteqrasiya dövründə innovativ proseslərin kompleks təhlili, idarə olunması və proqnozlaşdırılması böyük həcmli verilənlərin toplanması və emalı əsasında həyata keçirilməyə başlanılmışdır. Hazırkı vaxtda dövlət idarəetmə strukturları və innovasiya menecerləri mövcud informasiyaları tam həcmdə toplayıb strukturlaşdıra və təhlil edə bilmədiklərindən, onların qəbul etdikləri qərarlar və işlədikləri konsepsiyalar, layihələr və proqnoz məlumatları da müəyyən qədər səmərəlilik səviyyəsinə uyğun olunur [2, 3].

İnnovativ tədqiqatlar üçün lazım olan ilkin zəruri informasiyanın həcmi durmadan kifayət qədər sürətlə çoxalmaqdadır. Bu, həm də onunla izah olunur ki, informasiyanın ilkin toplanılma və qeydiyyata alınma mənbələrinin sayının dinamiki artmasıyla yanaşı, göstəricilərin say tərkibinin və müxtəlif təhlil kəsimlərinin çoxalması, ayrı-ayrı keyfiyyətlilik hallarının qiymətləndirilməsi üçün birbaşa göstəricilərin olmaması və s. kimi hallar da xeyli artmışdır. Bu isə, öz növbəsində sürətlə dəyişməkdə olan müxtəlif xüsusiyyətli informasiyanın emalı və idarə olunması üçün müasir idarəetmə mexanizmlərinin və vasitələrinin işlənilməsini və tətbiqini qaçılmaz edir.

II. TEXNOPARKLARDA İNNOVASIYA FƏALİIYYƏTİNİN XÜSUSİYYƏTLƏRİ

Texnoparklar iqtisadi infrastrukturun formalaşdırılması və innovativ proseslərin reallaşması vasitəsidir. Onlar yaradıldığı məqsədlərə müvafiq olaraq müxtəlif istiqamətlərdə, o cümlədən İKT, sənaye, nəqliyyat və s. sahələrdə ixtisaslaşırlar. Texnoparkların yaradılması iqtisadiyyatın davamlı inkişafı və rəqabət qabiliyyətinin artırılması, müasir elmi və texnoloji nailiyyətlərə əsaslanan yüksək texnologiyalar sahələrinin genişləndirilməsi, elmi tədqiqatların aparılması və innovativ texnologiyaların işlənilməsi üzrə müasir komplekslərin yaradılması, təhsil müəssisələri, elmi-tədqiqat institutları və sənaye arasında qarşılıqlı əlaqələrin təkmilləşdirilməsi və s. kimi spesifik məqsəd və xüsusiyyətlərə malikdir. Bunlardan asılı olaraq texnoparklar bir çox əsas funksiyaları yerinə yetirirlər: Onlar elmi-tədqiqat və təcrübi-konstruktor işləri aparır, kiçik innovasiya şirkətlərini formalaşdırır, yüksək ixtisaslı mütəxəssislər üçün yeni iş yerləri yaradır, ali məktəb müəssisələri, elmi tədqiqat institutları və sənaye arasında qarşılıqlı əlaqələri formalaşdırır, texnopark bazasında yeni yaradılmış elm və texnologiya tutumlu məhsulu kiçik seriyalarla və ya kütləvi istehsal edir və onun yerli və xarici bazarlarda satışı ilə məşğul olur.

Texnoparkların bu cür kompleks fəaliyyəti onun əsas parametrlərinin müəyyənləşdirilməsini, idarəetmə modelinin konseptual strukturunun və icra mexanizmlərinin işlənilməsini zəruri edir [4]. Bütün bunlar iqtisadi struktur kimi fəaliyyət göstərən texnoparklarların effektiv idarəçiliyi üçün onun informasiya təminatı sistemlərinin təkmilləşməsini tələb edir. Yeni iqtisadiyyatın inkişafında inovativ struktur kimi texnoparklar xüsusi rola malikdirlər. Ona görə də texnoparkların idarə olunması prosesində informasiya amilinin nəzərə alınmasının çətinliyi, əhəmiyyəti, informasiyanın qeyri-

müəyyənliyi, mürəkkəbliyi, çatışmamazlıqları və toplanması kimi məsələlərin həlli də aktual əhəmiyyətə malikdir.

Texnoparklarda ilkin verilənlərin operativ və analitik təhlili onun inkişafına xidmət edir. Ancaq həmin verilənlərin həcminin kifayət qədər böyüməsi onların nizamlanmasını, strukturlaşmasını və operativ təhlilini tələb etdiyindən, həmin proseslərə müasir İKT, başqa sözlə Big data texnologiyaları tətbiq olunmalıdır.

III. TEXNOPARKLARIN İDARƏ OLUNMASI PROSESLƏRİNİN XÜSUSİYYƏTLƏRİ

Qeyd etmək lazımdır ki, Big Data texnologiyalarının tətbiqi ümumi halda texnoparkların idarə edilməsi sisteminin funksiyalarından da asılıdır. Həmin funksiyalara aşağıdakılar daxildir [5]:

1) innovativ fəaliyyətin dəstəklənməsi sisteminin formalaşdırılması,
2) elmi-tədqiqat və təcrübi-konstruktor işlərinin yerinə yetirilməsi üçün əlverişli şəraitin yaradılması,
3) innovasiya müəssisələrinə konsaltinq, mühəndis, texniki, texnoloji, informasiya və digər xidmətlərin göstərilməsi,
4) bazarların tədqiqi, əlaqədar müəssisələrə marketinq xidmətlərinin göstərilməsi,
5) innovasiyalı məhsulun üzə çıxarılması, seçim və istehsalda tətbiqinə qədər onların müşayiət olunması,
6) qərar qəbuletməyə dəstək məqsədilə intellektual informasiya sisteminin işlənilməsi və s.

Texnoparkların idarə olunmasında istifadə olunan informasiyanı əsasən, aşağıdakı mənbələrdən əldə etmək olar:

1. informasiya mərkəzləri,
2. mətbuat, kütləvi informasiya vasitələri,
3. ticarət - sənaye palatası,
4. statistika komitəsi,
5. standartlaşdırma və metrologiya komitəsi,
6. iqtisadi inkişaf nazirliyi,
7. banklar,
8. konsaltinq firmaları və s.

İdarəetmə proseslərinin informasiya təminatı sistemi bir sıra elementləri özündə əks etdirir [6]:

1) informasiyanın toplanması,
2) informasiya dövriyyəsinin təşkili,
3) informasiyanın saxlanması və aktuallaşdırılması,
4) informasiya sisteminin texniki təminatı,
5) informasiya sisteminin riyazi proqram təminatı,
6) informasiyanın verilməsi və s.

Texnoparkların dayanıqlı fəaliyyətinin təşkilində idarəetmə qərarlarının düzgün hazırlanmasının əsası idarəetmə subyektinin kifayət qədər düzgün məlumatla təmin edilməsində və onun mükəmməl informasiya təminatına malik olmasındadır.

IV. TEXNOPARKLARIN İNFORMASİYA TƏMİNATININ STRUKTUR TƏRKİBİ

Texnoparkın idarə olunması proseslərində dövlət strukturları, patent, sertifikatlaşdırma, lisenziya, ixtira strukturları, maliyyə strukturları, konsaltinq təşkilatları, innovasiya generatorları, aktiv innovasiya müəssisələri, biznes inkubatorları, regional strukturlar, tədris strukturları, elmi-tədqiqat strukturları, bank, sığorta şirkətləri, marketinq və reklam mərkəzləri, sosial obyektlər, təcrübi-eksperimental müəssisələr arasında zəruri informasiya əlaqələri qurulmalı və müvafiq verilənlər bazası, həmçinin dəstəkləyici qərar qəbuletmə sistemləri yaradılmalıdır [6]. Texnopark strukturları arasında sirkulyasiya edən informasiya onun informasiya təminatı sisteminin (Şəkil 1) formalaşması üçün baza rolunu oynayır.

Şəkil 1. Texnoparkların informasiya təminatı sisteminin əsas struktur blokları

V. TEXNOPARKLARIN İNNOVASİYA FƏALİYYƏTİNİN VƏ POTENSİALININ QİYMƏTLƏNDİRİLMƏSİ GÖSTƏRİCİLƏRİ

Ümumiyyətlə, texnoparkların fəaliyyətinin qiymətləndirilməsi onların informasiya təminatı bazasınını təşkil edən bir çox göstəricilər əsasında hesablanır.

Texnoparkların fəaliyyəti müxtəlif tip göstəricilərlə qiymətləndirilir. Əsas göstəricilərdən biri innovasiya potensialının qiymətləndirilməsi göstəricisidir. Texnoparkların innovasiya fəaliyyətini həm ayrı-ayrılıqda olan göstəricilərlə, həm də qrup və istiqamətlərdə birləşən göstəricilərlə xarakterizə edirlər.

İnnovasiya fəaliyyəti üzrə əsas göstəricilərin dinamikasını təhlil etmək üçün aşağıdakılara diqqət yetirmək lazımdır [7]: 1)müəssisənin innovativ aktivliyi, 2)müəssisələrdə elmi-tədqiqat bölməsinin fəaliyyəti, 3)daxili və xarici bazarlardakı satışın həcmində innovasiya məhsul və xidmətlərinin strukturu, 4)innovasiya fəaliyyəti nəticələrinin reytinq göstəriciləri, 5)ölkə və regionlar üzrə innovasiya fəaliyyəti ilə məşğul olan müəssisələrin strukturu, 6)maliyyə mənbələri üzrə innovasiya xərclərinin strukturu, 7)innovasiya fəaliyyəti növləri üzrə xərclərin strukturu, 8)texnoloji innovasiyalara mane olan faktorların reytinqi.

Qeyd etmək lazımdır ki, informasiyanın sürətli emalı, onun faydalılığı innovativ strukturların fəaliyyət səmərəliliyi üçün müsbət nəticələrdən sayılır və ona müəyyən tələblər qoyulur. Texnoparklar geniş spektrli müəssisə olduğundan informasiyanın dəqiq strukturlaşmaya malik olması, orada informasiya təminatının müasir idarəetmə strukturuna uyğun olması böyük həcmli verilənlərlə işləmə texnologiyasından istifadə zərurətini yaradır.

VI. TEXNOPARKLARDA BİG DATA TEXNOLOGİYALARININ TƏTBİQİ XÜSUSİYYƏTLƏRİ VƏ PERSPEKTİVLƏRİ

Texnoparklarda ənənəvi üsul və vasitələrlə işlədilən mövcud verilənlər bazasının (VB) həcmi qiqabaytlar və terabaytlarla qiymətləndirilə bilərdisə, artıq Big Data texnologiyaları vasitəsilə emal olunan informasiyanın həcmi petabayt və exsabaytlarla ölçülür. İnformasiyanın saxlanması qeyri-mərkəzləşdirilmiş formada həyata keçirilir. Onların strukturlaşma dərəcəsi az, bəzi hallarda isə tamamilə qeyri-müəyyən formada olur. Verilənlər arasındakı daxili əlaqə kifayət qədər zəif formada özünü göstərir. Verilənlərin saxlanması və emalı texnologiyaları üfiqi modelləşməyə əsaslanır [8-10]. Ona görə də texnoparkların ikin verilənlərlə işləməsinin çətinliyi informasiya emalının həddən artıq olan həcmi, sürəti, dəqiqliyi, dəyəri, faydalılığı və çox müxtəlifliyidir.

Yaranmış mürəkkəb situasiyadan çıxmaq üçün innovativ texnoparklarda həm müasir modelləşdirmə vasitələrindən, həm də Big Data texnologiyalarından istifadə olunmağa başlanılmışdır. Big data texnologiyalarından geniş istifadə etməklə tədqiqat və idarəetmə obyekti haqqında ətraflı və dəqiq informasiya toplama, proseslərə təsir gücünə malik faktorları müəyyənləşdirmək, həmin faktorların çəki əmsallarını hesablamaq mümkündür. Beləliklə, texnoparklarda real zaman çərçivəsində daha dəqiq proqnozlaşdırma işlərini həyata keçirmək imkanı yaranır. Elə bunun nəticəsidir ki, innovativ tədqiqat proseslərinə hazırki dövrdə daha çox tətbiq olunan və getdikcə məşhurlaşan Big data texnologiyaları investisiya qoyuluşları üçün də əsas sahəyə çevrilməkdədir.

Müasir texnoparklarda ənənəvi strukturlaşmaya malik olmayan çoxlu sayda mətn sənədləri, cədvəllər, video yazılışlar, veb-jurnallar, maşın kodları, şəkillər, hesablamalar, ticarət məlumatları, qeydiyyat materialları və s. toplanmışdır [5]. Bütün bunlar müxtəlif formatlarda, müxtəlif yerlərdə, müxtəlif saxlama obyektlərində toplanmaqdadır. Kifayət qədər böyük həcmə və müxtəlifliyə malik elə informasiyaya müraciət etmək, onlardan səmərəli istifadə etmək olduqca mürəkkəbdir. Nəzərə alınmalıdır ki, indiki dövrdə layihələndirmə və idarəetmə proseslərinə sensorlar, videokameralar, intellektual sayğaclar, daimi informasiya istehsal edir və həmin informasiyanın operativ işlənilməsinə ehtiyac vardır. Belə olan şəraitdə Böyük verilənlər texnologiyaları vasitəsilə alınmış nəticələri daha asan təhlil etmək üçün onları müasir vasitələrlə vizuallaşdırmaq lazımdır. Hazırda teq buludlar, qrafiklər, diaqramlar, tarixi axın, fəza axını kimi üsullardan, istifadə olunur [5]. Teq buludlardan istifadə zamanı hər bir elementə onun mühümlüyündən asılı olaraq çəki əmsalı verilir. Həmin əmsalın böyüklüyünə uyğun şrift seçilir. Çəki əmsallarının ekspertlər tərəfindən verilməsi zamanı elementin dəyişilməsi tezliyi, ona təsir edən müxtəlif amillər nəzərə alınmalıdır. Bu isə informasiya çoxluğundan ən mühümlərini ayırd etməyə imkan verir. Qrafik və diaqramlar informasiyanı asan və başa düşülən formada verilməsinə kömək edir. Bəzi məlumat və sənədlərdə isə onun dəyişilməsi, redaktə olunması tarixini izləməyə ehtiyac yaranır. Həmin sənədin kim tərəfindən, nə vaxt, hansı vaxtda redaktə olunması, proseslərinin izlənilməsi müvafiq texnologiya vasitəsilə həyata keçirilir. Texnoparkların fəaliyyətində bəzi hallarda müəyyən informasiyanın ərazi prinsiplərinə görə paylanmasını və istifadə olunmasını təhlil etmək lazım gəlir. Bu zaman “ərazi informasiya axını” texnologiyalarından istifadə etməklə vizuallaşma işlərini həyata keçirmək olar. Elə hallar da olur ki, obyektlər arasında məzmun əlaqələrini aşkara çıxarmaq lazım gəlir. Bu zaman semantik şəbəkələrdən, istiqamətlənmiş qraflardan istifadə olunur. Bütün bunlara baxmayaraq, mövcud vizuallaşma metodları kifayət qədər tam işlənilməmişdir və onların həm təkmilləşməsinə, həm də yeni vizuallaşdırma metod və vasitələrinin işlənilməsinə ciddi ehtiyac vardır.

Texnoparklarda Big data texnologiyalarının tətbiqi nəticəsində bir çox uğurlu nəticələr əldə olunmuşdur ki, onlardan bəzilərinə aşağıdakıları aid etmək olar: 1)böyük verilənlərdən istifadə edən müəssisələrin istehlakçılarının daha da etibarlılığı, 2)yeni məhsullara olan tələbatın artması, 3)böyük bazar strategiyasının reallaşması, 4)rəqabət qabiliyyətli müəssələrə innovativ xidmətlərin göstərilməsi, 5)istehlakçıları başa düşməklə onların bazarda payının

artırılması, onların istəklərinə uyğun istiqamətlənmə, 6)istehlakçıların tam dəstəklənməsi və onların məmnuniyyətinin daha da artırılması, 7)proqnozlaşdırma üçün verilənlərin asan idarə olunması və əldə edilməsi, 8)yeni məhsulun işə salınması işlərinin daha da asan yerinə yetirilməsi, 9)real vaxt rejimində proses iştirakçıları olan tərəflərin birləşdirilməsi, 10)təchizat zəncirində şirkətlər ilə effektiv münasibətlərin qurulması, 11)strukturlaşdırılmamış məlumatlardan, sosial şəbəkələrdən effektiv yararlanma [11-13**]**.

Texnoparklarda Big data texnologiyalarının tətbiqi üçün onların müxtəlif proqram təminatı sistemləri haqqında müəyyən məlumatlara malik olmaq lazımdır.

Big data texnologiyalarının Proqram təminatı. Müasir Big data texnologiyalarının əksəriyyəti açıq proqram təminatı əsasında (open sourse) işlənildiyindən ilk növbədə Linux əməliyyat sisteminə və onun standart instrumentariyasına bələd olmaq lazımdır. Hazırda Böyük verilənlərin emalı üzrə populyar sistem Hadoop platforması və MapReduce paylanması hesablama paradigması hesab olunur. Son dövrlərdə Hive, Pig və s. kimi bir çox yeni proqram məhsulları işlənilmişdir ki, onlar böyük verilənlərlə işləməyi asanlaşdırmışdır. Buna baxmayaraq, həm SQL üzrə zəruri baza biliklərinə malik olmaqla, NoSQL kimi (HBase, Cassandra, MondoDB, Neo4j) müasir verilənlər bazası ilə iş təcrübəsinə də yiyələnmək olduqca vacibdir. Bu sahədə effektiv fəaliyyət üçün Pyhon, Java, Scala kimi əsas proqramlaşdırma dillərindən istifadə olunur. Maşın təlimi və statistik təhlillər üçün R dili daha çox istifadə olunur. Eyni zamanda, Pyhon (NumPy, SciPy, Nltk, Skikit) və Java-nın (Mahoot, Weka) müxtəlif proqram paketləri də tətbiq olunur [14-16].

Big data modellərinin tətbiq sahələri. Texnoparklarda Big data modellərinin bəzi tətbiq sahələrinə aşağıdakılar daxildir [17]: 1) bazarların əraziyə görə seqmentasiyası, 2) alıcı informasiyasının təhlili, 3) müəyyən əmtəə və xidmətlərə alıcı reaksiyasının öyrənilməsi, 4) kontentlərin təhlili, 5) innovativ ideyaların toplanması və sistemləşdirilməsi, 6) konkret innovasiyalara bazarın reaksiyasının müəyyənləşdirilməsi, 7) innovativ ideyaların səmərəli işlənilməsi variantının müəyyənləşdirilməsi, 8) mürəkkəb dinamiki qərar qəbuletmə prosesinin operativ təhlili və s.

Big data modellərindən innovativ proseslərdə istifadə olunmasının tipik variantlarına seqmentasiya, müştəri axıcılığının azaldılması, şirkətlərin idarəçiliyi, yeni innovativ məhsul və xidmət yaradılması, əmtəənin mümkün irəliləməsinin yollarının təyini, marketing tədqiqatlarını, qərar qəbuletmə proseslərini və s. aid etmək olar. Müasir texnoparklarda, xüsusən innovativ proseslərin idarə olunmasında Big data texnologiyalarının tətbiqi verilənlər arasındakı əlaqələrin müəyyənləşdirilməsi prosesinə yeni baxışı və yanaşmanı tələb edir. Bu zaman əsas məsələ bütün verilənlər spektrini kompleks təhlil edə bilmək, dəqiqliyin ziyanına olsa da, nizamlanmamış və strukturlaşdırılmamış verilənlərlə işləməyə hazır olmaq və mühakimə yürütmə formasının dəyişilməsi ilə korrelyasiyalara inanmağı həyata keçirməkdir.

Texnoparklarda daim artmaqda olan verilənlərin həcmi və onların emalı proseslərinin surətliliyi iqtisadi və biznes proseslərinin idarə olunmasında Big data texnologiyalarını zəruri etmişdir [17]. Əksər müəssisə və təşkilatlarda verilənlərin yığılması, qeydiyyatı, təşkili, saxlanması və bəzi təhlili ilə bağlı problemlər yaranır: 1)artmaqda olan verilənlərin səmərəli emalı, 2)istifadəçi və həll əlavələrin həddən artıqlığı, 3)çoxlu sayda qeyri-ənənəvi mənbəələrdən alınan informasiyanın inteqrasiyası, 4)müxtəlif strukturlar üçün verilənlərin geniş spektrdə təhlili, inteqrasiyası və ümumi istifadə rejiminin təmini, 5)artıq verilənlərin mövcudluğu və onlarla işləmək zərurəti, 6)müxtəlif sistemlərin reaksiyası və verilənlərin uzlaşmasında ləngimələr, 7)operativ ötürmələrin və inteqrasiya problemlərinin çətinliyi.

NƏTİCƏ

Hesab olunur ki, texnoparklarda Big data texnologiyalarının tətbiqi hesabına istehsalat, xidmət, tədqiqat və idarəetmə sahələrində xərclərə qənaət olunacaq, istehsal və idarəetmə prosesləri sürətlənəcək, əmək məhsuldarlığı yüksələcək, kadrlara və onların öyrədilməsinə olan tələbat müəyyən qədər azalacaqdır. Bu məqsədlə Böyük verilənlər və biznes-analitika instrumentariyasından, paket proqramından istifadə olunması tövsiyyə olunur.

Texnoparkların fəaliyyətində Big data texnologiyalarının tətbiqi nəticəsində onların innovasiya fəaliyyəti sahəsində idarəetmə qərarlarının qəbulu üçün keyfiyyətcə yeni və səmərəli nəticələr əldə etmək olar.

ƏDƏBİYYAT

[1] “Azərbaycan Respublikasında informasiya cəmiyyətinin inkişafına dair 2014-2020-ci illər üçün Milli Strategiya”nın təsdiq edilməsi haqqında Azərbaycan Respublikası Prezidentinin Sərəncamı, 02.04.2014-cü il, http://www.president.az

[2] “Azərbaycan - 2020: Gələcəyə Baxiş” İnkişaf Konsepsiyası. Bakı, 29 dekabr 2012-ci il. http://www.president.az

[3] R.M.Alguliyev, A.G.Aliyev, RO.Shahverdiyeva. The content of innovations and structural analysis of their features in the formation of information economy. Life Science Journal 2014;11(12):119-125.

[4] А.Г.Алиев, Р.О. Шахвердиева, В.А. Аббасова. Концептуальные основы разработки механизмов управления инновационных технопарков. Українська наука: минуле, сучасне, майбутнє. Тернопільський національний економічний університет МОН України. №2, 2014. стр.14-27

[5] Иванов П.Д., Лопуховский А.Г. Технологии Big Data и различные методы представления больших данных. Инженерный журнал: наука и инновации, 2014, вып. 9. № 9(33).

[6] R.M.Alguliyev, A.G.Aliyev, R.O.Shahverdiyeva. Development of information support systems for management of innovative structures. The 8th IEEE International Conference on Application of Information and Communication Technologies (AICT2014). Kazakhstan, Astana, 15-17 october 2014. p. 378-382

[7] Ə.Q.Əliyev, R.O.Şahverdiyeva. İnformasiya iqtisadiyyatında innovasiyalar üzrə monitorinq sisteminin ilkin göstəricilər bazasının işlənilməsinin konseptual əsasları. “İnformasiya cəmiyyəti problemləri” jurnalı. №2, 2014-cü il. səh.50-60.

[8] Frane Adam. Innovation: Measurement and Indicators. Measuring National Innovation Performance. Springer. 2014, pp. 9-13

[9] Т.В.Губарева, Е.И.Луковникова. Особенности применения новых информационных технологий Больших данных. Проблемы социально-экономического развития Сибири: научный журнал. БрГУ. 2015 г. №2 (20). Стр.24-31.

[10] С.Артемов. Big Data: новые возможности для растущего бизнеса. http://www.jet.msk.su

[11] В. Н.Севостьянов. Big data как инновационный путь развития компании. Ж. Молодой учёный. № 8.2 (67.2). 2014 г. стр.55-60.

[12] Sofia Berto Villas-Boas. Big Data in Firms and Economic Research. Applied Economics and Finance, vol. 1, No1, 2014. pp.65-70.

[13] Tushar M. Chavan, S. P. Akarte. Opportunities and challenges of Big data in economics research and enterprises. International Journal of Computer Science and Mobile Computing, vol.3 Issue.4, 2014, pg. 1155-1161.

[14] Maria Rosaria Carillo, Erasmo Papagni. “Little Science” and “Big Science”: The institution of “Open Science” as a cause of scientific and economic inequalities among countries. Economic Modelling, vol.43, 2014, pp.42-56.

[15] Л. Е. Жуков. Конференция Большие Данные в национальной экономике. Москва, 22 октября 2013, www.ospcon.ru. Стр. 48- 50

[16] Jelinek, M., Bergey, P. Innovation as the strategic driver of sustainability: big data knowledge for profit and survival. Engineering management review. vol.41, Issue:2. pp14–22. 2013. IEEE

[17] McConnell, Elizabeth Anne, Birkett, Michelle, Shattell, Mona. The future of big data: innovative methodological approaches. 2015, vol.36, Issue 6, pp. 478-480. 3p. http://web.a.ebscohost.com

[18] Bogdan Nedelcu. About Big data and its challenges and benefits in manufacturing. Journal Database Systems. Vol.IV, №3/2013. pp.10-19

“Big Data” Neft-Qaz Sənayesində Qərar Qəbul Etmənin Əsas Komponenti Kimi

Məkrufə Hacırəhimova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
makrufa@science.az

***Xülasə* — Böyük verilənlər neft-qaz sənayesi üçün yad məfhum deyildir. Neft-qaz sənayesinin liderləri verilənlərin gücünü anlamaqla rəqəmsal neft sahəsinin reallığa çevrilməsində israrlıdırlar. Məqalədə böyük verilənlər neft-qaz sənayesində kəşfiyyat və qazma işlərində əsaslandırılmış qərarların qəbul edilməsində əsas amil baxımından analiz olunur.**

Açar sözlər — Big data; neft və qaz sənayesi; big data analitika; prediktiv analitika, qərar qəbul etmə.

I. GİRİŞ

Neft-qaz sənayesi son dərəcə rəqabətqabiliyyətli və tənzimlənə bilən mühitdir. Məhdud təbii enerji ehtiyatları, tələbat və qiymət dəyişiklikləri ilə xarakterizə olunan qeyri-müəyyən mühitdə neft-qaz şirkətləri istehsalı artırmaq, xərcləri optimallaşdırmaq, ekoloji risklərin təsirini azaltmaq məcburiyyətindədirlər [1, 2]. Bu gün yerin təkinin daha yaxşı öyrənilməsi istəyi neft-qaz şirkətlərini böyük tezliklə, müxtəlif tip və daha çox informasiya toplamağa sövq etmişdir. Neft-qaz sənayesində kəşfiyyat, qazma, hasilat və emal zamanı çox böyük verilənlər massivi toplanmaqdadır. Təbiidir ki, bu verilənlər arasında qiymətli informasiyalar var və onların əsasında əsaslandırılmış qərarların qəbul edilməsi ciddi məsələdir. Bu baxımdan böyük verilənlər (*big data*) və böyük verilənlərin analitikası (*big data analytics*) neft-qaz sənayesi üçün daha vacibdir [3-6]. Çünki verilənlərin analizi biznes sahəsində rəqabət və müvəffəqiyyətin açarıdır, şirkətlərə neft və karbohidrogen yataqlarının kəşfiyyatı və işlənməsi zamanı planlaşdırma və qərar qəbul etmədə əvəzolunmaz alətdir. Bu işdə də əsas məqsəd neft-qaz sənayesində “big data” texnologiyalarının imkanlarını və ekspertlər tərəfindən əsaslandırılmış operativ qərarların qəbul edilməsində əsas şərt və komponent olduğunu göstərməkdir.

Neft-qaz sahəsi üzrə beynəlxalq səviyyədə keçirilən “SPE Big Data and Analytics”, informasiyanın idarə edilməsi üzrə hər il keçirilən “European E&P Data and Information Management Conference”, “EAGE workshop on Oil and Gas Business Analytics”, “Big Data Solutions & Analytics in Upstream Oil and Gas Industry” və s. tədbirlər və orada müzakirə olunan məsələlərə nəzər salsaq, bu mövzunun nə qədər aktual olduğunu, neft-qaz sənayesində kəşfiyyat və qazma işləri çərçivəsində problemlərin həllində “big data” texnologiyalarının geniş tətbiqinin şahidi oluruq [7,8].

II. NEFT-QAZ SƏNAYESİNDƏ BÖYÜK VERİLƏNLƏRİN ANALİTİKASI VƏ QƏRAR QƏBUL ETMƏ

Qeyd edildiyi kimi neft-qaz sənayesində çox böyük həcmdə verilənlər generasiya olunur, toplanır. Belə ki, neft-qaz şirkətləri minlərlə sensorlar vasitəsi ilə yeraltı və yerüstü obyektlərdən (*neft quyuları və s.*) daim verilənlər toplayır (*məsələn, Şevronun 1 gün ərzində İT trafiki 1.5 terbaytı keçir*), real-vaxt rejimində onlar aktivlərinə nəzarət və ətraf mühitin monitorinqini [9], kəşfiyyat mərhələsində seysmik araşdırmalar zamanı əməliyyatları dəstəkləmək üçün 2D, 3D, 4D vizuallaşdırma, qabaqcıl geofiziki modelləşmə, simulyasiya həyata keçirirlər. Verilənlər yığımı o qədər böyük və mürəkkəbdir ki, onları ənənəvi üsullarla emal etmək mümkün olmadığından, neft-qaz şirkətləri öz biznes strategiyalarını yaxşılaşdırmaq üçün böyük verilənlərdən maksimum istifadənin yollarını axtarırlar. Toplanan informasiya müxtəlif formatda *(strukturlaşdırılmış və strikturlaşdırılmamış)* olur. Bu da öz növbəsində onları istifadə və interpretasiya etməkdə ciddi problem yaradır. Çünki hazırda kəşfiyyat və hasilat verilənlərinin mövcud idarə etmə platforması əsasında böyük verilənlərdən biznes-analitika üçün kritik informasiyanın əldə olunması qeyri mümkündür. Ona görə müxtəlif mənbələrdən və müxtəlif formatlarda toplanan verilənlərdən maksimum yararlanmaq, dəyər yaratmaq üçün yeni analitik texnologiyalara ehtiyac yaranır. Problemin həllində “big data” texnologiyalarının tətbiqi xüsusilə çox vacibdir [10]. Verilənlərin inteqrasiyasını və sintezini təmin etməklə, lazımi informasiyanı lazım olan vaxtda əldə etmək şirkətlər üçün düzgün qərar qəbul etməyə imkan verir. Şirkətlər topladıqları böyük həcmli verilənlər üzərində dərin təhlillər aparmaqla qərar qəbul etmə prosesini reaktiv vəziyyətdən aktiv vəziyyətə gətirə bilir [2].

Neft-qaz yataqlarının kəşfiyyatı və hasilatı zamanı toplanan verilənlər də bir qayda olaraq “5V” ilə xarakterizə olunur [3.5]:

- *Həcm (volume):* kəşfiyyat, qazma, hasilat, seysmik və s. *verilənlər;*
- *Müxtəliflik (variety):* strukturlaşdırılmış (*PPDM, SEG-Y, WITSML, PRODML, RESML və s. kimi verilənlərin standart modelləri*); strukturlaşdırılmamış (*şəkil, aidio, video və s.*); qismən strukturlaşdırılmış (*analiz və*

interpretasiya zamanı emal olunan verilənlər, qazma işləri üzrə gündəlik hesabatlar və s.);

- *Sürət (velocity):* real-vaxt rejimində qazma qurğularından (EDR, LWD, MWD və s.) və sensorlardan (Flow, Pressure, ROP və s.) daxil olan verilənlər axını;
- *Həqiqilik- verilənlərin keyfiyyəti* (*veracity*– data quality):
  – verilənlərin keyfiyyətinin yaxşılaşdırılması;
  – verilənlərin inteqrasiyası modellərinin yerinə yetirilməsi;
  – seysmik, qazma və hasilat məlumatların uyğunlaşdırılması;
  – verilənlərdə anomaliyaları müəyyən etmək üçün ilkin emal.
- *Dəyər (value):*
  – ılkin neftin alınmasının sürətləndirilməsi;
  – istehsalın artırılması;
  – xərclərin azaldılması
  – risklərin minimallaşdırılması.

Neft-qaz sənayesində “Big Data” texnologiyalarının tətbiqində əsas məqsəd aşağıdakılardan ibarətdir [11,12]:

- böyük sürət və dəqiqliklə daha çox karbohidrogen yataqları tapmaq;
- ətraf mühitə minimum zərər verməklə təhlükəsiz, effektiv neft çıxarmaq və təkmilləşdirmək;
- təbii enerji ehtiyatlarını ən effektiv və ən münasib dəyərə saflaşdırmaq və paylamaq;
- müştərilərlə münasibətləri idarə etmək, yeni sərfəli bazar və imkanlar tapmaq.
- risklərin azaldılması, xüsusilə təhlükəsizlik, ekologiya və s. sahəsində.

III. İMKANLAR VƏ PROBLEMLƏR

Neft-qaz şirkətləri verilənlərin gücünü anlamaqla və istifadə etməklə aşağıdakı inkanları əldə edirlər [10,13,14]:

- neft-qaz yataqlarının planlaşdırılması, kəşfiyyatı, işlənməsi və hasilatı mərhələlərində rəqabətqabiliyyətliliyin saxlanılması;
- texniki xidmət və proqnozlaşdırma ilə əlaqədar istehsalın səmərəliliyinin artırılması;
- ilkin neftin çıxarılnası vaxtının, istismar xərclərinin azaldılması və bütün mərhələlərdə məhsuldarlığın artırılması;
- lazımi informasiyanın lazım olan vaxtda avtomatlaşdırılmış qaydada əlyetərliliyinin təmini.

Deməli, böyük verilənləri analiz etməklə neft-qaz şirkətləri istehsal xərclərini azalda, qərar qəbul etməni yüksəldə, əməliyyatların və biznes-proseslərin effektivliyini artıra, müxtəlif tip verilənləri konvergensiya etməklə yeni ideyalar əldə edə bilər, yeni biznes-modellər işləyə bilərlər. Bir sözlə, böyük verilənlər investisiyalardan əldə olunmuş böyük gəlirdir.

Neft-qaz sənayesi kəşfiyyat və istehsal (emal) mərhələsində üç əsas problemlə qarşılaşır:

- *Verilənlərin idarə olunması* (*data management*). Bu analitik və statistik metodlardan istifadə etməklə informasiyanın analizi və effektiv axtarışına imkan verən böyük həcmdə strukturlaşmış və strukturlaşmamış verilənlərin toplanması, emalı və saxlanmasını özündə birləşdirir.
- *Kəmiyyət verilənləri* (*quantification of data*). – proqnozların verilməsi və daha dəqiq proqnoz vermə əhəmiyyətinə malik ideyaların təyini üçün statistik və analitik metodların tətbiqini özündə birləşdirir.
- *Riskin qiymətləndirilməsi* (*risk assessment*). – məlum olmayan risklərlə necə işləməyi bilmək üçün toplanmış məlum riskli verilənlərin riyazi modellər vasitəsilə yerinə yetirilən proqnoz analizidir.

IV. KƏŞFİYYAT VƏ HASİLAT

Biz rəqəmsal meyllərin ənənəvi biznes-modelləri üstələdiyi, verilənlərin heç vaxt bu qədər mövcud olmadığı erada yaşayırıq [15]. Bu meyllər sənaye sahələrində böyük verilənlərin və daha mütərəqqi analitikanın yaranmasına imkan verdi. Neft-qaz sənayesi isə bu texnologiyaların tətbiqində aparıcı yerlərdən birini tutur. Böyük verilənlər istər geoloji kəşfiyyat və qazma, istərsə hasilat və istehsal, istərsə də nəqli mərhələlərini optimallaşdırmaq, təhlükələrdən sığortalanmaq imkanı verir. Yeni resursların tədqiqi zamanı böyük verilənlər və qabaqcıl analitikadan istifadə seysmik izləri taparaq onları əvvəlki unudulmuş seysmik nəticələrlə birləşdirir ki, bu da olduqca mühüm nəticələr verir. Burada seysmik verilənlərin vizualizasiyası üçün Hadoop və Greenplum platformasından istifadə etmək daha münasibdir. Qazma və hasilat ilə bağlı tarixi məlumatlara istinad etməklə ətraf mühitin qorunması qaydaları ilə məhdudlaşdırılan sahələrdə yeni tədqiqatların aparılması zamanı fərziyyə və mülahizələrin yoxlanması tədqiqatçılara böyük kömək ola bilər. Məsələn. iqlim dəyişikliyi və buz axınları kimi informasiyalar analitiklərə qazma işlərində fırtınaların proseslərlə təsirini öyrənməyə imkan verə bilər [14-16].

[17]-də böyük verilənlərdən və analitikadan düzgün istifadə etdikdə tipik səhvlərdən qaçmaqla böyük fayda əldə etmək yolları göstərilir. Tədqiqatda həmçinin neft-qaz şirkətlərində “big data” üçün dörd əsas tətbiqi məsələlərin ortaya çıxdığı qeyd olunur: rəqəmsal sahələr (*digital fields*), prediktiv qurğu və qazma analizi (*predictive plant and drilling analysis*), məsafədən əməliyyatlar və rezervuar modelləşdirmə (*Remote operations and Reservoir modelling*) və seysmik təsvirlərin (*seismic imaging*) emalı.

Neft-qaz sənayesində kəşfiyyat zamanı istifadə oluna bilən böyük verilənlərlə bağlı analitik işlərə isə aşağıdakılar daxildir:

• *Axtarışın keyfiyyətinin yüksəldilməsi.* Yeni fəaliyyət qrupu yaratmaq üçün şirkətin verilənlərin real-vaxt rejimində istehsal verilənləri ilə birləşdirilmək;

• *Sahənin qiymətləndirilməsi və yeni perspektivlərin yaradılması.* Coğrafi məkan məlumatları, xəbər yayımları və ya digər informasiya mənbələrindən istifadə etməklə düzgün kəşfiyyat aparmaq.

Böyük verilənlərdən istifadə qazma ilə bağlı ehtimal edilən maneələri qabaqcadan görməyə, qazmaya təsir edən şəraiti və ya anomaliyaları müəyyən etməyə imkan verir. Bu da öz növbəsində milyonlarıla vəsaiti və minlərlə avadanlıqları xilas edə bilər.

Neft quyularının qazılması prosesində qazma briqadaları bəzi situasiyalara sərbəst nəzarət edə bilməlidirlər. Bu, quyuların ilkin planında nəzərə alınmayan hadisələr üçün xarakterikdir. Əgər qazma briqadaları müəyyən məqamlar haqqında daha çox məlumatlandırılarsa, o zaman onlar belə situasiyalarda daha əsaslandırılmış qərarlar qəbul edə bilərlər. Real-vaxt rejimində qazma əməliyyatları zamanı prediktiv analitikanın köməyi ilə potensial problemləri müəyyən etmək, proqnozlar üçün statistik modellər, evristik mühakimələr əsasında düzgün qərar verməyə kömək edə bilər [18].

Neft-qaz kəşfiyyatında analiik alətlərin əsas istiqamətləri aşağıdakılar təşkil edir:

- perspektivlərin qiymətləndirilməsi;
- seysmik izlərin müəyyən edilməsi;
- kəşfiyyat işlərinin sürətləndirilməsi;
- yeni elmi modellərin yaradılması.

V. QAZMA VƏ TAMAMLAMA

Torpaq qatının quruluşunu və anomaliyasını əvvəlcədən bilmək qazma prosesində həm əmək, həm də təchizat baxımından milyonlara qənaət etmək deməkdir. Həmçinin qazma prosesi üçün mövcud quyuların verilənləri ilə araşdırılan yeni quyunun verilənlərinin müqayisəsi əvvəlcədən baş verə biləcək hadisənin qarşısını almaqda əsas amillərdəndirr. Məhz, bu amil neft-qaz şirkətlərinin yerin dərin qatlarında baş verənlərin anlanması üçün bütovlükdə, yüksək sürətlərlə müxtəlif tip böyük həcmli verilənlərin yığılmasına gətirib çıxardı. Nəticədə, neft-qaz şirkətləri öz biznes-strategiyalarının yaxşılaşdırılmasında böyük verilənlərdən istifadə etməyə başladılar. Neft sənayesində qazma işlərinə çəkilən xərclərin yüksək olduğunu və bəzi hallarda isə müvəffəqiyyətsizliklə nəticələndiyini nəzərə alsaq quyuların qazılması zamanı qərarların qəbul edilməsində risklərin azaldılmasında böyük həcmdə dəqiqləşdirilmiş verilənlərin rolu çox böyükdür.

Qazma platformaları və boru-kəməri infrastrukturunun uğursuzluqları, catısmazlıqlarını əvvəlcədən görmək və qarşısını almaqda, qazma əməliyyatlarına təsir edə biləcək şərait və ya anomaliyaları muəyyən etməkdə böyük verilənlər və analitikanın imkanlarından istifadə etmək çox önəmlidir. Bu da insan həyatını və təchizatları qorumağa kömək edə bilər.

Böyük verilənlər və analitikanın qazma və tamamlama əməliyyatlarını inkişaf etdirə bildiyi məsələlər aşağıdakılardır:

- *Qazma modellərinin qurulması və qiymətləndirilməsi*: bu, qazma qurğusunda yerləşdirilmiş sensorlardan daxil olan verilənlər əsasında modelləri yeniləmək və qazma parametrlərini optimallasdırmağa kömək edəcək;
- *Qazma dəqiqliyininin artırılması və təhlükəsizlik*: qazmaya təsir edən anomaliyaların erkən müəyyənləşdirilməsi və zərbə, partlayış və s. kimi arzuolunmaz hadisələrin qarşısının alınması;
- *Qazma prosesinin optimallaşdırılması:* qazma əməliyyatlarına təsir edən neqativ faktorların erkən müəyyən edilməsi;
- *Maliyyə vəsaitlərinin optimallaşdırılması*- hesablama texnologiyalarından istifadə etməklə optimal dəyərin müəyyən edilməsi;
- *Real zamanda qərar qəbulu*: qazma məlumatları və onun qiymətləndirilməsi real zamanda icra edilməli və bu məlumat qərar qəbul etmədə, yeni elmi modellərim yaradılmasında istifadə edilməlidir və s.

Böyük həcmli geoloji verilənlərə statistik və kəmiyyət analizi, tədqiqat və proqnoz modelləşdirməsi və faktlara əsaslanan idarəetmə yanaşması tətbiq etməklə qərarların qəbulu, həmçinin kəşfiyyat və qazma proseslərində yeni ideyalar əldə etmək olar.

Onu da qeyd etmək lazımdır ki, IBM, Oracle, Microsoft, Teradata və s. kimi aparıcı informasiya texnologiyaları nəhənglərinin neft-qaz sənayesində "big data" və analitikanın tətbiqini mümkün edən məhsulları mövcuddur [2,19-21].

NƏTİCƏ

Neft-qaz sənayesi hazırda sürətli dəyişikliklər dövründədir. Neft və qaz sənayesi sürətlə inkişaf edən və rəqabətli sənayedir. Burada eksponensial olaraq artan verilənləri idarə etmək isə vacib bir çağırışdır. Yeni metod və texnologiyaların qazma və yeni neft-qaz yataqlarının geoloji kəşfiyyatında tətbiqi bu sahənin mütəxəssislərinin sözləri ilə desək – "Bu günün qazma işləri verilənlərdən asılıdır. Toplanmış verilənlər qazma və hasilatın təlimatçısıdır". Böyük verilənlərin və qabaqcıl analitik üsulların mövcud olması neft-qaz şirkətləri üçün qərar qəbuletmənin əsas komponentinə çevrilmişdir. Lazım olan vaxtda qərar qəbul edən şəxslərin əlində olan düzgün verilənlərin köməyi ilə aşağıdakı nəticələrə nail olmaq olar: qazma dəqiqliyi və təhlükəsizliyin artırılmasına; qeyri-ənənəvi mənbələrdən neft çıxarılmasının gücləndirilməsinə; təhlükəsizliyin yaxşılaşdırılması və risklərin aradan qaldırılmasına; məhsuldarlığın artırılmasına; ilkin neftin əldə olunmasını sürətləndirməyə; qurğuların nasazlığını əvvəlcədən bilmək və aradan qaldırmağa; yeni quyuların təchizatının təkmilləşdirilməsinə; daha dəqiq proqnozlaşmadan gəlirlərin əldə olunmasına və s.

ƏDƏBİYYAT

[1] Feblowitz J., The Big Deal About Big Data in Upstream Oil and Gas, Paper & presentation, IDC Energy Insights, October 2012

[2] Unified Data Solution for Oil & Gas, http://www.hrgresearch.com/

[3] Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001. http://blogs.gartner.com/

[4] Big Data In Big Oil: How Shell Uses Analytics To Drive Business Success [http://www.forbes.com/sites/teradata/2015/09/15/the-big-data-management-challenge-for-oil-gas/]

[5] Murtaza H. Beyond the hype: Big data concepts, methods, and analytics // International Journal of Information Management, vol. 35, 2015, pp. 137–144.

[6] Amir S., Atle B. Knowledge Based Oil and Gas Industry, Knowledge Based Oil and Gas (research report 2011)

[7] http://oslo.spe.org/bigdata

[8] http://www.ecim.no/conference/ecim2015/

[9] Brulé M., 2013, Tapping the power of Big Data for the oil and gas industry, IBM Software White Paper for Petroleum Industry, May 2013

[10] Robert K.Perrons, Adam H. Cloud computing in the upstream oil & gas industry: A proposed way forward // Energy Policy, vol. 56, 2013, pp.732–737.

[11] X. Wu, X. Zhu, G.Q. Wu, W. Ding, “Data mining with bigdata” // IEEE Transactionson Knowledge and Data Engineering, 2014, vol.26, issue 1, pp.97–107.

[12] Baaziz A., Quoniam L., 2013, The information for the operational risk management in uncertain environments: Case of Early Kick Detection while drilling of the oil or gas wells, International Journal of Innovation and Applied Studies (IJIAS), Vol. 4 No. 1, Sep. 2013

[13] Seshadri M., 2013, Big Data Science Challenging The Oil Industry, CTO Global Services, EMC Corporation, March 2013

[14] Exploring Oil Data: A Reporter's Handbook. http://openoil.net/exploring-oil-data/

[15] Steed A., Ricciuto D.M., Shipman G., etc. Big data visualanalytics for exploratory earth system simulation analysis Chad // Computers &Geosciences, 2013,vol.61, pp.71–82.

[16] Bertocco R., Padmanabhan V. Big Data analytics in oil and gas: Converting the promise into value . http://www.bain.com/

[17] Gundersen O.E., Boe L.J. Successful implementation of Big Data and Advanced analytics in O&G, (Predictive Real-Time Analytics Drives Decision Support for Oil Well Drilling,

[18] Evensen O. E. Big Data in Oil & Gas for Improved Decision Making and Operational Efficiency, IBM

[19] Hems A., Soofi A., Perez E., 2013, Drilling for New Business Value: How innovative oil and gas companies are using Big Data to out maneuver the competition, A Microsoft White Paper, May 2013

[20] Hollingsworth J., 2013, Big Data For Oil & Gas, Oracle Oil & Gas Industry Business Unit, March 2013

[21] Panorama des solutions de BIG DATA: Hadoop et son écosystème, Journal du Net, http://www.journaldunet.com/, 2013.

Ali Təhsildə Big Data Texnologiyası Tətbiqinin Üstünlükləri

Havar Məmmədov[1], Zəfər Cəfərov[2], İqbal Babayev[3]
[1,2,3]Azərbaycan Texniki Universiteti, Bakı, Azərbaycan
[2]*c.zafar@mail.ru*

***Xülasə* — Təhsil böyük həcmli informasiya istifadə və hasil olunan sahələrdən biridir. Məqalədə ali təhsil müəssisəsi üzrə Big Data texnologiyasının tətbiqi zərurətləri və üstünlükləri təhlil olunur. Göstərilir ki, bu texnologiya tədrisin səmərəli təşkili və keyfiyyətinin yüksəldiməsi üçün faydalı ola bilər.**

Açar sözlər — böyük verilənlər, verilənlərin analizi, təhsil analitikləri

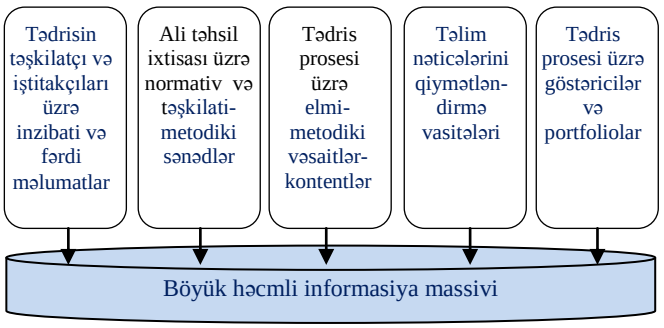
I. GİRİŞ

"Big Data" təhsilin informasiyalaşdırılması sahəsində perspektivli istiqamətlərindən biridir. Bu texnologiya ilə məşqul olan qabaqcıl şirkətlər (Knewton, Silicon Graphics və s.) ilk növbədə təhsil sahəsində böyük həcmli verilənlərin intellektual təhlili üzrə platformalar yaratmışlar [1].

Hazırda təhsil müəssisələrində əsasən MS Excel və ya Data Base cədvəllərində strukturlaşdırılan kiçik həcmli informasiya işlədilir. Buna səbəb xüsusi informasiya-kommunikasiya mühitin yaradılmaması və mövcud verilənlərin əksəriyyətinin (90%) struktursuz informasiya təşkil etməsidir. Universitetlər isə təhlil edilməli böyük həcmə və müxtəlif tərkibə malik məlumatlar hasil və istifadə olunan müəssisələrdəndir. Bu müəssisələr tədrisin səmərəli təşkili və keyfiyyətinin yüksəldiməsi məqsədilə həmin məlumatlardan faydalı biliklər əldə etmək üçün böyük elm və təhsil potensialına malikdirlər. Müasir informasiya emalı və yaddaş texnologiyaları böyük həcmli məlumatları toplamaq və təhlil etməyə imkan verir. «Big Data» istiqaməti müəssisənin səylərini bu məlumatların saxlanması, emalı və təhlininin təşkilinə konsentrasiya edir. Ona görə də "Big Data" texnologiyasının təhsil sahəsinə tətbiqinin araşdırılması aktual hesab edilir.

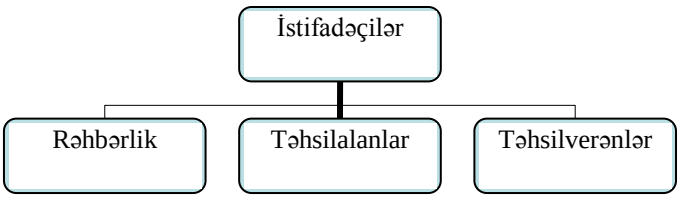
II. TƏHSİL SAHƏSİ ÜZRƏ BÖYÜK VERİLƏNLƏRİN FORMALAŞMASI

Təhsil insan fəaliyyətinin daha cox informasiya resursları toplanan sahəsidir (şək. 1). Təhsilin təşkili və planlaşdırılması üzrə standartların, normaiv, təşkilati-metodiki sənədlərin, elmi və tədris-metodiki kontentlərin, təlim nəticələrinin qiymətləndirmə vasitələrinin, tədrisin təşkilatçı və iştitakçıları üzrə inzibati və fərdi məlumatların hazırlanması, təhsilalanların bir neçə il həftədə beş və ya altı gün olmaqla çoxlu sayda təlim saatları üsrə tədris göstəricilərinin qeydiyyatı, təhsilalan və təhsiverənlərin portfoliolarının tərtibi və s. böyük həcmli informasiya massivlərinin - «Böyük verilənlər»-in yaradılmasına səbəb olur.



Şəkil 1. Böyük həcmli informasiya massivinin formalaşması

Təhsil sahəsində «böyük verilənlər» adətən üç kateqoriya istifadəçilər - təhsilalanlar, təhsilverənlər və təhsil müəssisəsinin rəhbərliyi tərəfindən formalaşdırlır və istifadə olunur (şəkil 2).



Şəkil 2. İnformasiya folmalaşdırıcı və istifadəçiləri

Rəhbərlik ali təhsil ixtisası üzrə inzibati, normativ sənədlər, standart lar, plan və proqramlar işləyib hazırlayır.

Təhsilverənlər təlim fəaliyyətinin cari və perspektiv planlaşdırılması üzrə təşkilati-metodiki sənədləri, elmi və tədris-metodiki kontentləri, təlim nəticələrinin qiymətləndirmə vasitələrini tərtib edirlər.

Təhsilalanlar məxsusi biliklərini – hesabatlarını təqdim edərək «Böyük verilənlər»in formalaşmasında fəall iştirak edirlər.

III. BÖYÜK VERİLƏNLƏR TEXNOLOGİYASININ TƏHSİLDƏ TƏTBİQİ

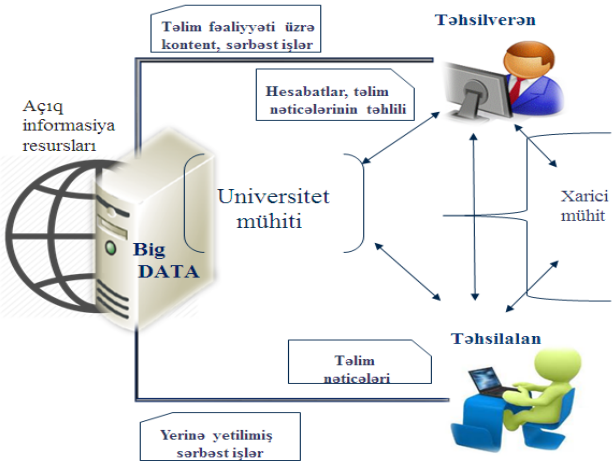
«Böyük verilənlər» anlayışı müxtəlif mənbələrdə yerləşərək mütəmadi yeniləşən fərqli tərkibə və formata malik iri həcmli informasiyanı təsvir edir. «Böyük verilənlər» texnologiyasnının mahiyyəti məhz bu cür informasiyanın təhlil olunmasından ibarətdir [2].

Big Data texnologiyası müxtəlif formata malik böyük həcmli informasiyanın istifadəçilərə ani (sürətli) təqdim

olunmasını təmin edir (şəkil 3). Bu isə cari təlim fəaliyyəti üzrə operativ zəruri məlumat mübadilə etmək və yüksək kommunikabellik imkanları yaradır. Təhsilalanlar Universitetin təqdim etdikləri biliklərlə yanaşı açıq informasiya (şəbəkə) resurlarından istifadə edir [3].

Şəkil 3. Böyük həcmli informasiya üzrə tələblər

Tədrisin keyfiyyətini yüksəltmək, təlim-tərbiyə işlərinin monitorinqin aparmaq məqsədilə Universitet üzrə «Böyük verilənlər» bazasında tədrisin bütün elementlərini əlaqələndirən, geniş funksional imkanlara malik elektron informasiya-təhsil mühititi yaratmaq olar (şəkil 4).



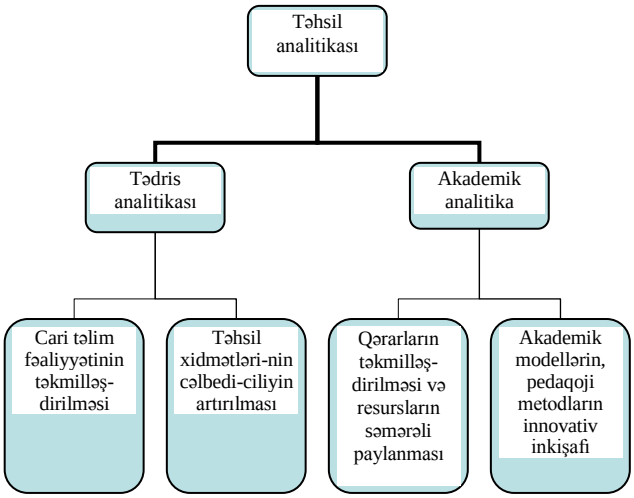
Şəkil 4. Böyük həcmli informasiya üzrə tələblər

Universitet üzrə toplanmış boyük həcmli verilənlər iki istiqamdə təhlil oluna bilər (şək. 5).

1) *Tədris analitakası*
2) *Akademik analitika*

Tədris analitakası təlim fəaliyyəti üzrə aparılır:

- ✓ Təhsilalanlar arasında fərdiləşdirmə, risk qruplarının müəyyənləşdirilməsi, təlim nəticəlrinin müstəqil və kollegial qiymətləndirilməsi, müvəffəqiyyətin proqnozu, "aqillı" tədris planı, proqramları və s. vasitəsilə cari təlim fəaliyyətinin təkmilləşdirilməsi.
- ✓ Yekun attestasiyadan sonra iş və təhsilin davam etdirilməsi üzrə konsaltinq və tövsiyə vasitəsilə təhsil xidmətlərinin cəlbediciliyin artırılması.



Şəkil 5. Təhsildə analitik istiqamətlər

Akademik analitaka idarəetmə prosesi üzrə aparılır:

- ✓ Təhsilalan və təhsilverənlərin bilik dövriyyəsinin və profillərinin daha dəqiq ölçmə vasitələrilə inzibati qərarların təkmilləşdirilməsi və təşkilati resursları səmərəli paylanması
- ✓ Müqaisəli təhlili əsasında akademik modellərin, pedaqoji metodların innovativ inkişafı

Təhsilverənin fəaliyyəti bütövlükdə ali təhsil proqramı üzrə təhsilalanların təlim nəticələrinin müvəffəqiyyətindən asılı olaraq qiymətləndirilir. Analitiklər təhsil proqramı üzrə hər bir tədris işştirakçısının təlim fəaliyyətini qiymətləndirməyə imkan verir [4].

Analitiklər təhsilverənlər üçün aşağıdakı imkanlar yaradırlar:

- ❖ təhsilalanların təlim fəaliyyəti üzrə tam məlumat almaq
- ❖ təhsilalanların fərdi və qrupdaxili təlim fəaliyyəti üzrə məlumat. almaq
- ❖ təhsilalanların təhsil proqramının müxtəlf mərhələlərdəki təlim nticələri haqqında informasiya əldə etmək
- ❖ təhsil proqramı üzrə digər təhsilverənlərlə qarşılıqlı əlaqə yaratmaq
- ❖ təlim nəticələrinin yüksəltmək məqsədilə təhsilalanllara qarşılıqlı əlaqə yaratmaq
- ❖ təhsilalanlların fərdi xüsusiyyətlərini araşdırmaq
- ❖ təhsilalanların təlim nəticələri üzrə müvəffəqiyyət göstəricilərini təhlil etmək
- ❖ təlim nəticələri üzrə universitet və auditoriya mühitinin əhəmiyyətini qiymətləndirə bilər.

Analitiklər təhsilalanlar üçün üçün aşağıdakı imkanlar yaradırlar:

- təlim nəticələrində maraqlı olduqları üçün təhsilalanlar təhsilverənlərin qiymətləndirmələrindən əlavə peşə standartları, sertifikat tələbləri, uğurlu karyera nümunələri kimi alternativ metodlardan istifadə etmək
- fərdi təlim metoduna üstünlük verdikləri üçün müntəzəm monitorinq aparmaq.
- məxsusi təhsil traektoriyası yaratmaq

NƏTİCƏ

Beləliklə, qeyd olunan mülahizələrə görə «Böyük verilənlər» təhsilalanlara müvəffəqiyyətin monitorinqinin aparılmasına və təhsil proqramlarını fərdiləşdirməyə imkan verir. Təhsilverənlər təhliledici vasitələrlə proqnozlaşdırma əsasında təlim metodkasınını təkmilləşdirilməsi imkanı əldə edirlər. « Böyük verilənlər»in təhlili əsasında rəhbərlərə təhsil proqramlarının rəqabət qabilyyətliyinin yüksəlməsi və onların uğurlu nəticələnməsinin proqnazlaşdırılması üzrə tövsiyyəvi metodika təqdim olunur

ƏDƏBİYYAT

[1] Шилина М.Г. Ридер по курсу Big Data: теория и методология современной коммуникации. Электронный носитель. М., 2014.

[2] Кореньков В.В., Ужинский А.В. Система мониторинга сервиса передачи данных (FTS) проекта EGEE/WLCG. Вычислительные методы и программирование: новые вычислительные технологии, том 10, 2009, с.96-100.

[3] 2. Mohanty S. Big Data Imperatives / S. Mohanty, M. Jagadeesh, H. Srivatsa - Apress, 2013, -311 c.

[4] 3. The Digital Universe In 2020: Big Data - Электрон. Дан – Режим доступа : http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

E-dövlət Sistemində “Big Data”: İmkanlar və Perspektivlər

Fərhad Yusifov
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
farhadyusifov@gmail.com

Xülasə **– “Big data” anlayışı, mövcud konseptual yanaşmalar, bu texnologiyanın imkanları, tətbiq sahələri və perspektivləri tədqiq olunur. E-dövlət sistemində formalaşan böyük həcmdə verilənlərin analizində “Big data” texnologiyalarının tətbiqi imkanları qiymətləndirilir.**

Açar sözlər – e-dövlət, açıq dövlət, açıq verilənlər, böyük verilənlər, bulud infrastrukturu

I. GİRİŞ

İnformasiya cəmiyyətinin formalaşması dövlət idarəçiliyi sistemində informasiya resurslarının istifadə imkanlarına dair mövcud təsəvvürləri əhəmiyyətli dərəcədə dəyişir. Açıq şəkildə təqdim olunan aktual, etibarlı informasiya hakimiyyət orqanları arasında informasiya mübadiləsinin təmin olunması şəraitində dövlət idarəçiliyi sisteminin “mühərrikinə” çevrilir.

Bununla yanaşı, informasiya cəmiyyətinə keçid dövlət idarəçiliyi sisteminin funksionallığına dair baxışları da dəyişmişdir. İKT-nin inkişafı, elektron dövlət (e-dövlət) quruculuğu dövlət idarəçiliyi sisteminə və siyasi proseslərə ciddi şəkildə təsir göstərməkdədir. Hazırda qabaqcıl texnologiyaların inkişafı və dövlət xidmətlərinə qoyulan tələblər e-dövlətin funksiyalarını transformasiya etməkdədir. Səmərəli dövlət idarəçiliyi mexanizmlərinin formalaşdırılması birbaşa demokratiyaya hədəflənən yeni e-dövlət konsepsiyasının yaradılmasını zəruri edir.

Hazırda açıq və şəffaflığı təmin etməyə imkan verən qabaqcıl texnologiyalar əsasında e-dövlətin yeni elektron idarəetmə mexanizmlərinin yaradılması ideyası həyata keçirilməkdədir. Bu prinsiplərin məntiqi ardıcıllığı kimi açıq verilənlər (ing. - *open data*, alm.- *offene daten*) göstərilir [1]. Açıq verilənlər e-dövlətin inkişafının əsasını təşkil etməklə, vətəndaş üçün dövlət (*G4C - government for citizens*) prinsipinin, eləcə də, açıq dövlət (open government) konsepsiyasının həyata keçirilməsinə imkan verir [1,2].

Açıq verilənlər ideyası böyük dövlətlər və beynəlxalq təşkilatlar tərəfindən dəstəklənir (*W3C, Open government partnership, Open knowledge foundation* və başqaları) [2]. Hazırda dövlətlər tərəfindən böyük həcmdə verilənlər İnternet üzərindən yayılır. Son dövrlərdə “böyük verilənlər”in (*Big data*) açıq verilənlər (*open data*) formatında emalı və təqdim olunması imkanı yaranmışdır. 2000-ci illərdə meydana çıxmış “Big data” termini ingilis dilindən Azərbaycan dilinə tərcümədə “böyük verilənlər” kimi başa düşülür. “Big data” termini struktrulaşdırılımamış, həcm baxımından mövcud intellektual analiz vasitələri ilə emal oluna bilməyən verilənləri təyin etmək üçün istifadə edilir [3,4].

Hakimiyyət orqanlarının fəaliyyətində və yerli özünüidarəetmə orqanlarında islahatların aparılması səmərəli idarəetmə sisteminin yaradılmasına yönəlmişdir. Hazırda səmərəli idarəetmə mexanizmlərinin formalaşdırılmasında “böyük verilənlər”dən istifadə imkanının olması xüsusi əhəmiyyət kəsb edir. İnformasiya texnologiyalarının inkişaf səviyyəsi “böyük verilənlər”lə iş üçün hakimiyyətin təşkilinin yeni və səmərəli sistemini təklif edir. 2014-cü ildə “Böyük verilənlərin böyük imkanları” (*Global Entrepreneurship Congress*) konfransı çərçivəsində “bulud texnologiyaları”nın (*Cloud technology*) istifadəsinə dair nümunələr təqdim olundu və əsas diqqət verilənlərin açıq, əlyetərəliliyin təmin olunmasında dövlət orqanlarının roluna yönəldildi [5].

Tədqiqat işində “Big data” konsepsiyasına dair baxışlar tədqiq olunur. Dövlət idarəçilinin səmərəliliyin artırılmasında “Big data” texnologiyalarının tətbiqi imkanları araşdırılır.

II. “BİG DATA” KONSEPSİYASI

“Big data” konsepsiyasına həsr olunan elmi-tədqiqat işləri, nəşr olunan məqalələr hazırda bu problemə kifayət qədər marağın olduğunu və geniş auditoriya tərəfindən müzakirə olunduğunu göstərir. Tədqiqatçıların əksəriyyətinin firkirinə görə “Big data” verilənlərlə işi - informasiya texnologiyalarının və innovasiyaların inkişaf konsepsiyası kimi baxılmalıdır. “Big data” anlayışını sadə halda böyük həcmdə verilənlərin saxlanması, emalı və analizi kimi başa düşülür. Ədəbiyyatlarda “Big data” anlayışının tam izahı “3V” kimi adlandırılır və aşağıdakı kimi şərh olunur [4,6]:

1. Böyük həcmə malik verilənlər (volume);
2. Yüksək templə artan verilənlər (velocity);
3. Müxtəlif mənbədən və fərqli formatlarda daxil olan verilənlər (variety).

Qeyd etmək lazımdır ki, analitiklər bəzən “Big data” anlayışını tam izahı “5V” kimi təsvir edilən həqiqlilik (*veracity*) və dəyər (*value*) xarakteristikalarını da daxil edirlər [3,7].

Bununla belə ədəbiyyatda verilən şərhlərin, anlayışların heç biri “böyük verilənlər”i tam şəkildə izah etməyə imkan vermir. “Böyük verilənlər”i ayırd etmək üçün adətən iki kəmiyyət parametri əsas götürülür:

I. Strukturlaşdırılan verilənlərin həcmi 100 terabayta çatmalı,
II. Daxil olan verilənlər axınının surəti saniyədə 60 Gbit olmalıdır.

[1] işində dövlət idarəçiliyində “Big data”nın istifadəsi məsələsinə baxılır. “Big data” sahəsində mövcud vəziyyət, imkanlar və perspektivlər araşdırılır.

[8] tədqiqat işində “Big data” analyışı şərh olunur və dövlət sektoruna uyğun “Big data” tipologiyası təklif olunur. E-dövlət xidmətlərinin fərdiləşdirilməsinin və səmərəliliyin artırılması üçün biznes analitikada “Big data”nın imkanlarından istifadə olunması araşdırılır. Eləcə də, müvafiq idarəetmə strukturlarında, müxtəlif məlumat mənbələrinin inteqrasiyası, rəqəmsal konfidensiallığı və təhlükəsizlik riski məsələlərində “Big data”nın tətbiqi və menecmenti məsələləri müzakirə olunur. Dövlət idaərçiliyi üçün “Big data” mövzusunda gələcək tədqiqat istiqamətləri müəyyənləşdirilir.

[9] işində “Big data” texnologiyası araşdırılır, onun üstünlüyü və çatışmazlıqları göstərilir. İqtisadiyyatda “Big data”nın tətbiqi imkanları və perspektivləri tədqiq olunur. Eləcə də, e-dövlət sistemində idarəedici qərarların qəbulunda “Big data”nın istifadəsi və dövlət orqanlarında “Big data” texnologiyalarının inkişaf perspektivləri araşdırılır.

[10] işində e-dövlət mühitində “böyük verilənlər”in saxlanması və emalı üçün bulud infrastrukturunun formalaşdırılması məsələsinə baxılır. Məqalədə bulud infrastrukturunun layihələndirilməsinə dair mövcud konsepsiyalar analiz olunur. E-dövlət mühitində “böyük verilənlər” sahəsində layihələrin həyata keçirilməsi, mövcud vəziyyət və perspektivlər araşdırılır.

“Big data” böyük həcmdə strukturlaşdırılmayan verilənlər massivini emal etməyə imkan verən texnologiyadır. Hazırda həcm, sürət, müxtəliflik, dəyər və həqiqilik kimi xarakteristikalara malik verilənlərin emalı və analizi məqsədilə yaradılmış *MapReduce, Hadoop, NoSQL, Data Discovery* və s. texnologiyalar mövcuddur [6,7,11,12]. Hazırda bu texnologiyaların tətbiqi və inkişaf etdirilməsi funksionallığa və müqayisə olunmayacaq dərəcədə səmərəliliyə nail olmağa imkan verir.

Wikibon analitiklərinin proqnozuna görə yaxın 5 ildə qoyulan sərmayənin həcmindən təxminən 7 dəfə çox gəlir gətirməklə “Big data” ən sərfəli investisiya sahələrindən birinə çevriləcək [13].

Hazırda bir çox ölkələr, o cümlədən ABŞ-da, Avstraliyada “böyük verilənlər” sahəsində investisiya qoyuluşuna dair xüsusi dövlət proqramları vardır. 2012-ci ildə ABŞ Prezident administrasiyası “böyük verilənlər” sahəsinə 200 mln. ABŞ dolları həcmində səmayə qoymaqla müəssisələrin böyük həcmli rəqəmsal verilənlərə çıxışı və emalı üçün zəruri olan metod və vasitələrin səmərəliliyin artırılmasını hədəf seçmişdir [14,15]. Öz növbəsində Avstraliya hökuməti 2009-cu ildə böyük həcmdə məlumatların analizi məqsədilə “Avstraliya dövlət xidmətinin böyük verilənlər strategiyası” (*The Australian Public Service Big Data Strategy)* adlı sənəd qəbul etmişdir [16]. Bu sənəddə dövlət orqanlarının işində “Big data” metodlarının tətbiqi mərhələləri verilmişdir. Avstraliya hökuməti “verilənlər”i vətəndaşların həyat səviyyəsinin yüksəldilməsində istifadə etməyi əsas hədəf kimi seçmişdir.

III. E-DÖVLƏT MÜHİTİNDƏ “BÖYÜK VERİLƏNLƏR”İN İSTİFADƏSİ

E-dövlət sistemləri mürəkkəb struktura malik olmaqla vətəndaşlara və biznes sektoruna müxtəlif xidmətlər göstərir. Bu sistemlərdə verilənlərin həcmi terabaytlarla ölçülür və bu səbəbdən “böyük verilənlər” bu gün üçün böyük aktuallıq kəsb edir. Vətəndaşlara və biznes sektoruna səmərəli xidmətlərin təklif olunması və gələcək üçün effektiv tətbiqlərin yaradılması baxımından e-dövlət layihələrinin həyata keçirilməsində “böyük verilənlər” texnologiyasının tətbiqi imkanları nəzərdən keçirilməlidir.

Hal-hazırda böyük şirkətlərdə, o cümlədən Adobe Amazon, Yahoo, Facebook və s. uğurla tətbiq olunan “Big data” layihələri vardır [10,11]. Bu şirkətlər tərəfindən baxılan bir çox məsələlər e-dövlət mühitinə də tətbiq oluna bilər. “Böyük verilənlər” və “Hadoop” artıq bir çox sahələrdə, o cümlədən verilənlərin saxlanması, tibb, təhsil, ticarət, energetika, loqistika, səyahət, maliyyə xidmətləri, siyasət və s. istifadə olunur. Təcrübədə tətbiq olunan “Big data” məsələlərindən bəziləri, xüsusilə e-dövlət üçün böyük əhəmiyyət kəsb edir.

Sosial şəbəkələrin populyarlığını nəzərə alaraq sosial şəbəkələrin analizi olduqca mühüm məsələdir. Geniş imkanlara malik analitik alətlər sosial mediada strukturlaşdırılmamış verilənləri analiz etməyə və konkret məsələyə uyğun istifadəçi marağını, profilini müəyyən etməyə imkan verir. Məsələn, müxtəlif marketinq kampaniyalarının səmərəliliyini müəyyən edilməsi üçün “böyük verilənlər”in analizində istifadə oluna bilər [10].

Bir başqa sahə elektron tibbdir və “böyük verilənlər”in analizi əhəmiyyətli dərəcədə bu sahənin inkişafına təsir göstərə bilər. Tibbi məlumatların yadda saxlanması və analizindən başqa “Big data” analitikası tibb müəssisələrinə, klinikalara xəstələrə fərdiləşdirilmiş xidmətlərin göstərilməsinə kömək edə bilər [17].

Həyata keçirilən layihələrdən biri “ağıllı” şəhər (*smart city*) konsepsiyası intellektual idarəetməyə, İKT-nin inteqrasiyasına əsaslanır və vətəndaşların aktiv iştirakını tələb edir. Müxtəlif mənbələrdən emal olunan böyük həcmdə verilənlər “Big data” texnologiyasının istifadəsini zəruri edir [10].

Bununla yanaşı, “böyük verilənlər”in mövcud olmasını təsdiqləyən çox sayda misallar (avto-naviqasiya, e-ödəniş, smart dayanacaq və s.) göstərilə bilər və bu proseslərə “Big data” texnologiyalarının tətbiqi səmərəliliyin artırılmasında mühüm amil hesab olunur [18]. Dövlət təhlükəsizliyi xidmətləri təsvir və videoların emalı üçün “böyük verilənlər”in analizindən istifadə edə bilər ki, bu xüsusilə intellektual nəqliyyat sistemlərində təhlükəsizliyin təmin olunması üçün əhəmiyyətli olar.

“Big data” texnologiyalarının tətbiqi bütövlükdə həyata keçirilən e-dövlət layihələrinin səmərəliliyini artıra bilər. Burada mühüm aspektlərdən biri xərclərin azaldılması olsa da, digər mühüm aspekt isə vətəndaşla biznes sektoru və dövlət orqanları arasında qarşılıqlı informasiya mübadiləsinin səmərəliliyinin artırılması hesab olunur.

E-dövlət layihələrində “böyük verilənlər”in istifadəsini aşağıdakı kimi xarakterizə etmək olar:

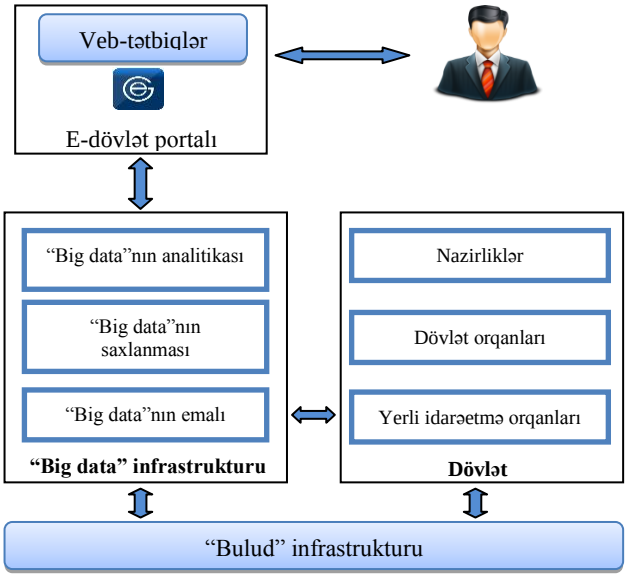
Məqsəd: E-dövlət portalında formalaşan böyük verilənlərdən istifadə və istifadəçilərin profilinin analizi.

Əhəmiyyəti: E-xidmətlərə çıxışın və səmərəliliyinin artırılması; e-dövlət portalının istifadəçiləri ilə əks əlaqənin yaradılması; istifadəçilərə göstərilən dövlət xidmətlərinin genişləndirilməsi.

Verilənlərin mənbələri: e-dövlət portalı; vətəndaşların elektron müraciətləri; informasiya mərkəzlərinə edilən müraciətlər, mesajlar və zənglər; sosial şəbəkələr və digər mənbələr.

Analizin nəticələri: e-xidmətlər haqqında müxtəlif məlumatlar (məsələn, reytinqlər, səmərəlilik, tələb edilən xidmətlər və s.); əldə olunan verilənlər əsasında dövlət xidmətlərinin səmərəliliyinin artırılması üçün tövsiyələrin işlənməsi; əks əlaqə mexanizminin qurulması.

Ədəbiyyatda “böyük verilənlər” infrastrukturunun konseptual modelinə dair müxtəlif yanaşmalar mövcuddur (Şəkil 1) [10].



Şəkil 1. “Böyük verilənlər” infrastrukturunun konseptual modeli

Vətəndaşlar və müəssisələr e-dövlət portalı vasitəsilə xidmətlərdən istifadə edir. Vətəndaşların sorğuları emal üçün “Big data” infrastrukturuna yönəldilir. “Big data” infrastrukturunun mərkəzləşməmiş verilənlər bazası vardır. Nazrliklərin, dövlət və yerli özünüidarəetmə orqanlarının verilənlər bazaları ilə inteqrasiya həyata keçirilir. “Big data”nın emalından sonra vətəndaş (istifadəçi) e-dövlət portalı vasitəsilə sorğuya uyğun cavab və visallaşdırılmış məlumat almış olur.

NƏTİCƏ

İnformasiya xidmətlərinin və texnologiyalarının, xüsusilə mobil rabitə, əşyaların interneti, sosial media, e-xidmət sahələrinin sürətli inkişafı böyük həcmdə struturlaşdırılmamış verilənlərin yaranmasına səbəb olmuşdur. Bu verilənlərin analizi yalnız biznes sektoru üçün deyil, eləcədə elm, dövlət idarəçiliyi, təhsil və bütöblükdə cəmiyyət üçün böyük əhəmiyyət kəsb edir. E-dövlət sistemlərində formalaşan böyük həcmdə verilənləri ənənəvi üsullarla emal etmək bu gün üçün praktiki olaraq mümkün deyil. “Big data”nın düzgün və ardıcıl istfadəsi e-dövlət sisteminin səmərliliyinin artırmaqla yanaşı, ümumilikdə vətəndaş cəmiyyətinin inkişafının yüksək səviyyəsinə nail olmağa imkan verəcək.

Hazırda “Big data” analizi üçün proqram həllərinin və alətlərinin inkişaf etdirilməsinin əsasını şəbəkə infrasturukturu təşkil edir. Şəbəkə infrastrukturu isə öz növbəsində verilənlərin saxlanması və emalı üçün etibarlı, təhlükəsiz mühitin formalaşmasını təmin etməlidir.

ƏDƏBİYYAT

[1] Е.В. Булгакова, В.Г. Булгаков, В.С. Акимов, Использование “больших данных” в системе государственного управления: условия, возможности, перспективы, Журнал Юридическая наука и практика: Вестник Нижегородской академии МВД России. Выпуск № 3 (31), 2015, стр. 10-14

[2] А.М. Игнатова, Открытые данные как новый способ взаимодействия государства и общества, Вопросы теории и практики, № 1 (51), 2015, стр. 78-80, http://scjournal.ru

[3] R.M. Əliquliyev, M.Ş. Hacırəhimova, “Big Data” fenomeni: problemlər və imkanlar, İnformasiya texnologiyaları problemləri, 2014, №2, s. 3-16

[4] J.S. Ward and A. Barker, Undefined By Data: A Survey of Big Data Definitions. Retrieved from http://arxiv.org/pdf/1309.5821.pdf

[5] Global Entrepreneurship Congress, Retrieved from www.gec.co

[6] Д.В. Кондратов, С.Б. Можейко, Концепция Big Data: проблемы и перспективы, Местное самоуправление в системе публичной власти, 2015, стр. 88-89

[7] What is big data? - Bringing big data to the enterprise, 2013. Retrieved from www-01.ibm.com

[8] S. Kalbande, S. Deshpande and M. Popat, Review paper on use of big data in e-governance of India, International journal for research in emerging science and technology, Vol. 2 (1), 2015, pp. 395-403

[9] Т.В. Губареваа, Е.И. Луковникова, Особенности применения новых информационных технологий Больших данных, Проблемы социально-экономического развития Сибири, 2015.

[10] J. Šuh, V. Vujin, D. Barać, and et al. Designing cloud infrastructure for big data in e-government, Journal of Universal Excellence, No 1, 2015, pp. A26–A38.

[11] Hadoop Wiki PoweredBy. Retrieved from http://wiki.apache.org/hadoop/PoweredBy

[12] Worldwide Big Data Technology and Services 2013–2017 Forecast, Retrieved from www.idc.com

[13] http://wikibon.org/wiki/v/Big_Data

[14] Д.Ю. Яблоков, Применение технологий больших данных в налоговом администрировании, Региональное развитие, № 2(6) 2015

[15] Obama administration unveils “big data” initiative: announces $200 million in new r&d investments, www.whitehouse.gov

[16] The Australian Public Service Big Data Strategy, www.finance.gov.au

[17] P. Groves, B. Kayyali, and et al. The big data revolution in healthcare, 2013, Retrieved from www.mckinsey.com

[18] The Case for Smarter Transportation - IBM Whitepaper, 2010, Retrieved from www-07.ibm.com

Hava Məkanında Münaqişə Vəziyyətinin Müəyyənləşdirilməsi Sistemlərində Big Data Texnologiyalarından İstifadə

Nadir Ağayev[1,2], Nəsrəddin İsgəndərov[3]
[1]Milli Aviasiya Akademiyası, Bakı, Azərbaycan
[2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[3]Azərbaycan Hava Yolları Qapalı Səhmdar Cəmiyyəti, Bakı, Azərbaycan
[1,2]nadir_avia@yahoo.com, [3]nesiazeri@yahoo.com

***Xülasə* — Məqalədə havadakı hərəkətin avtomatlaşdırılmış idarəetmə sistemlərində baş verən catışmazlıqların aradan qaldırılması üçün lambda arxitekturalı verilənlər bazasına malik Big Data texnologiyalarından istifadə edilməsi təklif edilmişdir**

Açar sözlər — hava gəmisi; münaqişə vəziyyəti; Big Data texnologiyaları

I. GİRİŞ

Müasir mülki aviasiyanın əsas məqsədi istənilən vəziyyətdə uçuşlarını təhlükəsizliyinin yüksəldilməsi üçün ən yeni texnologiyalar, o cümlədən süni intellektdən istifadə edən metod və texnologiyaları bu sahəyə tətbiq etrməkdir. Müasir uçuş aparatları üçün təhlükələr bir qayda olaraq:

a) xarici faktorlar - (bura əsasən meteoşəraitin dəyişmələri, havada hərəkət edən digər obyektlərin təsiri ilə yaranan təhlükələr daxildir),
b) texniki faktorlar - (bura hava gəmisinin texniki göstəricilərinin dəyişmələri nəticəsində yarana bilən təhlükələr daxildir);
c) insan faktoru - (bura uçuş zamanı pilot və ya dispetçerlərin fəaliyyəti ilə yaranan təhlükələr daxildir);

kimi kompleks faktorların təsiri ilə yaranan təhlükələrdir.

Ümumilikdə, normativ sənədlərə əsasən uçuş vəziyyətini normal vəziyyət və xüsusi vəziyyət olmaqla iki hissəyə ayırmaq olar. Xüsusi vəziyyətlər təhlükə dərəcələrinə görə 4 yerə bölünür [1]:

- Uçuş şəraitinin çətinləşməsi ;
- Mürəkkəb vəziyyət ;
- Qəzalı vəziyyət ;
- Faciəvi vəziyyət.

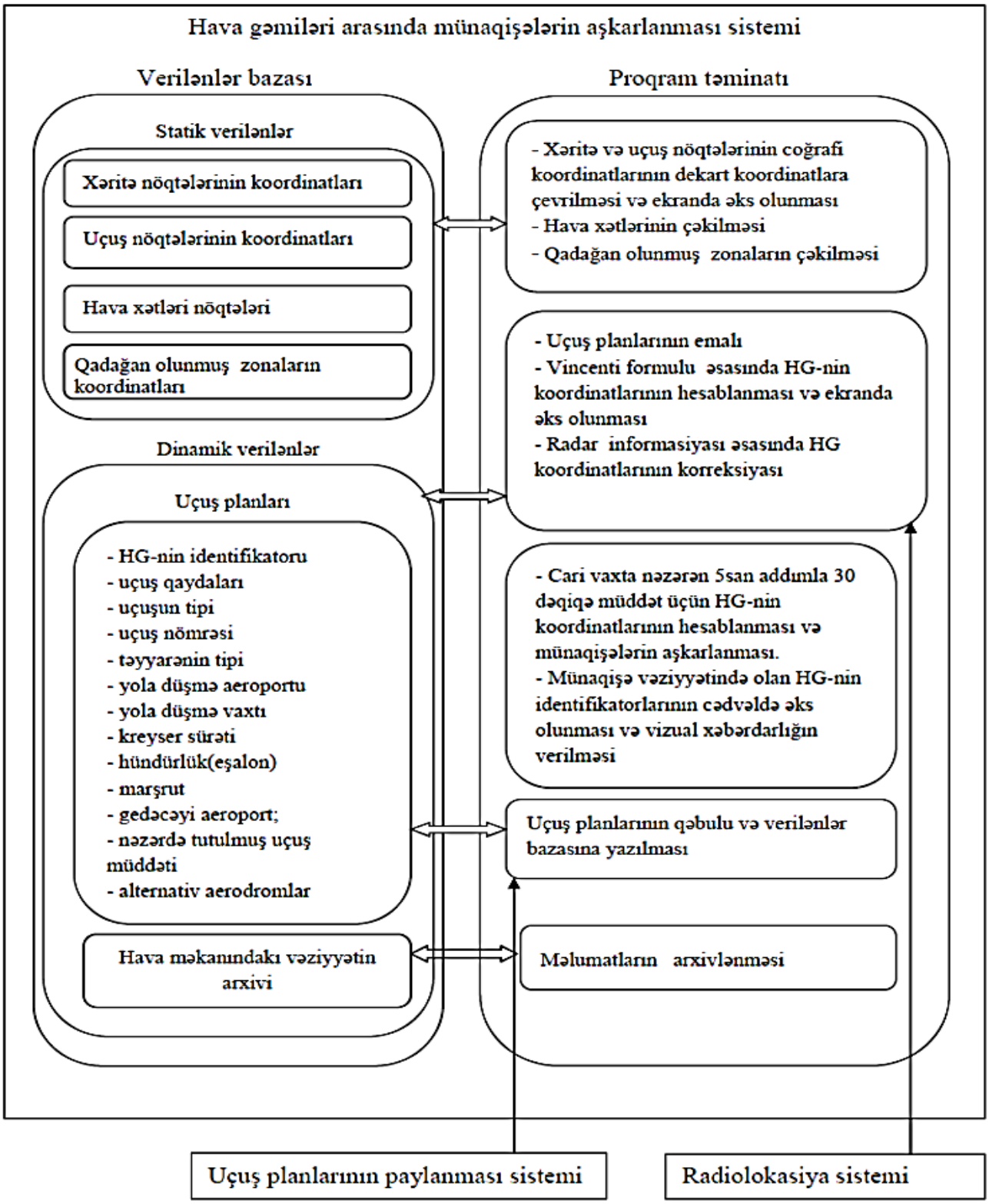
Bu halda uçuş vəziyyəti funksiyası vektor funksiya kimi qəbul edilir və aşağıdakı faktorlardan asılı olaraq formalaşır :

1) ekipajın psixofizioloji yüklənməsi;
2) stabillik və idarəetmə xarakteristikaları;
3) uçuş- texniki xarakteristikalar ;
4) uçuş parametrinin və şərtlərinin dəyişməsi ;
5) texniki vəziyyət göstəricilərinin dəyişməsi ;
6) uçuş planının dəyişilməsi ;
7) ekipaj üzvlərinin professional səviyyəsi ;
8) təyyarənin zədələnmə səviyyəsi ;
9) ekipaj və ya sərnişinlər arasında zədələnmə və ya ölüm hadisəsinin baş verməsi;

II. HAVADAKI HƏRƏKƏTİN MÜASİR AVTOMATLAŞDIRILMIŞ İDARƏETMƏ SİSTEMLƏRİ

Havadakı hərəkətin avtomatlaşdırılmış idarəetmə sistemlərinin (HHAİS) [2] əsas funksiyalarından biri bu faktorların hər birinin yaratdığı təhlükələri identifikasiya etməklə münaqişələri əvvəlcədən aşkarlayaraq uçuşların təhlükəsiz həyata keçirilməsini təmin etməkdir. Qeyd edilən faktorlar nəinki mənbələrinin müxtəlifliyinə, həm də təbiətləri etibarı ilə müxtəlif olurlar. Buna misal olaraq, uçuş zamanı xarici faktorların təsirinin qeyri-müəyyən və stoxastik dəyişmələri, operatorun fəaliyyətinin qeyri-səlisliyi, qurğuların və avtomatik idarəetmə sisteminin fəaliyyətinin determinik və ehtimallı dəyişmələrini və s. göstərmək olar [3]. Digər tərəfdən faktorların sayca coxluğu və dinamik dəyişmələri HHAİS-in bazasında qısa müddətdə terabaytlarla ölçülən informasiyaların toplanmasına səbəb olur. Bu informasiyaların bütövlükdə və qısa zaman müddətində emal edilərək əvvəlcədən məlum olmayan münaqişə vəziyyətinə uyğun olaraq doğru və operativ qərar qəbul edilməsi HHAİS qarşısında duran əsas vəzifələrdən biridir.

HHAİS-in tərkibində olan münaqişələrin qabaqcadan xəbərdarlıq alt sisteminin sxematik quruluşu şək.1-də göstərilmişdir. Qeyd edək ki, əvvəllər bu sistemlərdə yalnız radar məlumatından istifadə edərək qısa müddət üçün (STCA) [4,5] münaqişə vəziyyətinin müəyyələşdirilməsi nəzərdə tutulsa da, hazırki sistemlərdə həm də plan məlumatları əsasında hava gəmiləri (HG) arasında orta müddətli münaqişələri (MTCD) [6,7] aşkarlama funksiyasından istifadə olunur.



Şəkil 1. Münaqişələrin qabaqcadan xəbərdarlıq alt sisteminin sxematik quruluşu

Qeyd edilən sistem ənənəvi arxitekturaya malikdir. Bu sistemlərdə informasiyanın həcminin artması ilə əlaqədar yaranan çatışmazlıqlar aşağıda göstərilmişdir:

1) İmtinayadavamlığın azalması. Sistemdə istifadə olunan kompüterlərin sayı artdıqca, imtina hallarının sayı da artır. Bu halda işçi vəziyyətə qayıtmaq üçün sistemin əl ilə idarə edilməsi nəzərdə tutulur, imtinalar növbəli idarə olunur və s. Bu bütövlükdə sistemin imtinayadavamlığını azaldır.
2) Dayanıqlığın pozulması. Məlumatlar paylanmış (həm zamana, həm də məkana görə) olduğundan, sistem daxil olan hər bir məlumatın nə vaxt və haradan qəbul edildiyini dəqiq bilməlidir. Bu halda məlumatların həcminin artması onun bazaya qəbul edilməsində və bazadan oxunmasında imtinaların və səhvlərin artmasına səbəb olur.
3) Hesablama çətinliyi. İnformasiyanın həcmi artdıqca məhdud zaman ərzində verilənlər arasında bütün mümkün əlaqələri nəzərə almaqla klassik hesablama metodları köməyilə düzgün qərar qəbuletmə çətinləşir. Bu hal xüsusi ilə strukturlaşmamış verilənlərlə işlədikdə daha çox rast gəlinir.
4) Texniki xidmətin mürəkkəbliyi. İnformasiyanın həcmi artdıqca sistemdə texniki əlavələr (yaddaş həcminin artırılması, giriş-çıxış qurğularının yenilənməsi və s.) sistemə texniki xidməti çətinləşdirir. Vaxtında və lazımi həcmdə yerinə yetirilməyən texniki xidmət sistemdə müxtəlif imtinaların və səhvlərin alınmasına səbəb olur

Beləliklə hazırda istifadə edilən ənənəvi arxitekturaya malik sistem, informasiyanın həcminin sürətlə artımı şəraitində tələbləri ödəmir. Münaqişələri müəyyən edən və klassik yanaşma üsulları tətbiq etməklə qurulan hər hansı metod böyük həcmli informasiyalarla işlədikdə yuxarıda qeyd edilən səbəblərdən praktiki tələbləri ödəməyəcək. Bunu nəzərə alaraq biz Big Data texnologiyalardan [8] istifadə etməklə bu problemi həll etməyi təklif edirik.

III. HAVADAKI HƏRƏKƏTİN LAMBDA ARXİTEKTURALI AVTOMATLAŞDIRILMIŞ İDARƏETMƏ SİSTEMİ

Big Data texnologiyaları emal edilən informasiyanın həcminə görə əsasən üç hissəyə bölünür:

- Fast Data - informasiyanın həcmi ən çoxu terabaytla ölçülür.
- Big Data Analytics - informasiyanın həcmi petabaytlarla ölçülür
- Deep Insight - informasiyanın həcmi ekzabaytlarla və zettabaytlarla ölçülür

Hazırda informasiyanın həcminin ekzabayt və zettabaytla ölçüldüyü bir şəraitdə sistem emalın və qərar qəbuletmənin yüksək sürətini də təmin etməlidir. İnformasiyanın saxlanılması və emalında belə sistemlər- SAP HANA, Oracle Big Data Appliance, Oracle Exadata Database Machine, Oracle Exalytics Business Intelligence Machine, Teradata Extreme Performance Appliance, NetApp E-Series Storage Technology, IBM Netezza Data Appliance, EMC Greenplum, HP Converged Infrastructure bazasında Vertica Analytics Platform [9,10] yaradılmış və müxtəlif məsələlərin həllində istifadə edilir. Belə sistemlərdən biri də lamda arxitekturalı sistemdir Sistem istənilən təbiətli verilənlərdən istifadə etməklə istənilən funksiyanı (sorğunu) reallaşdırmaq qabiliyyətinə malikdir və üç təbəqədən ibarət olaraq formalaşır:

- Paket təbəqəsi
- Xidmət təbəqəsi
- Hesablama təbəqəsi

Sistem “Sorğu=F(bütün məlumatlar)” prinsipi ilə işlədiyindən böyük həcmli məlumatlar üçün hər bir yeni məlumat emal edilərək bazaya daxil edilməlidir. Bunun üçün bütün baza paketlərə bölünür və hər paketin xüsusiyyətləri (xarakteristikaları) indeksləşdirilir. Xüsusi üsullarla sorğu funksiyası qurulur və o paket funksiyaları ilə yoxlanılır (birölçülü halda bu indekslərin yoxlanılması kimidir) Baza paketlərə bölündükdə məlumatların elə dəyişməz xarakteristikaları götürülür ki, yeni məlumat daxil edildikdə əvvəlki məlumat dəyişilməz qalsın. Bu ilkin verilənlərin xüsusi emalını həyata keçirən və Data Mining texnologiyaları əsasında yaradılmış metod əsasında qurulan proqram təminatıdır. Bu proqram sorğu funksiyasını hesablayır. Sorğu funksiyası paket təbəqəsində yerləşən məlumatlar əsasında periodik dəyişilən “məlumatlar partiyası” funksiyası ilə yoxlanılır. ...”Məlumatlar partiyası” o vaxta qədər dəyişdirilir ki, onun funksiyası ilə sorğu funksiyası üst-üstə düşsün. Beləliklə, hesablama təbəqəsində sorğunun özü yox, onun rəqəmli analoqu olan funksiya vasitəsi ilə reallaşır. Bu sürətin artımına və nəticənin daha dəqiq və səhvsiz alınmasına səbəb olur.

Münaqişələrin aşkarlanması sistemlərində sorğular əvvəlcədən müəyyənləşdirilir. Bu International Civil Aviation Organization (İCAO) sənədlərinə görə HG hərəkətinə qoyulan tələblər əsasında formalaşır, məs. HG arasındakı məsafə, eşalonların ölçüləri, eyni eşalonda hərəkət zamanı sürət və istiqamətə qoyulan tələblər və s. Beləliklə, biz sorğuların çoxluğunu almış oluruq. Bu sorğular sistemə daxil olan məlumatlar (şəkil 1) əsasında hesablanmış paket funksiyası ilə yoxlanılır. Yoxlama alqoritmi hər hansı bir üsulla [5] yerinə yetirilə bilər. Beləliklə lambda arxitekturalı verilənlər bazası ilə işlədikdə yuxarıda qeyd edilən çatışmazlıqlar aradan qalxır:

- Sorğular real vaxt rejimində cavablandırılır
- Emalda geçikmələr bazaya yazma və oxumada cəld alqoritmlərin tətbiqi nəticəsində aradan qaldırılır.
- Sistem miqyaslaşdırma xüsusiyyətinə malikdir- yaddaşın həcmi yalnız əvvəlki yaddaş tamamilə istifadə edildikdə artırılır.

NƏTİCƏ

1) Müasir havadakı hərəkətin avtomatlaşdırılmış idarəetmə sistemləri tədqiq edilmiş, informasiyanın həcminin

sürətlə artımı şəraitində sistemdə yaranan catışmazlıqlar göstərilmişdir.

2) Müasir havadakı hərəkətin avtomatlaşdırılmış idarəetmə sistemlərində baş verən catışmazlıqların aradan qaldırılması üçün Big Data texnologiyalarından - lambda arxitekturalı verilənlər bazasının idarə edilməsi sistemindən (VBİS) istifadə edilməsi təklif edilmişdir.

3) Göstərilmişdir ki, Lambda arxitekturalı VBİS HG uçuşu zamanı münaqişələrin həll edilməsində İCAO tərəfindən qoyulan vaxt və dəqiqlik məhdudiyyətlərini ödəyir.

ƏDƏBİYYAT

[1] Руководство по управлению безопасностью полетов. ИКАО, Doc. 9859 AN/460, 2006

[2] Р.М. Ахмедов, А.А. Бибутов, А.В. Васильев и др. Автоматизированные системы управления воздушным движением: Новые информационные технологии в авиации: Учеб. пособие – СПб. Политехника, 2004, 446с.

[3] Hussein A. Abbass, Jiagnjun Tang, Rubai Amin, Mohamed Ellejmi, and Stephen Kirby The Computational Air Traffic Control Brain: Computational Red Teaming and Big Data for Real-time Seamless Brain-Traffic Integration.

[4] EUROCONTROL Guidance Material for Short Term Conflict Alert. Appendix A: Reference STCA System” Edition number:2.0, Document identifier: EUROCONTROL-GUID-123, 2009.

[5] N.B.Ağayev, N.İ.İsgəndərov “Havadakı hərəkətin avtomatlaşdırılmış idarə edilməsi zamanı münaqişələrin aşkarlanması və qərar qəbul edilməsi üsullarının sistem struktur analizi” Elmi məcmuələr. Cild 15, № 2, “Azərbaycan Hava Yolları” Qapalı Səhmdar Cəmiyyəti Milli Aviasiya Akademiyası. Bakı 2013, səh. 21-27.

[6] EUROCONTROL Specification for Medium-Term Conflict Detection. Edition 1.0, EUROCONTROL-SPEC-139, 2010.

[7] Nadir Agayev, Nasraddin Isgandarov “Flight plan based aircraft conflict detection method” International Journal of Latest Research in Science and Technology. Volume 3, Issue 5: Page No 63-66. September-October 2014. http://www.mnkjournals.com/ijlrst.htm.

[8] Nathan Marz James Warren.Big Data Principiles and best practices of scalable realtime datasystems. Maning Publications 2012.

[9] Frédéric Kaplan A map for big data research in digital humanities.\ Front. Digit.Humanit. 06 may 2015

[10] Майер-Шенбергер, В. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер ; пер. с англ. Инны Гайдюк. — М. :Манн, Иванов и Фербер, 2014. — 240 с.

Big Data - to Address The Business Issues

Dr.Yashar Hajiyev[1], Dr. Ali Şaintaş[2]
[1,2]Qafqaz University, Baku, Azerbaijan
[1]*yhajiyev@qu.edu.az*, [2]*asahintas@qu.edu.az*

***Abstract* — Paper is focused on key success elements of big data analytics, business needs and demands, fact-based decision making culture. The efficiency of big data dedicated analytical tools was considered in conjunction with analyze and getting deep insight on current business data for improved performance and results.**

Keywords — Big Data, desition making technology; decision support data management, big data projects.

I. INTRODUCTION

The Big Data ─ is it panacea for companies of business problems, especially during economic crisis? Is it practically helpful? What are the specific capabilities of big data to leverage business activity, to increase revenues from enlarged client basis? What might be an ultimate goal of exploiting of big data as a new business enabler?

The definite answer to these questions would help business stakeholders to avoid thoughts of Big Data as kind of cut-end science experiment. There should be clear business insight into real benefits of Big Data specific opportunities that might drive upgrating of enterprises' current business processes. Big Data is promised to be very instrumental to empower managers to concatenate disembodied data and fragmented, subtle information to reveal current latent processes that could directly affect their business. Variety of data sources, foreseeable analytics ─ are key to big data potentialities that strengthen their current business intelligence to improve the quality of decision making on cases that incorporate their customer and products. From Big Data related experience of many advanced companies it is obvious that the Big Data based market researches and monitoring methods are far more powerful and precise than that were used in past [1]. It gives more flexible tools to develop smarter decisions and make better predictions in areas where so far the only intuition and oversee tooling is prevailed over. As the approaches and analysis tools of big data spread the smart business stakeholders see using big data for what it is ─ deliver new functionality and capabilities to meet the demands and challenges of dynamically changed business situation.

Overwhelming growth of easier accessible of new sources of information favours the gradual digitizing of business activity when the intelligently utilization of huge amounts of on-line information helps to improve business performance. New sources of various data are expected to be able to shed new light on almost any business issues ─ from customer sympathies to supplier relationships.

As large corporations look for leverage for their information assets, the importance of data predominating in any virtual field of interest of business is increasingly magnified. Numerous surveys on Big Data [4] initiatives on analytical application of data among peer institutions discovered that the companies shared a common interest and common objective to utilize effectively data and analytics in supporting of their main business aims. The surveys also revealed that 70% of world famous brand companies have either already deployed or are working on plans to deploy big data-related specific programs. There are huge amounts of Big Data technologies behind new data processing, visualization, and analytical instruments, are aimed to take market share away from well-knowing traditional business intelligence vendors.

A tremendous part of surveys 80% underscore [3] that the dedicated Big Data tools or architectures are already used or planning to use in companies' production environment to cope with the come in of huge amounts of data.

II. BİG DATA AS PROVİDER OF DECİSİON SUPPORT TECHNOLOGİES

The greatest benefit of Big Data for the business intelligence is the leveraging of massive and aggregated data sets from different sources which are difficult to process by traditional methods. Big Data is intended to help companies understand definitely which client groups are unsatisfied with offered service, what is the customer expectation and fix faulty processes or retailer that caused dissatisfaction. To retrieve the relevant information, the huge amount of unstructured bytes from different sources should be handled before the information strips give us the full picture. In principle the similar market analysis can be done by traditional data warehousing approaches in weeks or months. But big data technologies do it several times faster, they are not held captive by their current user base. The most of considerable sources for big data are quite new. The images, multimedia products, comments, messages nowadays are main sources of information. The real business impact is that big data technologies empower managers to measure, and hence to get to know, radically more about own businesses, and immediately convert obtained knowledge into improved decision making and performance. That way [3] business executives could track not what customers purchased, also what else they are seeking for; what they navigated for through on-line shopping; how much they were influenced by promotions, reviews, and page layouts; and similarities across individuals and groups. Big data technologies make available

unexpected sources of data accessible. Those in turn open up opportunities for newcomers to find better way to use big data than an insider.

The technical aspects of big data usage are very practical. It is invaluable to help the enterprises to figure out what to offer to market, and where to go next. However the managerial challenges in most cases are dealing with the role of senior executive staff. Big data related analytic [2] techniques introduce new culture of decision making that encompass:

-Big Data technologies do not diminish role of manager in decision making and do not deny the need for vision or human insight. Especially, when data not available, the experience and intuition of managers have distinguished value;

-To be confident in quality of fetched data and how far you might rely on big data analytical outcomes;

-Critical aspects of big data is its impact on how decisions are made and who gets to make them;

-Data professionals specialized at working with large amount of information. New generation of computer scientists are bringing to bear techniques for working with very large data sets;

-Employing of available open-source software combined with commodity hardware to integrate all relevant internal and external sources of data; to gather, handle and analyze of incoming streams of data and distributes them onto cheap monitoring disks;

-In era of big data, the information is created and transferred, and expertise is often not where the relevant decision making rights are used to be;

-Employees having clear insight into problems to be incorporated with the right data, and with people having problem-solving techniques to exploit them effectively;

-Suppression of imitations of being more data-driven company than they actually are. It's quite easy to make an error to find cause-and-effect relation, and find the numbers that would justify the decision;

The outcomes are self-evident: data-driven decision-making tend to be better decisions. Business executers should decide how harness big data to transform decision making for sector after sector. But the data tell us that's the surest bet.

Foremost the principal point is that any data collecting and collating should be separated from Big data dedicated analytics. Business entities demand immediate insight into their momentary situation — strike moment in time where information and current data fuels decision-making, prompts a client to purchase, and get companies to present brand immutability across services. However, companies need current data from a big group of sources in order to feed processes through the combination of big data and analytics. The global supply chain management provides opportunities to [3] demonstrate how relationship between big data and analytics generates more efficient business processes.

What is much demanded ─ is to have actionable business intelligence retrieved from big data. To find the clue idea held therein, the big data specific tools are required for the predictive analytics to mine them. Intelligent predictions of business processes in future can be elaborated on the base of predictive analytics platforms. That is highly competitive market that includes world famous brands as well as several new companies that are specializing in data visualization and geospatial analysis. Business analysts apply algorithms to company's data sets to yield granular and real-time predictions. It takes form of recursive process for predictive analytics

III. LEVELS OF BIG DATA IMPLICATION INTO BUSINESS

Different companies are moving ahead at different paces with respect to how they fast are adopting big data to their business and how wide they are implementing the advanced analytics to create competitive advantage. As many surveys revealed [5] the most advanced companies in their strive to be data-driven went through several levels:

- to monitor current business activity to deploy BI (Business Intelligence) solutions. BI utilizes basic analytical methods to set alert under- or over-performance areas of business, and automates sending alerts with relevant information to business executers;

- through the Business Monitoring do observation on a specific area of business where these particular actions can be taken to improve business performance;

- enterprises develop and implement embedded analytics enabling automatically optimize parts of their business processes;

- companies try to sale to different organizations the packet of their firsthand data including analytical insights. Integrate analytic results directly into manufactured products to create more "sophisticated" products. They incorporate actionable insights and outcomes to deepen their client relationships and seriously review their "customer experience."

-ultimate aim of advanced companies in using dedicated Big Data tools is to leverage the insights that they are gathering about their clients' usage patterns, product performance behaviors, and dominated market trends [5] to convert their business models into advanced services in new markets;

The business opportunities have much of future with regards to how and where companies can exploit big data and how they can benefit from advanced analytics to navigate their business perspectives – from observation over the monetization of data for business usage to building intercorporated newest net business models

Presumably enterprises need to get grasp what they want to achieve from business perspective. Before commences travel through big data space it is required to know intended waybill destination.

The companies proceeded on the assumption that Big-data related projects have a several layers of formalization from abstracting of big data up to exercising analytics against the formalized data. General elements of analytical Big-data and are their relation to each other. The higher level elements allow to simplify big data projects and do them more productive. Underlying thought is Hadoop is usually at the core of Big-data projects, but it is not obligatory

IV. BENEFİTS OF BİG DATA SPECİFİC ANALYTİCS

Collecting and storing Big Data is just collecting and storing data. Big Data by its own cannot create business value. Business value could be created only when data is handled, processed analyzed, collated and acted on. The benefits from big data analytics [2] can be varied, and could be a basis for competitive advantage. Big data has strong perspective in business forecasting due to its potential benefits of using data and analytics in decision making. It is revealed that publically traded firms found that enterprises that have implemented "data-driven decision-making" gain output and productivity that is 5-6% higher than different companies. It is included relationship expanded to performance elements like asset market value, return on equity and utilization. Findings indicate [6] that top-performing organizations use analytics five times more than lower performers.

There are also potential benefits from governments' use of big data, following the scenario how to take advantages of big data analytics: to reduce strain on services and infrastructure because of growing population but a reduced; to improve health care to stimulate its efficiency and less investment; to strengthen levels of security through effective predictions of threats to public safety and national borders; to increase accuracy in prediction and managing of country's economy risks.

Impact of big data might to fuel development of the Internet itself. Abstracting from the specifics of each project, we can identify their common features that ensure their success

-projects should be focused on business issues rather than technologically oriented. The business needs should be addressed ─ solving a problem or seeking for opportunity.

-it is dead-end road when some executives to push big data projects without clearly identified goals;

-big data analytics projects focus on customer-centric objectives and exploit current and newly accessible internal sources of data;

-big data analytics projects are particularly useful for enterprises that want to get better understanding of customers demands, develop adequate relationships with customers, and intensify operations that enlarge customer experience.

-apart from focus, successful big data initiatives have to begin with a narrowly specified set of objectives rather than a sluggish " create it — they will come up" approach.

CONCLUSSİONS

Big data can be considered as very recently generated tool that support computer-based decision making. It supports a data-centric approach to modeling business operations and processes. Big data is 4th generation of decision support data management system. It introduces the capacities to grasp, store, and analyze high-volume, high-velocity, and high-variety data. That capacities provide support to automatic decisions making processes. Big data devoted analytic introduced mew type of fact-based decision-making culture. Business is getting to be able to "run by the numbers" and to do experimentation to see what works best. The shaping and supporting of this culture depends on senior management. Big data begets different [7] new data management technologies, systems and platforms. The Big data dedicated analytics requires traditional tools and as well as newest data analysis and visualization languages. However, most important are there business users, analysts, and data scientists who can work with and handle big data.

REFERENCES

[1] Dr. Hugh J. Watson, "Big data: concepts, technologies, and applications," Available at
http://www.bacollaborative.com/sites/default/files/learningcenterfiles/Watson,%20Tutorial,%20Big%20Data,%20Business%20Analytics%20Collaborative.pdf

[2] J. Power, D.J. (2007). "A Brief History of Decision Support Systems," DSSResources.com, http://DSSResources.COM/history/dsshistory.html, version 4.0.

[3] Eckerson, W. (2004) "Gauge Your Data Warehousing Maturity," DM Review (14)11, pp. 34.

[4] Franks, B. (2012) Taming the Big Data Tidal Wave, New York: Wiley.

[5] Ayers, I. (2007) Super Crunchers, New York: Bantam Books.

[6] O'Brien, J. (2012) Presentation at The Data Warehousing Institute conference, Las Vegas. February 2012..

[7] Miller, S., S. Lucas, L. Irakliotis, M. Ruppa. T. Carlson, and B. Perlowitz (2012) "Demystifying Big Data:. A Practical Guide to Transforming the Business of Government," Washington: TechAmerica Foundation. Available at
http://breakinggov.com/documents/demystifying-big-data-a-practical-guide-to-transforming-thebus/

Biznes Sahəsində BİG DATA-nın Tətbiqi Perspektivləri

Rasim Mahmudov
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
rasimmahmudov@gmail.com

***Xülasə* - Tədqiqat işində Big Data texnologiyalarının biznesin inkişafında yaratdığı imkanlar araşdırılır. Dünyanın qabaqcıl analitik şirkətlərinin Big Data-nın inkişaf perspektivləri ilə bağlı proqnozları şərh olunur. Bank, pərakəndə satış və telekommunikasiya biznesi sahələrində Big Data-nın tətbiqinin müsbət nəticələri analiz edilir və ümumiləşdirilir.**

Açar sözlər - Big Data texnologiyaları, Big Data-nın üstünlükləri, Big Data-nın problemləri.

I. GİRİŞ

İnformasiya cəmiyyətinin inkişaf etdiyi, informasiya iqtisadiyyatının formalaşdığı bir şəraitdə böyük həcmdə müxtəlif xarakterli məlumatlar strateji resursa çevrilir. İstehsal, kommersiya, xidmət sahələrində İKT-nin tətbiqinə əsaslanan idarəetmə üçün zəruri olan texnoloji, audio, video, foto informasiyanın həcmi dinamik olaraq genişlənir. Nəticədə, İnternetin, sosial şəbəkələrin, audio, video, geolokasiya servislərinin inkişafı sayəsində informasiya məhsullarına və xidmətlərinə tələbat sürətlə artır. Müştərilərə müvafiq xidmətləri təklif etmək üçün müəssisələr müxtəlif mənbələrdən əldə edilən böyük həcmli informasiyanı analiz etməli olurlar. Toplanan informasiya həmçinin telekommunikasiya şirkətləri, banklar, topdansatış ticarəti müəssisələri kimi biznes sahələri üçün də mühüm strateji aktivə çevrilir.

Müasir şəraitdə biznesi ənənəvi qaydalarla idarə etmək qeyri-mümkündür. Bu gün yalnız düzgün qərar qəbul etmək yetərli deyil, həm də onu operativ şəkildə etmək lazımdır. Qeyri-dəqiq proqnozlar biznesi fəlakətə sürükləyə, iflasa gətirib çıxara bilər. Ona görə də texnologiyanın və iqtisadiyyatın sürətli inkişafı şəraitində vaxtında düzgün qərarların qəbul edilməsi üçün real istrumentariya işləyib hazırlamaq lazımdır.

Big Data bunun üçün strukturlaşdırılmış və strukturlaşdırılmamış verilənlərin qanunauyğunluqlarını müəyyən etməklə bağlı yeni və perspektivli yanaşma təklif edir. Digər perspektiv istiqamətlərdə olduğu kimi, *Big Data* sahəsində də marketinq, texnoloji, iqtisadi yanaşmaların uzlaşması müşahidə olunur ki, bunlar da yeni iqtisadi modelin əsasını təşkil edir.

Big Data-nın xidmət, marketinq sahələrində praktiki tətbiqi sahəsində böyük təcrübə olmasa da, müvafiq layihələrə qarşı maraq davamlı olaraq artmaqdadır. Bu texnologiyalar innovasiya şirkətlərinin rəqabətədavamlılığı artırmaq, yeni xidmət sahələri yaratmaq, müştərilərlə qarşılıqlı əlaqələrin idarə edilməsinin təkmilləşdirilməsi ilə bağlı müxtəlif məsələlərin həllində uğurla tətbiq edilir.

II. BİG DATA-NIN PERSPEKTİVLƏRİ İLƏ BAĞLI ANALİTİK TƏDQİQATLAR

McKinsey analitik şirkətinin araşdırmalarına görə, *Big Data* yaxın gələcəkdə dünya iqtisadiyyatında böyük dəyişikliklərə səbəb olacaq. Bu şirkət 2020-ci ilə qədər olan müddətdə ABŞ-ın iqtisadi və sosial həyatına təsir edəcək beş əsas faktor sırasında energetika, ticarət, infrastruktur və təhsil sahələri ilə yanaşı, *Big Data*-nı da göstərib. *McKinsey*-nin hesablamalarına görə, *Big Data* 2020-ci ilə qədər ABŞ-da Ümumi Daxili Məhsulun həcminə 325 mlrd. dollar əlavə edilməsini təmin edəcək. Şirkət hesab edir ki, *Big Data*-nın verəcəyi dəqiq proqnozlar hesabına təkcə iri pərakəndə satış şəbəkələri 55 milrd. dollar gəlir əldə edəcəklər [1, 2].

IDC marketinq şirkəti *Big Data* texnologiyaları və xidmətləri bazarının 2013-cü ildən etibarən hər il orta hesabla 27% artığını və 2017-ci ildə 32,4mlrd. dollara çatacağını proqnozlaşdırır [3].

Wikibon analitik şirkətinin hesabatında isə 2017-ci ildə müvafiq bazarın həcminin 50 mlrd. dolları keçəcəyi qeyd olunur [4].

IDG Enterprise media şirkətinin 2014-cü ildəki açıqlamasına görə, dünyanın nəhəng şirkətlərinin 70%-i *Big Data* texnologiyalarını tətbiq edir və ya bunu planlaşdırır. Şirkət kiçik və orta şirkətlərlə bağlı həmin göstəricinin 56%-ə bərabər olduğunu qeyd edib [5].

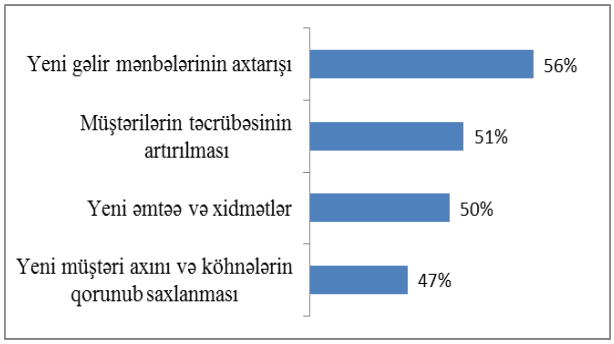
İnternet mühitində inkişaf edən və genişlənən media və əyləncə sənayesində də böyük həcmli verilənlərin saxlanmasına dair tələblər çox sürətlə artmaqdadır. HD standartının yayılması və mobil videoya olan maraq müvafiq rəqəmsal kontentə olan tələbatı yüksəldir. Bununla bağlı olaraq videoteka arxivlərinin yaradılması üçün verilənlərin HDD-də saxlanmasına dair tələbat də artır. *Coughlin Associates* şirkətinin proqnozuna görə, 2017-ci ilə qədər media və əyləncə sənayesində verilənlərin rəqəmsal saxlanc həcminə dair tələbat 5-6 dəfə artacaq [6].

İrlandiyanın *Accenture* konsaltinq şirkətinin 2014-cü ildə *Big Data*-nın biznes sahəsində üstünlükləri və çatışmazlıqlarını aşkara çıxarmaq üçün dünyanın 19 ölkəsindən olan 1000 nəfər biznes meneceri arasında sorğu keçirmişdir [7]. Rəyi soruşulanların 56%-i *Big Data*-nın yeni

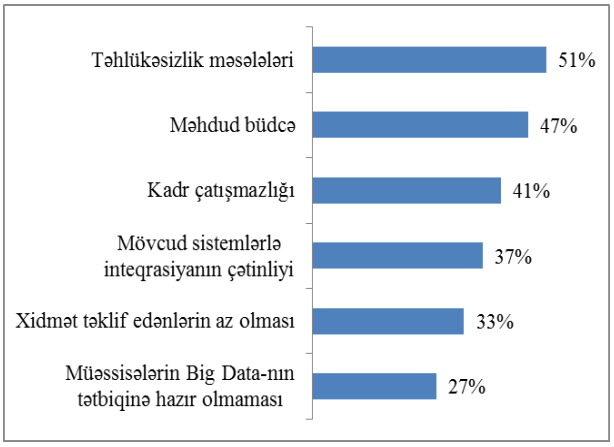
gəlir mənbələrinin axtarışı, 51%-i müştərilərin təcrübəsinin artırılması, 50%-i yeni əmtəə və xidmətlərinin yaradırlması, 47%-i isə yeni müştəri axını və köhnə müştərilərin qorunub saxlanması kimi üstünlüklərini qeyd ediblər (Şəkil 1).

Sorğuda rəyi soruşulanların 51%-i *Big Data*-nın əsas problemləri qismində təhlükəsizlik məsələlərini, 47%-i maliyyə vəsaitlərinin çatışmazlığını, 41%-i kadr çatışmazlığını, 37%-i mövcud sistemlərlə inteqrasiyanın çətinliyini, 33%-i xidmət təklif edənlərin az olmasını, 27%-i müəssisələrin bu texnologiyaların tətbiqinə hazır olmamasını qeyd edib (Şəkil 2).

Sorğudan da göründüyü kimi, *Big Data* ilə bağlı əsas problemlərdən biri də kadr təminatı ilə bağlıdır. Bu texnologiyalarla işləmək üçün müvafiq ixtisaslı mütəxəssislər yetişdirmək lazımdır. Verilənlərin analizi üzrə mütəxəssislər biznesin inkişafı üçün yeni ideya və imkanların effektivliyini qiymətləndirməyi mümkün edən məsələlər qoya bilərlər.



Şəkil 1. *Big Data*-nın biznes imkanları [7]



Şəkil 2. *Big Data*-nın əsas problemləri [7]

Onlar konkret biznes sahəsində verilənlərin analizi metodlarına və texnologiyalarına və xüsusi biliklərə malik peşəkar olmalıdırlar.

Gartner şirkətinin proqnozuna görə, 2015-ci ilə qədər dünyada *Big Data* sahəsində çalışmaq üçün 4,4 milyon iş yeri açılacaq ki, onlardan da 1,9 milyonu ABŞ-ın payına düşür. Həm də hər bir müvafiq iş yeri informasiya texnologiyaları sferasından kənarda əlavə üç iş yerinin açılmasına səbəb olur [8].

Gartner ekspertlərinin fikrincə, bu sahədə əsas problemlərdən biri kadr çatışmazlığıdır. Dövlət və özəl təhsil sistemi bu sahəni lazımi qədər ixtisaslı kadrlarla təmin etmək iqtidarında deyil. Hazırkı potensial hesabına müvafiq iş yerlərinin yalnız 30%-ni kadrlarla təmin etmək mümkündür.

III. PƏRAKƏNDƏ SATIŞ BİZNESİNDƏ BİG DATA

Big Data-nın pərakəndə ticarətdə tətbiqi təcrübəsi göstərir ki, bu texnologiya müvafiq ənənəvi və şəbəkə müəssisələrində qiymətləri optimallaşdırmağı, loqistika proseslərini təkmilləşdirməyi, ehtiyatları və nəqliyyatı effektiv şəkildə idarə etməyi təmin edir. Böyük həcmli verilənlərin müasir analiz texnologiyaları toplanmış informasiya massivindən və cari satış haqqında məlumatlardan istifadə edərək analitik məsələlərin operativ şəkildə həllinə kömək edir [9].

Pərakəndə ticarət şəbəkələrinin marketinq xidmətlərində alıcılar haqqında məlumatların analizi üçün imkanlar yaranır. Bunun əsasında reklam göndərişləri üçün müştəri bazasını dəqiq seqmentləşdirmək mümkün olur.

Müasir dövrdə rəqabət gücləndikcə şirkətlər üçün vacibdir ki, müştərilərə öz məhsul və xidmətlərini məhz onların bunlara daha çox ehtiyac duyduqları zaman təklif etsinlər. Ona görə də biznes mühitində marketinqin rolu artmaqdadır. *IBM* şirkətinin hesablamalarına görə, şirkət direktorlarının 63%-i öz biznes strategiyalarını uğurla həyata keçirmək üçün marketinq üzrə müavinlərin xidmətlərindən istifadə edirlər [10].

Öz növbəsində, marketoloqlar da işlərini uğurla qurmaq üçün yüksək analitik imkanlara malik olan *Big Data* texnologiyalarından istifadə edə bilərlər. Əgər əvvəllər onların sərəncamında kiçik həcmdə informasiya bazası var idisə, indi vəziyyət xeyli dəyişib. Marketoloqlar daxili və xarici mənbələrdən böyük həcmdə informasiya toplayırlar. Böyük həcmdə informasiyanın analizi müştərilərə fərdiləşdirilmiş təkliflər irəli sürməyə imkan verir. *IBM* şirkətinin dünyanın 59 ölkəsində 19 sahə üzrə fəaliyyət göstərən şirkətlər arasında keçirdiyi sorğunun nəticələri göstərir, bu biznes subyektlərinin 94%-i qarşıya qoyulan məqsədlərə çatmaq üçün analitikanın mühüm rol oynadığını bildiriblər. Tədqiqatın yekunları həmçinin müəyyən edib ki, şirkətlərin marketinq bölmələri informasiya texnologiyaları ilə qarşılıqlı əməkdaşlıq şəraitində fəaliyyət göstərdikdə fəaliyyət effekti daha yüksək olur.

Off-layn ticarətdə *Big Data*-dan alıcıların davranışlarını analiz etmək, biznes strategiyalarını müəyyənləşdirmək, məhsulları düzgün yerləşdirmək və paylaşdırmaq, təchizatı optimallaşdırmaq və son nəticədə satış həcmini yüksəltmək üçün istifadə edilir.

On-layn ticarətdə isə satış mexanizminin özü *Big Data* üzərində qurulur: alıcılara məhsullar əvvəlki alqı-satqının bazası və onların fərdi yanaşmaları əsasında təklif edilir. Hər iki halda böyük həcmli verilənlərin analizi satış məsrəflərini azaltmağa, müştəri rəğbətini artırmağa və daha geniş auditoriyanı əhatə etməyə imkan verir. Mürəkkəb iqtisadi

vəziyyətlərdə, o cümlədən iqtisadi böhran zamanı xərclərə qənaət etmək və müştəri rəğbətini artırmaq məsələləri ön plana çıxır. *Big Data* texnologiyaları da məhz bu məsələlərin uğurla həlli üçün çox effektlidir [8].

Kütləvi tələbat mallarının satışı ilə məşğul olan Amerikanın *Guess* şəbəkəsi bu texnologiyaları satış haqqında informasiyanı və malların hərəkətini operativ izləmək üçün tətbiq edir. *HP Vertica* sisteminin şəbəkəsindən istifadə etməklə hansı malların daha yaxşı satıldığını analiz edir, tələbat olmayan malları ayırd edir. Satış haqqında aktual məlumatlar növbəti günün səhəri təqdim olunur [11].

Wal-Mart supermarketlər şəbəkəsi *Big Data* texnologiyalarından tələbin proqnozlaşdırılması, qiymətlərin optimallaşdırılması, aksiyaların və marketinq kampaniyalarının hazırlanması üçün istifadə edir. Bunun üçün bir saat ərzində 2,5 petabayt həcmində informasiya toplanır və analiz edilir [12].

Big Data-nın pərakəndə ticarətdə tətbiqi təcrübəsinin analizi göstərir ki, bu texnologiya qiymətləri optimallaşdırmağa, assortimentlərlə işləməyə, loqistik prosesləri təkmilləşdirməyə, ehtiyatları və nəqliyyatı effektiv idarə etməyə imkan verir.

Pərakəndə satış şəbəkələrinin marketinq xidmətlərində müştəri davranışlarının analizi üçün yeni imkanlar yaranır. Alıcılar haqqında məlumatlar müştəri bazasını reklam göndərişləri üçün dəqiq seqmentləşdirməyə imkan verir.

IV. BANK BİZNESİNDƏ BİG DATA

Bank biznesində *Big Data* texnologiyalarının tətbiqi xidmət keyfiyyətinin yüksəldilməsinə, yeni bank məhsullarının və xidmətlərinin hazırlanmasına, təhlükəsizliyin təmin edilməsinə və xərclərin optimallaşdırılmasına yönəlib.

Məsələn, *ING* maliyyə holdinqi *Big Data* texnologiyalarının köməyi ilə sayt ziyarətçilərinin hərəkətləri barədə informasiyanı toplayır və analiz edir. Bunun əsasında da istifadəçi interfeysi optimallaşdırılır və müştərilər üçün fərdi marketinq təklifləri hazırlanır. “*Next Best Action*” (“Növbəti ən yaxşı fəaliyyət”) konsepsiyasının reallaşdırılması çərçivəsində *Big Data* texnologiyaları konkret müştərilərə təklif edilən bank məhsullarının seçimi və fərdiləşdirilməsi həyata keçirilir [13].

HSBC bankında dələduzluqla mübarizə məqsədilə tətbiq edilən *Big Data* texnologiyaları təhlükəsizlik xidmətinin effektivliyinin 3 dəfə, insidentlərin aşkarlanmasının dəqiqliyinin isə 10 dəfə artırılmasına səbəb olub. İqtisadi effekt isə 10 mln. dollar təşkil edib [14].

Bütövlükdə, bank sektorunda *Big Data* layihələrinin analizi göstərir ki, rəqabətin kəskinləşməsi şəraitində bu texnologiyaların tətbiqinin əsas məqsədi hesabatların operativ əldə edilməsi, şübhəli əməliyyatlarla, bədbiyyətli əməllərlə mübarizə, həmçinin xidmət və məhsulların fəardiləşdirilməsidir.

Bank biznesində *Big Data* texnologiyalarından istifadənin iki əsas istiqamətini fərqləndirmək olar. Birinci istiqamət daxili bank sistemlərinin və xarici mənbələrin analizi əsasında müştərinin profilinin müəyyən edilməsinə yönəlib. İkinci istiqamət çərçivəsində bank transaksiyaları, bankomat və digər bank qurğularından əldə edilən məlumatlar analiz edilir.

V. TELEKOMMUNİKASİYA BİZNESİNDƏ BİG DATA

Böyük həcmdə informasiyanın analizi və işlənməsinə əsaslanan əsas bazar xidmətləri sahələrdən biri də telekommunikasiyadır.

Big Data texnologiyalarının tətbiq olunmadığı vaxtlarda çoğrafi şəbəkələr yalnız tikililərin və əhalinin sıxlığı haqqında informasiya əsasında genişləndirilirdi. Hazırda isə bunun üçün *Big Data* texnologiyaları tətbiq edilməklə, trafikin real yüklülüyü, xidmətə olan tələbat, müştəri “dəyəri”, rabitənin keyfiyyətinə dair tələb, müxtəlif xidmət kateqoriyalarına tələbat haqqında informasiya və s. əsasında çoxfaktorlu analiz aparılır.

Big Data texnologiyalarını uğurla tətbiq edən telekommunikasiya şirkətlərindən biri də Çinin *China Unicom* mobil operatorudur. Bu şirkət biznes üçün daha yüksək səmərəlilik əldə etmək üçün *Big Data* texnologiyalarından dörd əsas istiqamətdə istifadə edir. Bu istiqamətlərdən üçü şirkətin daxili faəliyyətini yaxşılaşdırmağa, biri isə müştəri məmnunluğunu və sayını artırmağa hesablanıb:

- yüksək dəqiqilikli marketinq;
- müştəri xidmətinin keyfiyyətinin idarə edilməsi;
- şirkətin daxili fəaliyyətinin və inkişafının optimal planlaşdırılması;
- informasiya aktivlərinin pula çevrilməsi [15].

IBM şirkətinin *Watson* superkompüteri isə real zaman rejimində pul transaksiyaları üzrə informasiya axınlarını analiz edir. Şirkətin məlumatına görə, *Watson*-un fəaliyyəti sayəsində aşkar edilən bədniyyətli əməliyyatların sayı 15%, saxta əməliyyatlardan sığortalanan pul vəsaitlərinin sayı isə 60% artıb [16].

NƏTİCƏ

Müxtəlif beynəlxalq anailitik şirkətlərin araşdırma və proqnozlarında rəqəm fərqləri olsa da, yekdil fikir ondan ibarətdir ki, dünyada *Big Data* texnologiyalarına olan tələbat sürətlə artmaqdadır. \bu texnologiyalar sosial-iqtisadi fəaliyyətin səmərəliliyinin artırılmasının ən mühüm vasitələrindən birinə çevrilir. Biznes sahələrində, o cümlədən bank, pərakəndə satış və telekommunikasiya kimi yüksək rəqabətli sahələrdə *Big Data* texnologiyalarının uğurla tətbiqi müştəri dairəsini, gəlirləri artırmağa, idarəetməni, planlaşdırmanı daha effektiv şəkildə həyata keçirməyə imkan verir. Bütün bunlar yaxın gələcəkdə də yeni *Big Data* texnologiyalarının işlənməsinə və tətbiqinə, bu sahədə kadr hazırlığının həyata keçirilməsinə böyük həcmdə vəsaitlərin cəlb edilməsini stimullaşdırır.

ƏDƏBİYYAT

[1] J.Manyika, M.Chui, B.Brown et all. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, June 2011, 156 p.

[2] http://www.mckinsey.com/insights/big_data_and_advanced_analytics

[3] http://www.idc.com/prodserv/4Pillars/bigdata

[4] http://www.wikibon.org/wiki/v/Big_Data

[5] http://www.idgenterprise.com/marketing-services

[6] http://www.tomcoughlin.com

[7] Accenture. Big Success With Big Data. Executive Summary, Accenture 2014, 12 p.

[8] http://www.gartner.com/technology/home.jsp

[9] E.G. Ularu, F.C. Puican, A. Apostu et al. Perspectives on Big Data and Big Data Analytics // Database Systems Journal, 2012, vol. III, no. 4, pp. 3-15.

[10] http://www.ibm.referrals.selectminds.com/landingpages/marketing-communications-opportunities-at-ibm-15

[11] http://www.cnews.ru/news/top/gloriya_dzhins_nachala_analizirovat_big_data

[12] https://www.datafloq.com/read/walmart-making-big-data-part-dna/509

[13] IBM Corporation Software Group. ING achieves 3x gain in marketing response with personalization, 2012, 4 p.

[14] http://www.hsbc.com

[15] W.Huang, Z.Chen, W.Dong et all. Mobile Internet Big Data Platform in China Unicom, Tsinghua Science and Technology, February 2014, Volume 19, Number 1,

pp. 95-101

[16] http://www.ibm.com/big-data/au/en/big-data-and-analytics/watson-foundations.html

Böyük Verilənlərin Saxlanılma Problemləri və Mövcud Həllər

Məkrufə Hacırəhimova[1], Mərziyə İsmayılova[2]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]makrufa@science.az, [2]imarziya@gmail.com

***Xülasə* — Verilənlərin saxlanması və emalı problemi hər zaman movcud olmuşdur. İnformasiya-kommunikasiya texnologiyaları vasitəsilə yaradılan böyük həcmli verilənlər bu gün saxlama problemini bir daha gündəmə gətirmişdir. Məqalədə verilənlərin saxlanma sistemləri müqayisəli şəkildə araşdırılmış, həmçinin böyük verilənlərin saxlanılma problemləri və mövcud yanaşmalar analiz edilmişdir.**

Açar sözlər — böyük verilənlər, böyük verilənlərin saxlanması, verilənlər mərkəzi, bulud hesablama, bulud saxlama.

I. GİRİŞ

Son onillikdə informasiyaya əlyetərliyin asanlaşması və kommunikasiya vasitələrinin çoxalması nəticəsində hər il rəqəmsal informasiyanın həcmi həndəsi silsilə ilə artır. Statistik rəqəmlərə nəzər salaq. IDC (International Data Corporation) analitik şirkətinin məlumatına görə dünyada yaradılmış və replikasiya olunmuş informasiyanın həcminin təxminən 7,5 zetabayta bərabər olması və informasiyanın 90%-dən çox hissəsini strukturlaşdırılmamış verilənlərin (foto, video, mp3 faylları, sosial media faylları və s.) təşkil etdiyi bildirilir, 2020-ci ildə dünyada rəqəmsal informasiyanın həcminin 44 zetabayta çatması haqqında proqnoz verilmişdir [1, 2]. Bu rəqəmlər həqiqətən də verilənlərin həcminin çox böyük olduğunu bir daha sübut edir.

Böyük verilənlər (BV) həcm, sürət və müxtəliflik ilə xarakterizə olunur. Həcm böyük verilənlərin ən əsas xarakteristikasıdır. Həcm isə ilk növbədə saxlama problemi yaradır ki, bu da genişmiqyaslı saxlama və paylanmış emal tələb edir. İnformasiyanın həcmi həndəsi silsilə ilə artdığı üçün informasiyanın etibarlı saxlanılması, təhlükəsizliyinin təmini, idarə olunması və onlara operativ müraciətin təşkili hər bir təşkilatın qarşılaşdığı problemdir [1].

Yeni elmi modellər, optimallaşma üsullları, təhlükəsizlik modelləri BV saxlanması və əlyetərliyinin təmin edilməsindən doğan böyük problemlərlə başlayır. Belə ki, həm saxlama sistemləri arxitekturu səviyyəsində (bulud saxlamaları və ehtiyat surətlərin alınması, paylanmış saxlanclar), həm də fiziki səviyyədə yeni həllər tələb olunur.

II. BÖYÜK VERİLƏNLƏRİN SAXLANILMASI QURĞULARI

BV əsri üçün analitik alətlərlə yanaşı böyük həcmli və sürətli saxlama sistemləri tələb olunur. Ümumiyyətlə verilənlərin saxlanma qurğuları 4 tipə bölünür: sərt disklər (Hard Disk Drive, HDD), optik disklər (Optical Disk Drive, ODD), sərtgövdəli qurğular (Solid State Drive, SSD) və maqnit lentlər [3].

A. Sərt disklər

HGST (Hitachi Global Storage Technologies) disklərin satış şirkətinin tədqiqatlar üzrə vitse-prezidenti K. Muncenin (Currie Munce) proqnozuna görə sərt disklər (”onlar hər yerdə var, praktikdir və ucuzdurlar”) 2020-ci ilə qədər əsas saxlama sistemi olaraq qalacaqdır. Sərt disklərin tutumunun artırrılması istiqamətində daima işlər davam etdirilir. Hal-hazırda istifadə olunan iri sərt diskin tutumu 4 terabaytdır. Sərt disklər üçün nanolitoqrafiya, helium disklər və HAMR **(***Heat-assisted magnetic recording***)** kimi üç yeni yanaşma daha çox imkanlar vəd edir [3].

Nanolithography – iki innovativ *“nanoimprinting” və* “*molecularself-assembly”* kimi texnologiya vasitəsilə yaradılan sərt disk yazı sıxlığını artırmaqla disk tutumunu ikiqat böyüdür [3].

Helium drives. 2013-cü ilin sonunda HGST şirkəti helium elementi ilə hazırlanmış ilk sərt disk platformasını təqdim etmişdir. İstehsalçının qeyd etdiyinə görə bu texnologiya verilənlərin emal edilmə mərkəzi və bulud infrastrukturu üçün sərt diskin tutumunu və onun ümumi dəyərini əhəmiyyətli dərəcədə artırmağa imkan verir. Bu innovasiya sayəsində 2017-ci ilə HDD-də 1 kvadrat düymə yazı sıxlığının 5 tera bayta çatacağı proqnozlaşdırılır. Hava ilə hazırlanmış disklərdən fərqli olaraq helium elementli disklər daha tutumludur. K. Munceyə görə bu disk yeddi lövhə dəstəkləməyə imkan verir, həm də enerjidən istifadəni 23 %-ə qədər azaldır. Ancaq hava ilə hazırlanmış disk ən çoxu beş lövhəyə malikdir [3,4].

HAMR. texnologiyası sərt diskin tutumunu artırmağa imkan verir. Bu texnologiya ilə hazırlanmış disklər hazırda istifadə olunan disklərdən fərqli olaraq böyük sıxlıqla saxlamaya imkan verir [3].

Verilənlərin saxlanması üçün sərt disklərin ən böyük istehsalçılarından olan Seagate Technology şirkəti laboratoriya sınaqlarında disk sahəsinin hər kvadrat düyma 1 tera bayt olmaqla saxlama sıxlığını artırmaq üçün HAMR-dan istifadə edir. Bu günki disklər hər kvadrat düyma maksimum 620 qiqa

bayt sıxlıq təklif edir. Şirkət həm də HAMR texnologiyalı 6 terabayt informasiya saxlayan və sonda ola bilsin ki, maksimum 60 terabayt həcmli sərt disklərin ilk məhsulunun satışa çıxarılma planının 2016-cı ilə gözlənildiyini bəyan etmişdir [3, 4].

B. Solid state memory

Təşkilat üçün ciddi problemlərdən biri böyük həcmli verilənlərin dinamik olması və onların sürətli əldə edilməsidir. Bu problem SSD disklərindən istifadə edilməklə həll edilir, belə ki, onlar sərt disklərdən fərqli olaraq hərəkət edən hissələrə malik olmadığından verilənlərə sürətli müraciət əldə edilir və analiz üçün işçi yaddaş qismində istifadə olunur. Hazırda tədqiqatçılar SSD üçün yeni yanaşmalar üzərində işləyirlər. Məsələn, 3D NAND qurğusu əhəmiyyətli dərəcədə saxlama sıxlığını artırır və ilk qurğunun istehsalının 2015-ci ilə gözlənildiyi bildirilir [3-5].

C. *Maqnit lentlər*

Qeyd etmək lazımdır ki, saxlama texnologiyalarının inkişafına baxmayaraq, maqnit daşıyıcıları əvvəlki kimi qiymət/həcm cəhətdən ən yaxşı həll kimi təklif edilir, bundan başqa, onlar yüksək etibarlılığa malikdirlər və xidmətdə çox yararlı hesab edilirlər. Göstərilən səbəblərdən, onlardan hələ də, xüsusilə nadir hallarda istifadə edilən verilənlərin arxivləşdirilərək saxlanması üçün geniş istifadə edirlər [5].

D. Optik disklər

Optik disklər BV-nin saxlanmasında çox da effektiv deyildir. Yaddaşın həcmini artırmağa imkan verən daha yüksək sıxlığa nail olmaq üçün holoqrafik, yaxud çoxlaylı həllərdən istifadə edilməlidir. Bu da mürəkkəb texniki problemdir [5].

III. BULUD SAXLAMALARI

Bu gün saxlama məsələsinin həllində informasiyanın qurğular arasında miqrasıyasını həyata keçirən bir sıra texnologiyalar: DAS (Direct-Attach-Storage), NAS (Network Attached Storage), SAN (Storage Area Networks), HSM (Hierarchical Storage Management), ILM (Information Life-cycle Management) mövcuddur [1, 6, 7]. Lakin bu texnologiyalar BV əsri üçün o qədər də yararlı deyildir. Son zamanlar isə saxlama qurğularının yaddaş tutumunun artması, çoxsaylı kompüterlərin (server, kompüter və s.) hesablama və yaddaş resurslarının klasterləşməsi və virtualllaşdırılmasını həyata keçirməklə, verilənlərin emalı və yadda saxlanılmasına xidmət edən “grid” və “cloud computing” texnologiyalarının tətbiqi saxlama sahəsindəki problemləri demək olar ki, nisbətən aradan qaldıra bilmişdir [1, 8].

Müasir İT faktorları: böyük verilənlər, analitika və bulud texnologiyalarını bu gün bir-birindən ayrı təsəvvür etmək mümkün deyildir. Buludlarda saxlama, buludlarda hesablamalarsız böyük verilənlərlə işləmək mümkün deyildir. Qeyd etmək lazımdır ki, bulud texnologiyaları böyük hesablamaların aparılmasında son dərəcə müvəffəqiyyətli yanaşmalardandır. Burada böyük həcmli rəqəmsal informasiya IaaS (Infrastructure as a service), PaaS (Platform as a service), SaaS (Software as a service) bulud xidmətləri vasitəsi ilə mərkəzləşdirilmiş qaydada idarə olunur və saxlanılır [1, 9].

Bulud saxlama əksər təşkilatlara böyük həcmli informasiyaların saxlanmasına kömək edir. Belə ki, təşkilatların çoxu öz informasiyalarını müstəqil saxlamaqdan yayınaraq buludda, böyük verilənlərin emalı mərkəzində saxlayırlar. Bulud saxlama xidməti tələb üzrə resursların ödəniş əsasında əldə edilməsinə xidmət edən ehtiyat və saxlama modelidir. Bu əsasən böyük sayda verilənlər toplayan və onların saxlanması üçün resursları çatışmayan təşkilatlar üçün sərfəlidir [10, 11].

A. Bulud saxlama növləri

İstifadəçilərə göstərilən bulud saxlamanın müxtəlif formaları mövcuddur. Onlardan aşağıdakıları fərqləndirmək olar:

Şəxsi bulud saxlama (*Personal Cloud Storage*) – və ya mobil bulud saxlama kimi tanınmış saxlama növü buludda informasiyanın fərdi qaydada saxlanması və onlara istifadəçilərin dünyanın istənilən nöqtəsindən daxil olmasını nəzərdə tutur. Bu da verilənlərin çoxsaylı qurğular vasitəsilə sinxronlaşdırılması və mübadilə imkanlarını təmin edir. ICloud şəxsi bulud saxlamanın bariz nümunəsidir.

Ümumi bulud saxlama (*Public Cloud Storage*) – müəssisənin və saxlamanın servis provayderləri ayrıdır və müəssisənin verilənlərin emalı mərkəzində saxlanan heç bir bulud resursu yoxdur. Bulud saxlama provayderi tamamilə müəssisənin şəxsi bulud saxlamalarını idarə edir.

Özəl bulud saxlama (*Private Cloud Storage*) – vahid təşkilat daxilində fəaliyyət göstərən bulud infrastrukturudur. Burada bütün resurslar toplusu yalnız bir təşkilata məxsus olur.

Hibrid bulud saxlama (Hybrid Cloud Storage) –-dövlət və özəl bulud saxlamaların birləşməsidir. Bir sıra qiymətli məlumatlar müəssisənin özəl buludunda saxlanılır. Eyni zamanda digər məlumatlar saxlama provayderilə ictimai buludda saxlanılır [11-13].

B. Bulud saxlamanın üstünlükləri və çatışmamazlıqları

Verilənlərin bulud saxlamalarının başlıca üstünlükləri, ondan ibarətdir ki, aparat və proqram təminatına çəkilən xərclər azalır. Elektrik enerjisi üçün xərclər, serverlər üçün sahə, sistem administratorlarının işi və s. bulud saxlama provayderi tərəfindən aparılır. Verilənlərin bulud saxlama üstünlüklərinə aşağıdakılar aiddir:

- Fərdi komputerlərin hesablama gücünə olan tələbat aşağı düşür (vacib şərt yalnız İnternetə girişin olmasıdır);
- Imtinalara davamlılıq;
- Təhlükəsizlik;
- Verilənlərin qorunması və əlyetərlik saxlama arxitekturuna məxsusdur;
- Verilənlərin emalında yüksək sürət;

- Disk fəzasına qənaət (verilənlər və proqramlar İnternetdə saxlanılır).

Bulud saxlama catışmamazlıqlarına isə aşağıdakıları aid etmək olar:

- Saxlanan verilənlərin bulud hesablamaları xidməti təqdim edən şirkətlərdən asılılığı;
- Yeni bulud inhisarçılarının meydana gəlməsi;
- Müxtəlif bulud saxlama provayderləri arasında qarşılıqlı əlaqəni təmin edən ümumi qəbul edilmiş standartların olmaması [11].

IV. MÖVCUD YANAŞMALAR

Hal-hazırda BV saxlanılması sahəsindəki problemlər (Big Data storage) nüfuzlu beynəlxalq təşkilatlar (IEEE, Storage Networking Industry Association, SNIA v.s.) tərəfindən keçirilən konfransların, simpozziumların, seminarların və forumların əsas mazakirə mövzusu sırasındadır [14, 15].

BV-nin saxlanması problemlərinin həllinə həsr olunmuş çox sayda yanaşmalar irəli sürülmüşdür [16–26].

[16]-da paylanmış saxlama sistemlərində replikasiyanın alternativi kimi təmizləmə kodlarından (*erasure coded)* verilənlərin itki və izafiliyinin qorunması üçün yanaşma təklif edilmişdir. Belə ki, bu kodlar cüzi xərclərlə yüksək elastikliyi təmin edir. Lakin bu təmizləmə kodunda verilənlərin effektiv bərpası və yenilənməsi imkanlarının olmaması onun ən əsas çatışmayan cəhətidir. Reed-Solomon kodlaşmasından istifadə etməklə verlənlərin bərpasının dəyərini kifayət qədər azaltmaq olar.

Bulud saxlamalara tələbat get-gedə sürətlə artır. Belə ki, böyük sayda müəssisələrin və müştərilərin verilənləri buludda saxlanılır və emal edilir. Bu səbəbdən paylanmış obyektlərin verilənlərinin saxlanmasının həlli (məsələn, (Quantcast File System, QFS), OpenStack Swift, (Hadoop Distributed File System, HDFS)) istənilən bulud infrastrukturunun çox mühüm tərkib hissəsini təşkil edir. Bu sistemlər çoxlu sayda serverlər arasında informasiyaları paylamaqla yaxşı etibarlılıq təklif edir. Ancaq bu sistemlərin çatışmayan cəhəti onların vahid verilənlərin emalı mərkəzi daxilində yerləşdirilmək üçün nəzərdə tutulmasıdur. Açıq kodlu əksər saxlama sistemləri üçün geo-replikasiya (verilənlərin emalı mərkəzində lazımsız informasiyzların paylanması) qeyri-asinxron olaraq şəffaf həyata keçirilir. Geo-replikasiyanın çox yüksək dərəcədə etibarlılığı təmin etməsini nəzərə alaraq [17]-də bilavasitə qlobal şəbəkə şəraitində yerləşdirilmiş bu saxlama sistemlərinin həyata keçirilməsinə baxılmışdır. Bu məqsədlə Quantcast-QFS, Swift və Tahoe-LAFS kimi üç populyar paylanmış saxlama obyekti götürülmüş və üç böyük verilənlərin emalı mərkəzində testdən keçirilmiş və qiymətləndirilmişdir.

Sənaye və elmi dairələr üçün [18]-də çoxlu passiv verilənlər saxlayan BV sistemində kodlaşdırma texnologiyası əsasında həll təklif edilmişdir. Kodlaşdırma texnologiyalarının tədqiqi sistemin etibarlılığını qoruyur, həm də təhlükəsizliyi və verilənlərin saxlama sistemlərindən istifadəni yaxşılaşdırır, boş yerlərə qənaət edir. Tədqiqatçılar yükləmə tarazlığını saxlamaq üçün saxlama qovşağını bir neçə virtual qovşağa bölmüşdürlər. Müxtəlif “server codec” üçün müxtəlif virtual qovşaqların saxlama qruplarını yaratmaqla sistemin əlyetərliyi təmin edilmişdir. Qovşağın və verilənlər blokunun şifrinin açılmasında paralel hesablamadan istifadə etməklə sistemin bərpası üçün vaxt azaldılmışdır. Bundan başqa, çoxlu istifadəçilər müxtəlif kodlaşdırma parametrlərindən istifadə etməklə BV-nin saxlama sisteminin etibarlılığını yüksəldir. Müəlliflər müxtəlif m verilənlər bloku və k kalibrləmə bloku yaratmaqla külli miqdarda eksperimentlərdə istifadə əmsalını yaxşılaşdırmışdır. İstifadə edilən ICRS (International Standard Recording Code) şifrləmə ilə kodlaşdırmanın effektivliyi CRS-dən (Cyclic redundancy check) 34,2% və RS (Reed-Solomon) şifrlənmədən isə 56,5% yüksəkdir. ICRS-lə deşifrlənmənin sürəti CRS-dən 18,1%, Rs-dən 31,1 % yüksəkdir.

[19]-da təqdim edilmiş klaster saxlama sistemi adlanan intellektual şəbəkə saxlama diskləri BV-nin saxlanması, əlyetərliyi, mühafizəsi və idarə edilməsi kimi problemləri həll edir. Yaddaş tıxacının (*bottleneck*) aradan qaldırılması üçün ucuz (*cheap*) və dəqiq (*accurate*) kanal işlənilmişdir.

Hal-hazırda verilənlərin buludda təhlükəsiz və effektiv saxlanması bulud hesablamaları sahəsinndə ən böyük cağırışlardan biridir. Buludda saxlanan verilənlərin etibarlı şəkildə qorunmasına zəmanət yoxdur. [20]-də bulud hesablama sistemində təhlükəsizliyi təmin edən etibarlı hesablama mühiti yaratmaq üçün metod təklif edilmişdir. Bu metod istifadəçilərə verilənlərini buludda təhlükəsiz və effektiv saxlamağa imkan verir. BV-nin idarə edilməsi və təhlükəsizlik problemləri verilənlərin bulud saxlamaya yüklənməsi zamanı şifrlənmə və sıxma (*compression*) metodu ilə həll edilir.

BV-nin həcmi və müxtəlifliyi kimi özəllikləri saxlanma və analitikada problemlər yaradır. [21]-də bulud mühitində BV-nin saxlanması üçün universal arxitektura təqdim edilmişdir. Burada klaster analizindən istifadə edilməklə müxtəlif qovşaqlar arasındakı kommunikasiya xətlərinin qiymətinə görə bulud qovşaqları bir neçə klasterə bölümüşdür. Bu məqsədlə istifadəçilər üçün universal saxlama və sorğu interfeysi təmin edən, ən böyük hesablama gücünə malik klaster seçilmişdir. Digər klasterlərin hər biri relyason verilənləri, açar-qiymət (key-value) verilənləri (DynamoDB, Redis, Riak), sənəd (document) verilənləri (MongoDB, CouchDB) və s. kimi konkret modellə verilənlərin saxlanmasına cavab verir.

Bulud hesablamaları və mobil İnternetin inkişafı BV ilə əlaqəli problemlər elmi sahələr kimi sənayenin də diqqət mərkəzinə çevrilmişdir. Mövcud işlərin analizi əsasında BV-nin saxlanmasının qarşılaşdığı problemlərin həllində paylanmış fayl sistemləri, o cümlədən dörd əsas: kiçik faylların saxlanması, yüklənmənin balanslaşdırılması, surət çıxarmanın daimiliyi (*consistency*) və təkrarlanmaların aradan qaldırılması (*deduplication*) üsullları ilə tədqiqat inkişaf etdirilmişdir [22].

Yeni hesablama paradiqması kimi formalaşan bulud hesablamaları, xüsusilə BV nin emalı üçün resurslara çevik və geniş miqyasda əlyetərliyi təmin edir. Burada həll edilməmiş əsas problem verilənlərin buludda səmərəli emalı üçün uzun

müddət ərzində müxtəlif coğrafi məkanlarda yerləşmiş verilənlərin effektiv miqrasiyasıdır. Sərt disklə daşınma yanaşması qıvraq və etibarlı deyildir. Parakəndə verilənlərin emal etmə mərkəzini əhatə edən buluda yönəldilmiş bu iş verilənlərin miqrasiyasına çəkilən xərclərin minimallaşdırılması problemini modelləşdirir və iki onlayn alqoritm təklif edir: verilənlərin emal etmə mərkəzində istənilən vaxt verilənlərin emal və aqreqasiya üçün seçilməsinin minimallaşdırılması; verilənlərin ötürülməsində marşrutun seçilməsi [23].

Verilənlərin tamlığı məzmunla əlaqəli istənilən verilən və hesablamanı praktik olaraq təmin etmək üçün mühüm amildir. O yalnız xidmətin keyfiyyətindən biri kimi xidmət etmir, həm də verilənlərin təhlükəsizliyinin və məxfiliyinin mühüm hissəsidir. Bulud hesablamaların yayılması və BV-nin analizinə artan tələbat verilənlərin, xüsusilə kənar (outsourced) verilənlərin tamlığının yoxlanması daha da zəruridir. Buna görə də verilənlərin tamlığının yoxlanması ilə əlaqədar tədqiqat mövzuları tədqiqatçıların böyük marağına səbəb olmuşdur. Bütün göstəricilər arasında iki ən maraqlı meyar effektivlik və təhlükəsizlikdir. [24]-də autentikliyin təhlili əsasında verilənlərin tamlığının effektiv yoxlanması analiz edilir və bu problemin əsas aspektləri üzrə icmal verilir, tədqiqat metodologiyaları, həmçinin bir neçə xarakterik yanaşmaların əsas nailiyyətləri ümumiləşdirilir.

Yer kürəsi müşahidələri repozitorisinin həcmi hər gün bir neçə terabayt artır. Qiymətli informasiyaları və bilikləri təqdim etmək üçün bu böyük həcmli verilənlərin emalı və saxlanması üçün effektiv metodun yaradılmasına ehtiyac vardır. [25] bulud virtuallaşması əsasında arxitektur həll təqdim edilir və Yer kürəsinin bu təsvirlərinin böyük həcmli verilənlərindən biliklərin əldə edilməsi üçün elastiki (*flexible*) və adaptiv (*adaptive*) metod təklif edilir.

[26]-da BV-nin saxlanması və idarəedilməsi problemləri analiz edilir və həllər təklif edilir.

NƏTİCƏ

Son illərdə verilənlərin həcmi çox böyük sürətlə artaraq. yeni imkanlar və problemlər yaradır. Onların saxlanılması sahəsində problemlər daha da aktuallaşmışdır. Aparılmış tədqiqatlardan da görünür ki. ümumiyyətlə verilənlərin ənənəvi saxlama sistemləri (HDD, SSD və s.) hələ geniş istifadə olunmaqdadır. Bu sahənin istehsalçıları və mütəxəssisləri mövcud saxlama sistemlərinin həcm və sürət kimi xarakteristikalarının yaxşılaşdırılması istiqamətində işlər aparırlar. BV-nin mövcud olması ilə “bulud saxlama” sistemlərinə maraq artmışdır. Bu sistemlərin isə təhlükəsizlik, əlyetərlik, effektivlik və s. kimi problemləri mövcuddur. Problemlərin həllində yeni yanaşmalar, metodlar təklif olunmuşdur. Nəticə olaraq qeyd etmək olar ki, BV saxlanma problemi həmişə olduğu kimi, indi də alimlərin maraq göstərdiyi mövzulardandır.

ƏDƏBİYYAT

[1] R.M.Əliquliyev, M. Ş Hacırəhimova,. “Big Data” fenomeni: problemlər VƏ İMKANLAR // İnformasiya Texnologiyaları Problemləri, 2014, №2, s. 3–16.

[2] T The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe.

[3] Л. Черняк. Большие Данные — новая теория и практика, М.: Открытые системы, 2011, №10.

[4] N. Leavitt, Storage Challenge Where Will All That Big Data Go, Computer? 2013, vol 48, no 9, pp. 22-25.

[5] Е. О. Тютляева, М. М. Тютляев, Системы хранения данных лидирующих суперкомпютеров ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ.

[6] X Fətəliyev, M.İ. İsmayılova, SAN texnologiyasının korporativ şəbəkədə tətbiqi haqqında / PCI 2006 "Kibernetika və informatika problemləri" Beynəlxalq konfransı. Cild II..- B., 2006. - S. 87-89.

[7] T.X Fətəliyev, M.İ. İsmayılova, Korporativ verilənlərin saxlanılmasının bəzi məsələləriI / 3-rd International Conference on Information Technologies and Telecommunication ITTC-2007 - Ganja., 2007. - S. 258-260.

[8] D. Agrawal, S. Das, A. Amr El, R. Big Data and Cloud Computing: Current State and Future Opportunities // EDBT, march 22–24, 2011, Uppsala, Sweden.

[9] Y. Foster, C.Kesselman, S. Tuecke, The Anatomy of the Grid: Enabling Scalable Virtual Organizations // Intern. J. of High Performance Computing Applications, 2001, 15(3), 200–222, www.globus.org.

[10] Big Data Processing using Apache Hadoop in Cloud System, www.researchgate.net.

[11] О.С. Коваленко Обзор состояний, проблем и перспектив хранения и анализа данных в облаке. Информатика, вычислительная теяника и инженерное образование. – 2011. №5(7).

[12] A. Elmustafa Sayed Ali, R.. Ali.. Saeed, A Survey of Big Data Cloud Computing Security // International Journal of Computer Science and Software Engineering (IJCSSE), vol. 3, no 1, 2014, pp: 78-85.

[13] Implementing, Serving, and Using Cloud Storage. www.snia.org.

[14] www.ieeebigdata.org

[15] www.snia.org

[16] K. S. Esmaili, A. Chiniah, A.. Datta, Efficient updates in cross-object erasure-coded storage systems, Big Data, 2013 IEEE International Conference on 2013, pp. 28 – 35.

[17] C. Yih-Farn; S. Daniels, M. Hadjieleftheriou, P. .Liu; C. Tian, V. Vaishampayan, Distributed storage evaluation on a three-wide inter-data center deployment**,** Big Data, 2013 IEEE International Conference on, 2013, pp. 17 – 22.

[18] C. Yin, J. Wang, C. Xie, J. Wan, C. Long, W. Bi, Robot: An efficient model for big data storage systems based on erasure coding. Big Data, 2013 IEEE International Conference on, 2013 , pp. 163 – 168.

[19] L Haixia, L. Chuiwei , S. Sheng, Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014 International Conference on, 2014, pp. 1 – 6.

[20] A. Kumar, L.Hoonjae; R. P. Singh, Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in, 2012, pp. 162 – 166.

[21] Q. Zhang, Z. Chen, A. Lv, L. Zhao, F. Liu, J. Zou, Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, 2013, pp.476 – 480.

[22] X. Zhang, F. Xu, Distributed Computing and Applications to Business, Engineering & Science (DCABES), 2013 12th International Symposium, 2013, pp. 76 – 80.

[23] J. Li, Z. Xu, Y. Jiang, R. Zhang, The overview of big data storage and management, Cognitive Informatics & Cognitive Computing (ICCI*CC), 2014 IEEE 13th International Conference on, 2014, pp. 510 – 513.

Применение Технологий Big Data в Медицине

Гюляра Мурадова
Азербайджанский Технический Университет
gularamu@aztu.edu.az

Абстракт – в статье рассматриваются вопросы применения больших данных в медицине и биомедицинских исследованиях. Приведены примеры, демонстрирующие как анализ больших массивов данных может эффективно снизить проблемы в сфере здравоохранения. Показаны перспективные направления применения Big Data в медицине.

Ключевые слова – источники Big Data, анализ “больших данных”, персональная медицина, носимая электроника для диагностики, биомаркер, трехмерная цифровая медицина

I. ВВЕДЕНИЕ

Мир больших данных радикально изменяет все и применение технологий Big Data позволяет не просто трансформировать применяемые ранее методы, но и найти новые модели принятия решений, которые раньше были попросту невозможны. Устаревшие методы анализа и сбора данных не соответствуют современным требованиям, поэтому в настоящее время наблюдается повышенный интерес к ИТ-решениям в Big Data. Развитие технологий больших данных и адаптация под новые задачи является целью изучения Big Data. Постоянное накопление и анализ больших массивов данных позволяет создавать модели, с которыми можно проводить эксперименты, т.е. исследователи, извлекая выгоду из данных, получают результаты, представляющие научную ценность. Анализируя большие массивы данных, установив закономерности, можно создать модель объекта, которую можно использовать в самых разных целях. С помощью модели можно изучать свойства уже описанного реального объекта, а также делать вероятностные прогнозы. Эффективность использования больших объемов данных поможет улучшению системы здравоохранения, в частности при выборе соответствующих путей лечения, диагностики и т.д.[1]. Для того чтобы справиться с огромным объёмом данных, необходимы новые решения. Анализ “больших данных” может улучшить планирование медицинского обслуживания. Big Data требует пересмотра существующего подхода с учетом новых технологий. Количество данных в таких областях, как биология, медицина и генетика увеличивается с возрастающей скоростью. Медицина является одной из отраслей, которая становится потребителем технологий больших данных. Источники «больших данных» могут быть разные данные. С медицинской точки зрения источниками данных являются:

1. Данные соц. медиа: Facebook, Twitter, LinkedIn, медицинские веб-сайты, приложения и т.д.

2. Данные отслеживающих устройств: с удаленных датчиков, томографов, счетчиков и др.

3. Данные транзакций: данные страховой медицины, биллинговые операции и др.

4. Биометрические данные: отпечатки пальцев, почерк, сетчатка глаза, рентгеновские и другие медицинские изображения, артериальное давление, пульс др.

5. Данные, генерируемые человеком: заметки врачей, бумажные документы, рецепты врачей.

Электронная форма работы в медицинских учреждениях сделала возможным отслеживание взаимосвязи между принимаемыми препаратами, их эффективностью, а также другие факторы, которые столь важны для фармакологических компаний. Главной задачей использования больших данных является сокращения времени и средств при разработке новых препаратов и вывод их на рынок за счет многофакторного анализа их эффективности. Теперь врач в реальном времени изучая данные, полученные от множества специалистов, занимающихся лечением пациентов с единичными заболеваниями технологий может позволить себе полагаться не только на собственные знания. А это позволит намного лучше анализировать эффективность применяемых методов лечения. Примеры использования технологий больших данных это создание регистров пациентов с наследственными заболеваниями для счета объемов и потребностей в препаратах, для упрощения диагностики, для поиска скрытых зависимостей. Существующие множество ограничений на распространение данных о пациентах, накладывает определенные ограничения на возможность такого анализа [2].

Для успешного принятия решения о внедрении решений Big Data необходимо четкое планирование первоначальных действий. Культурный уровень сводится к осознанию возможностей, которые открываются при извлечении новых знаний из накопленных данных, и, как следствие, увеличению когнитивной способности человечества. Научный уровень образован математическими методами анализа данных, позволяющими совершать открытия в химии, биологии, медицине и других дисциплинах. Технологический уровень предполагает переориентацию ИТ со счетных

задач как основного приложения на сбор, хранение, обработку больших массивов данных с поддержкой тех или иных аналитических методов[3].

II. НАПРАВЛЕНИЯ ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ BIG DATA В БИОМЕДИЦИНСКИХ ИССЛЕДОВАНИЯХ

Современная медицина превращается в науку, где большие объемы информации о пациенте позволяют специалистам вести персонифицированное лечение. Это новая модель организации медпомощи, основанная на выборе диагностических, лечебных и профилактических средств, которые являются оптимальными для конкретного пациента, с учётом его генетических, физиологических, биохимических, поведенческих и других особенностей. Исходя из этого мохно утверждать, что теперь любой процесс в организме человека можно наблюдать и измерять. Медицине нужны удобные инструменты для обработки, для хранения, для передачи массива данных. Модель организации медицинской помощи людям за счет рационального подбора лекарств актуальное направление технологий Big Data. Геномный Проект (PGP), использующий большие данные — поставила перед собой цель секвенирования или иными словами «прочесть» молекулу, 100000 добровольцев, чтобы направить исследования на персональную геномику и персональную медицину [4].

Это исследование разработано в Гарвардской Медицинской Школе. В рамках проекта планируется выполнить расшифровку персональных геномов множества добровольцев, их генетические данные вместе с записями из медицинских карт будут доступны для дальнейшего анализа. В рамках проекта расшифровка персональных геномов добровольцев, генетические данные которых вместе с записями из медицинских карт будут доступны для дальнейшего анализа [5,6]. Применение технологий Big Data в биоинформатике позволяет использовать методы компьютерного анализа в сравнительной геномике. ПроектGenePool создан для поиска биомаркеров в геноме человека, а также анализа и хранения полученной в ходе исследований информации[7]. С помощью биомаркеров возможно оценить общее состояние организма, выявить отклонения и уничтожить конкретный вредоносный ген. Для определения наличия или отсутствия раковых клеток вводится в организм специальное вещество, в состав которого входит онкомаркер, затем оно выявляет или не выявляет зараженные клетки. Биомаркер может определяет и степень поражения тканей, и область распространения метастазов вокруг них. В результате эти вещества позволяют точно скорректировать зону лечения и избежать удаления здоровых клеток.

Проект GenePool использует для поиска биомаркеров последовательности ДНК молекулы, объясняющей ее строение. Один человеческий геном содержит в себе около трех миллиардов символов GenePool систематизирует весь огромный поток полученных данных. [8]. Применение технологий Big Data, расшифровывающей этот массив данных и выявляет закономерности в геномном мире. С помощью технологий Big Data учёные проанализировали геномы многих тысяч человек, больных и здоровых. Основываясь на анализе огромного количества информации генетического кода, система подбирает лечение. Так, например, болезнь Альцгеймера зависит от генов, участвующих в развитии рассеянного склероза и паркинсонизма[9,10,11].

В определении наиболее эффективных в клиническом аспекте и рентабельных применяемых методов лечения может помочь анализ больших данных. Для фармацевтических и медицинских секторов обнаружение заболевания на ранних стадиях дает возможность моделировать будущий спрос на свою продукцию. Это является основой для принятия решений об инвестициях в научные исследования и разработки. Важнейшей задачей повышения уровня здравоохранении является анализ глобальных закономерностей заболеваемости с целью выявления тенденций закономерностей заболеваемости на ранней стадии [12].

Фармацевтические компании, используя потенциал «больших данных» сокращают время и средства при разработке новых препаратов на этапе их вывода на рынок и затраты на проверку заведомо неперспективных компонентов при разработке лекарств. [13]. Благодаря технологиям «Больших данных» Германии, например, по анализу крови пациентов и доноров выявляются онкологические заболевания либо предрасположенность к ним. В результате своевременной диагностики существенно повышается эффективность лечения и снижаются затраты государства и самих людей. Российские разработчики «Элайд Инновейшенс» занимаются развитием проекта интеллектуальной компьютерной системы для прогноза течения онкологических заболеваний на базе технологий обработки больших данных. Помогают врачам выбирать наиболее эффективную методику лечения в каждом конкретном случае[14]. Глобализация в мире привело к тому, что население становится мобильнее, а географические границы — более размытыми и для медицинских работников всё более важно иметь возможность получить доступ к данным приезжих пациентов. Анализ и обработка “больших данных”, относящихся к определённым регионам и группам населения, улучшит планирование медицинского обслуживания. Раньше медицинские учреждения в основном имели дело с отдельными компаниями, которые разрабатывали и поставляли им аппаратуру, программное обеспечение к ней, создавали прикладные решения «местного значения». Сейчас в эту сферу стали активно вторгаться государственные организации и им требуется доступ к данным о пациентах в любой точке и в любое время. Например, в Бостоне в местном госпитале хранится история болезни. Но если в Нью-Йорке попасть в аварию, то врачу скорой помощи будет очень важна информация о хронических болезнях или аллергии на лекарства.

Быстрый доступ к таким данным поможет оказать квалифицированную помощь. В США есть программа по созданию региональной инфраструктуры здравоохранения. InterSystems участвовала в реализации этой программы в Нью-Йорке. В созданной системе содержатся записи по 25-30 млн пациентов обо всех случаях взаимодействия пациента с врачами.

При этом также важным является распознавать речь (неструктурированные данные) и превращать ее в текст (структурированные данные). Объемы данных получаются гигантскими, так как человек говорит примерно в три-четыре раза больше, чем пишет. InterSystems предлагает технологии, позволяющие фильтровать сырые данные и извлекать из них информацию, имеющую ценность[15]. Эта технология применима и для психически неполноценных людей. Так, например, можно обнаружить депрессию, так как депрессивные лица, как правило, говорят более медленно. Люди все больше занимается самолечением. Развиваются открытые информационные системы для самодиагностики. Все больше и больше людей приобретает носимую электронику, а различные мобильные девайсы постоянно развиваются и наращивают функционал (уровень сахара, частота сердечных сокращений, давление, активность, диагностические анализы и прочее). В здоровых людей могут встраиваться USB-порты для диагностики. Уже сейчас в сттранах Запада анализы делаются дома самостоятельно, клиническая картина сохраняется в облаке, а общение с врачом происходит дистанционно – через интернет. Больные диабетом, могут использовать различные устройства с подключением к сети, которые помогли бы людям периодически отправлять такие данные, как уровень сахара в крови. Таким образом, посредством внутренних датчиков человек в режиме реального времени отслеживает свое состояние и программа контролирует уровень сахара в крови пациента. Агрегировать (объединять полученные данные) данные об уровне сахара в крови, данные о питании и состоянии здоровья, полученные из различных источников - это возможность получения медицинской документации, необходимой врачам и диабетологам. Комбинируя эти данные с возможностями аналитических инструментов, предоставляется возможность миллионам больных диабетом получить более качественного медицинского обслуживания[16].

Внедрение и развитие подобных датчиков порождает огромное количество данных, целый поток, который необходимо оперативно не просто обрабатывать, но и хранить для дальнейшего анализа.

Другой пример обработки больших потоков данных в реальном времени. Цель медицинского центра Колумбийского университета состоит в том, чтобы обеспечить своевременной информацией медицинских работников. Если в режиме реального времени отслеживать свое состояние больных, то передовая аналитика диагностирует серьезные осложнения о повреждении аневризмы головного мозга на целых 48 часов раньше[13].

В настоящее время функционирует система Watson претендующая на замещение человека при принятии решений в цифровом мире система такого класса, реально претендующая на замещение человека при принятии решений в цифровом мире. Watson является платформой IBM BIG DATA. Watson реализован на идеях массово-параллельных вычислений, адаптированных под вычислительную платформу. С 2006 года здесь работает огромный совместный коллектив медиков и програмистов. Более 100 различных методов используются для анализа естественной речи, определения источников данных, генерации гипотез, поиска и ранжирования фактов, объединения и подтверждения гипотез. В медицине Watson помогает анализировать специфические симптомы, историю болезни, историю наследственных заболеваний, и выполнять синтез и обработку этих данных с доступной неструктурированной и структурированной медицинской информацией, включая медицинские книги и статьи[17]. Применяются технологии Big Data для обработки данных геопозиционирования, которые накапливаются операторами сотовой связи. Alex Pentland и David Laser из Массачусетского технологического института, в 2009 году обосновали, что анализ геоданных не менее информативен для прогноза социально значимых заболеваний, чем геном. В происхождении мультигенного заболевания значительную роль играют факторы внешней среды, которые называют экспозом. Анализируя перемещения человека с достаточной степенью детализации можно получить информацию характерных для определённого синдрома, заболевания, например, с развитием сердечно-сосудистых заболеваний или ожирения[14].

Основная задача системы трехмерной цифровой медицины — обеспечение работы с параметризованной, максимально подробной (вплоть до тканевого и клеточного уровней) 3D-моделью органов и систем человека в норме и в патологии. Это обеспечивает адресность оказания лекарственной помощи пациентам. Рост во всем мире числа раковых заболеваний — главная причина применения томографии и технологий трехмерной медицины[18]. Процесс биомеханического анализа любой биологической системы на основе разработанной физико-математической модели требует ввода данных о пространственной геометрии объекта. Для работы с полученными изображениями, современные компьютерные томографы оснащаются специальным программным обеспечением, применяемые в области лучевой терапии и онкологии[19].

III. СЛОЖНОСТИ ПРИ РАБОТЕ С BIG DATA

Множество ограничений на распространение данных о пациентах существует в медицине. Даже в США не все учреждения здравоохранения готовы к работе с большими

данными. По результатам опроса, который проводила MeriTalk среди топ-менеджеров из сферы государственного здравоохранения США менее 25% считают, что их ведомства готовы работать с большими данными, 59% уверены, что успешная работа с большими данными станет фактором повышения эффективности их учреждения в ближайшие годы, а 63% опрошенных считают, что технологии больших данных позволят более эффективно отслеживать состояние здоровья пациентов[2]. По оценкам IBM, в мире ежедневно генерируется 2,5 квинтиллиона (2,5*1018) байт данных. Большая часть из них – это Big Data. Экспоненциально увеличился и объем данных, управляемых компаниями по всему миру. При отсутствии четкой структуры и политики использования информация беспорядочно нагромождается, и, как следствие, организациям становится сложно управлять ею. «Большие данные» не гарантируют «большой смысл» них отсутствует структура корректность достоверность [20]. По определению Майкла Франклина из Университета в Беркли «Большие данные – это любые данные, работа с которыми требует значительных затрат и из которых трудно извлечь информацию»

Ряд вопросов, которые приходится решать при работе с большими данными:

1. Большую часть данных в биологии и медицине состовляют сложные для анализа из-за неполноты описания явления. Необходима творческая достройка аналитической модели при помощи гипотез. Следует развивать и логику, и оптимальные стратегии работы.
2. Вопрос защиты персональной информации клиентов. Позволят ли люди обрабатывать компаниям их персональные данные и вообще всю поступающую от клиентов информацию.
3. Нужны технологии, позволяющие повысить предсказуемость затрат времени, ресурсов, а также итоговую результативность работы дата-центров. Значительное присутствие непредсказуемых и случайных факторов особенно в медицине, где многое меняется «на лету». Если клиенты генерируют большое число запросов, то это приводит к совместному использованию дата-центров, что приводит к непредсказуемости их работы[21].

Таким образом, для организации эффективного управления данными необходимо комплексное решение всех этих задач.

ЗАКЛЮЧЕНИЕ

Анализ больших потоков данных из самоотчетов и личных медицинских записей, в том числе медицинских тестов, прописанных лекарств, а также томографов, всевозможных датчиков, дает возможность резкого улучшения в сфере здравоохранения. Количество информации, требующей анализа, продолжает непрерывно расти. Доступность, непрерывность, удобство использования, масштабируемость, способность управлять на различных уровнях детализации, конфиденциальность и безопасность решений, а также контроль качества данных имеют важное значение для эффективного управления данными. Применение технологии «больших данных» крайне востребовано, так как последние смогут решить актуальные задачи в медицине и ряде смежных областей. Учитывая экспоненциальный рост объёмов данных и новые средства их расширенного анализа возможно преобразование системы здравоохранения. Поэтому применение современных компьютерных методов решения этих проблем в будущем значительно увеличит продолжительность жизни. Применение современных компьютерных методов решения этих проблем в будущем значительно увеличит продолжительность жизни.

ЛИТЕРАТУРА

[1] Kyoungyoung Jee, Gang-Hoon Kim Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System / Healthcare Informatics Research, *Korea*, 2013, v.19, Issue 2

[2] http://www.cnews.ru/reviews/new/bi_bigdata_2014/articles/chto_bolshie_dannye_mogut_dat_biznesu/

[3] Черняк Л. Серьезно о технологиях для Больших Данных/Открытые системы http://www.osp.ru/os/2014/01/13039646/

[4] http://en.wikipedia.org/wiki/Personal_Genome_Project

[5] Васильков А. Медицина становится персонализированной под влиянием «больших данных» , http://www.computerra.ru/88609/big-data-improve-personalized-medicine/

[6] 6.http://rusangels.ru/naba/pressroom/smi-about-naba/v-kakie-startapy-rekomenduyut-vkladyvat-rossiyskie-investory.html

[7] http://www.stationxinc.com/how-genepool-works

[8] http://mygenome.su/articles/112/

[9] Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease // Nature genetics, http://www.nature.com/ng/journal/v45/n12/full/ng.2802.html

[10] Swarup V.Geschwind D.H. Alzheimer's disease: From Big Data to mechanism // Nature, 2013, v500

[11] Bettens Karolien, Sleegers Kristel, Van Broeckhoven, Genetic insights in Alzheimer's disease, 2013, v12, Issue 1

[12] "Большие данные", большое дело, большая проблема, ITU News https://itunews.itu.int/Ru/Note.aspx?Note=4879

[13] Wullianallur Raghupathi , Viju Raghupathi Big Data analytics in healthcare: promise and potential / Health Information Science and Systems, 2014,v2

[14] Трифонова О. П.,Ильин В. А.«Большие данные» в биологии и медицине/Acta Naturae / Медицина и здравоохранение, ГРНТИ: Научная библиотека, 2013, № 3

[15] Большие Данные в медицине и здравоохранении http://www.osp.ru/medit/2014/04/13040834.html

[16] Новости и статьи с тегом Сколково http://leptandra58.rssing.com/chan-33366061/all_p1.html

[17] Шмид А. Big Data http://www.interface.ru/home.asp?artId=34991

[18] Турлапов В., Гаврилов Н.,Сапрыкин. В. Платформа цифровой медицины «Открытые системы», № 05, 2014http://www.osp.ru/os/2014/05/13041824/

[19] Skripcak T., Belka C., Bosch W. Creating a data exchange strategy for radiotherapy research: Towards federated databases and anonymised public datasetsRadiotherapy and Oncology, 2014,v 113, Issue 3

[20] http://datareview.info/article/10-zadach-kotoryie-ne-smogut-reshit-bolshie-dannyie/

[21] Microsoft о будущем интерфейсов и Big Data http://ria.ru/technology/20130927/966349131.html

Big Data в Электронном Образовании: Большие Возможности и Большие Риски

Фирудин Агаев[1], Гюлара Мамедова[2]
[1,2]Институт Информационных Технологий НАНА, Баку, Азербайджан
[1,2]*depart10@iit.ab.az*

Аннотация **− в статье показаны уникальные возможности применения Big Data в электронном образовании. Рассмотрены риск и препятствия, связанные с использованием больших объемов данных в сфере образования.**

Ключевые слова *− электронное образование, персонализация электронного обучения, преимущества больших данных в электронном образовании, риски больших данных в электронном образовании.*

I. ВВЕДЕНИЕ

В октябре 2013 года в Катаре прошел Всемирный саммит инноваций по образованию (WISE) с призывом провести радикальные перемены. В саммите участвовали политики, члены правительств и ведущие педагоги из более чем 100 стран мира. Основным призывом саммита было утверждение о необходимости подготовки молодых специалистов к миру завтрашнего дня. Одна из тем, которая вызвала значительный интерес у делегатов, была посвящена большим данным (Big Data). Руководитель саммита Джон Фэллон [1] высказал мнение, что “большие данные могут привести и привели к одному из самых творческих периодов в истории с точки зрения инноваций. Времени использования их в сфере образования”. Big Data обозначает большой объем информации, которая поступает в электронное образование каждую секунду через различные каналы, как правило, онлайн. Это − данные, которые являются слишком большими, сложными и динамичными. Обычными способами их трудно улавливать и управлять ими. Термин возник в открытых сетях, где специалисты пытались найти быстрые и более масштабируемые решения для хранения и обработки огромных объемов данных. Благодаря достижениям в информационных технологиях эти данные теперь могут быть интерпретированы и проанализированы, обеспечивая большую пользу в различных сферах человеческой деятельности, в том числе и в электронном образовании. [3] Данные "большими" (и по размеру, и по значимости) делает то, что они позволяют анализировать и прогнозировать поведение обучаемого через огромное разнообразие демографических, личностных показателей, обучающих стилей, процессов, уровней интеллекта, академических намерений, генетической предрасположенности, навыков и т.д. [4]

II. КАКОВЫ ПРЕИМУЩЕСТВА БОЛЬШИХ ДАННЫХ В ЭЛЕКТРОННОМ ОБРАЗОВАНИИ?

В сфере образования большие данные в настоящее время используются, чтобы помочь разобраться в стратегиях обучения, оценить влияние этих стратегий на студентов и преподавателей, на основе фактических данных сформировать научно-обоснованный подход для экспериментирования, а также создать персонализированную среду обучения.

Термин "большие данные в сфере электронного образования" охватывает три аспекта: объем, скорость и разнообразие:

1) Большой *объем* данных может дать информацию о тысячах обучающихся, изучающих один и тот же курс или имеющих одинаковый учебный опыт. Эта информация с течением времени также может пролить свет на некоторые данные об одном ученике. Из-за своей масштабируемости большие данные могут когда-нибудь собрать данные обучения сотен учебных организаций, чтобы обеспечить глобальную перспективу образования.

2) *Скорость* изменения больших данных в образовании позволяет учащимся и преподавателям получать быстрый доступ к данным в режиме реального времени. Представьте себе студента, дающего неправильный ответ на экзамене. Скорость мгновенно обеспечит исправление и обогащение вариантов теста, основанных на предыстории и закономерностях обучения и успешных стратегий от тысяч других учеников, которые тоже неверно ответили на этот вопрос. Это также позволит преподавателям на постоянной основе внести коррективы в доставку контента на основе экспресс-анализа пользовательского опыта.

3) Большие данные в образовании дают *разнообразную* информацию о студентах с различными уровнями подготовки. Это позволяет видеть, как зависит выполнение заданий студентами от уровня их подготовки. Без этой разнообразной информации мы имеем ограниченное представление о студентах.

III. ПРЕИМУЩЕСТВА, КОТОРЫЕ БОЛЬШИЕ ДАННЫЕ ПРЕДЛАГАЮТ ПРЕПОДАВАТЕЛЯМ КУРСОВ ЭЛЕКТРОННОГО ОБУЧЕНИЯ

Есть множество преимуществ [2], которые большие данные предлагают для специалистов, разрабатывающих

курсы электронного обучения. Вот лишь некоторые из наиболее значительных преимуществ, связанных с большими данными:

A. Отслеживание информации обучающихся в социальных сетях

Большие данные обеспечивают анализ того, какие модули электронного обучения наиболее посещаемые и на какие разделы курса учащиеся поместили ссылки в социальных сетях, например на Facebook.

B. Своевременность полученных данных

Применение больших данных позволяет получать информацию почти сразу же, вместо того чтобы ждать в течение длительного периода времени. Это означает, что педагоги не будут долго ждать, чтобы точно настроить свою стратегию обучения, они смогут приступить к реализации изменений или использованию информации сразу.

C. Улучшение курса обучения

Большие данные дают возможность специалистам электронного обучения точно определить области, которые, возможно, должны быть точно настроены в ходе выполнения электронного курса или модуля. Например, если некоторая часть учащихся тратит чрезмерное количество времени, чтобы закончить конкретный модуль, это, вероятно, означает, что модуль должен быть улучшен и следует сделать его более приемлемым для них. На основании шаблонов поведения учащихся педагоги имеют возможность предсказывать, где учащиеся могут отставать или преуспевать. Позволяет им понять, как учащиеся усваивают информацию и нуждаются ли некоторые из них в помощи. Тем самым они смогут развивать свои курсы электронного обучения таким образом, чтобы учащиеся могли получить шанс для достижения наилучших результатов. [5]

D. Персонализация данных

Большие данные позволят специалистам разрабатывать более персонализированные курсы электронного обучения. Если они будут иметь возможность узнать, как их учащиеся получают информацию и что работает лучше всего для них с точки зрения содержания и доставки, то это приведет к разработке более персонализированных и привлекательных курсов электронного обучения. Таким образом, модули могут быть специально адаптированы для удовлетворения на высоком уровне индивидуальных потребностей обучаемого.

IV. ОРИЕНТАЦИЯ НА ЭФФЕКТИВНЫЕ СТРАТЕГИИ И ЦЕЛИ ЭЛЕКТРОННОГО ОБУЧЕНИЯ

Большие данные в электронном обучении дают нам увидеть то, какие стратегии электронного обучения работают, а какие не обязательно полезны в плане достижения целей обучения. Например, можно определить, какие курсы электронного обучения способствуют развитию навыков, а какие модули или элементы, может быть, следует исключить из курса подготовки. Можно выяснить, на разработку каких модулей нужно выделять ресурсы, чтобы учащиеся смогли получить подготовку, необходимую для выполнения своих карьерных целей.

V. ОТСЛЕЖИВАНИЕ МОДЕЛЕЙ УЧЕНИКА

Используя большие данные, специалисты электронного обучения получают редкую возможность отслеживать учащихся в течение всего процесса, от начала и до конца. Другими словами, можно увидеть, насколько хорошо они выполняют задания тестов или как быстро они закончили сложные модули курса. Это поможет больше узнать о поведении не только отдельно взятого учащегося, но и групп обучаемых в целом.

VI. РИСКИ БОЛЬШИХ ДАННЫХ В ЭЛЕКТРОННОМ ОБРАЗОВАНИИ

Попав в чужие руки, большие данные могут принести больше вреда, чем пользы. Независимо от того, намеренно или без умысла были совершены противоправные действия, простым щелчком кнопки можно подорвать всю систему образования.

Вот некоторые из рисков и препятствий, связанных с использованием больших объемов данных в сфере образования:

1) *Конфиденциальность.* Компании, как, например, Google, предлагают такие услуги, как электронная почта, хранение документов и их обработка, новости, веб-браузер, отслеживание местоположения, видео и обмен фотографиями, голосовая почта, магазины, социальные сети и все остальное. Все это может представлять интерес для своих, а также и сторонних пользователей. Даже информация, которая доступна для общественности, когда она собрана из различных источников в совокупном досье, может создать ясную картину человека. Простой поиск Google может дать огромное количество информации о человеке, хотя точность многого из этого находится под вопросом.

2) *Дегуманизация.* Помимо очевидных возможностей для ошибок и предвзятости, использование профиля личности нежелательно, потому что это обезличивает тех, кого оцениваем, а также тех, кто принимает решения. Расчет заменяет человеческую рассудительность о том, что некоторая информация может быть очень чувствительной для человека и затрагивает сугубо личные интересы. Т.е. в этом случае к человеку относятся как к объекту или коллекции фактов.

3) *Обман в цифрах.* Профессор психологии из Лондонского университета Сирил Людовик Берт был человеком, ответственным за внедрение стандартизированного экзамена в Великобритании. Впоследствии он был дискредитирован из-за публикаций в журнале, в котором сам являлся редактором, из-за того, что фальсифицировал не только данные, на которых

основывал свою работу, но и коллег по исследованию. Многие утверждают, что стандартизированные тесты стали разрушительной силой в образовании. Ограниченная академическая оценка стала почти навязчивой идеей в некоторых странах [6].

ЗАКЛЮЧЕНИЕ

Инновационные технологии революционным образом меняют практику электронного образования. Использование больших данных является следующим логическим шагом в этом эволюционном процессе. Теперь у нас есть доступ к объемам больших данных, и мы должны разобраться в том, чем они могут нам помочь, в каком направлении необходимо проводить дальнейшие исследования и какие потенциальные проблемы и сложности могут возникнуть при работе с большими данными.

ЛИТЕРАТУРА

[1] www.wise-qatar.org/2014-wise-summit#5

[2] В.Майер-Шенбергер, К.Кукьер, “Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим”, 2014.

[3] J.Wassan, Discovering Big Data Modelling for Educational World Maitreyi College”, University of Delhi, New Delhi-21, India, IETC 2014.

[4] M.West, Big Data for Education: Data Mining, Data Analytics, and Web Dashboards, 2012.

[5] L. Maacfadyen and S.Dawson, Mining LMS Data to Develop an ‘Early Warning System’ for Educators, Computers & Education, Volume 54, 2010.

[6] U.S. Department of Education Office of Educational Technology, Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics", , 2012.

[7] F.Castro, A.Vellido, A.Nebot and F/ Mugica, Applying Data Mining Techniques to e-Learning Problems", Studies in Computational Intelligence, Volume 62, 2007, pp. 183−221.

[8] J.Manyika, M.Chui, B.Brown, J.Bughin, R. Dobbs, C. Roxburgh and A. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, May, 2011.

İnkişaf Etmiş Ölkələrin “Big Data” Təşəbbüsləri

Məkrufə Hacırəhimova[1], Aybəniz Əliyeva[2]
[1,2] AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*makrufa@science.az*, [2]*aybeniz63@rambler.ru*

***Xülasə* — Məqalə bəzi inkişaf etmiş dövlətlərin “big data” strategiyalarının tədqiqinə həsr olunur. Məqalədə ABŞ, Avstraliya, Cin, Koreya, Böyük Britaniya və s. kimi inkişaf etmiş dövlətlərin “big data” təşəbbüsləri şərh edilir.**

Açar sözlər — big data; big data texnologiyaları; təşəbbüslər; strategiya; analitika; Digital Roadmap; kiberfəzanın təhlükəsizliyi

I. GİRİŞ

Müasir informasiya texnologiyalarının geniş istifadəsi sayəsində sürətlə artan böyük həcmli verilənlər yığımı elm, sənaye, biznes və s. kimi sahələrlə yanaşı dövlət dairələrinin də diqqətini cəlb etmişdir [1]. Artıq bir çox ölkələr milli təhlükəsizlik, sosial-iqtisadi inkişaf, səhiyyə və s. kimi və sahələrdə böyük verilənlərin (BV) imkanlarından istifadə etmək təşəbbüsləri ilə çıxış edirlər. Gələcəkdə ölkələr arasındakı iqtisadi və siyasi rəqabət böyük verilənlərin potensial istifadəsinə əsaslanacaqdır. Başqa sözlə, böyük həcmli verilənlərin tədqiqi və tətbiqi istənilən ölkənin rəqabətqabiliyyətliliyinin artması üçün zərurətə çevriləcəkdir [2]. Ona görə də dövlətlərin bu sahədə qəbul etdikləri strategiyaların, təşəbbüslərin araşdırılması vacibdir.

II. BİG DATA BEYNƏLXALQ TƏŞKİLATLARIN GÖZÜ İLƏ

“Böyük verilənlər” mövzusu beynəlxalq səviyyədə müzakirə olunan mövzulardandır. Mütəxəssislər tərəfindən informasiya kommunikasiya texnologiyaları sahəsinin lokomotivi adlandırılan və “yeni neft” kimi dəyərləndirilən (BV) strateji əhəmiyyətə malik resurs kimi nüfuzlu beynəlxalq təşkilatların diqqət mərkəzindədir. Son illər beynəlxalq təşkilatlar tərəfindən BV-yə yüksək dəyər verilmiş, onun potensial imkanlarından istifadə sahəsində bir sıra strateji planlar və layihələr qəbul edilmişdir. 2012-ci ildə Davosda keçirilən Dünya İqtisadi Forumunda BV valyuta və ya qızıl dəyərində yeni iqtisadi aktiv kimi dəyırləndirilmiş, beynəlxalq inkişaf üçün BV-nin imkanlarını təqdim edən sənəd hazırlanmışdır [3].

2009-cu ildə Birləşmiş Millətlər Təşkilatı tərəfindən “Global Pulse" təşəbbüsü irəli sürülmüşdür. Təşəbbüs ölkələrin inkişafı, yoxsulluğun, böhranlı vəziyyətlərin aradan qaldırılması, insanların yaşayış səviyyəsinin yüksəldilməsi sahəsində BV-nin imkanlarından istifadəni nəzərdə tutan bir konsepsiyadır [1, 4]. İqtisadi Əməkdaşlıq və İnkişaf Təşkilatının “İnformasiya, Kompüter və Kommunikasiya Siyasəti üzrə Komitə”si (OECD - Committee for Information, Computer and Communications Policy) tərəfindən 2012-ci ilin oktyabrında təqdim edilən “Exploring data-driven innovation as a new source of growth“ hesabatda əhəmiyyətli dərəcədə rəqabət üstünlüyünün yaradılması, innovasiyaların, iqtisadiyyatın və cəmiyyətin dayanıqlı inkişafını stimullaşdıra bilən biliklərə əsaslanan kapitalın formalaşdırılması üçün verilənlərin və onların analizinin potensial rolu tədqiq olunur [5]. Hesabatda verilənlərin istifadəsinin onlayn reklam, dövlət idarəçiliyi, səhiyyə, kommunal xidmətlər, loqistika və nəqliyyat kimi sahələrdə innovasiya və əməyin məhsuldarlığının yüksəldilməsi mənbəyi kimi postensial dəyərə malik olduğu göstərilir [5]. Avropa Komissiyası elmi tədqiqatlar və innovasiyalar sahəsində Horizon 2020 proqramını qəbul etmişdir. Proqramın icrası üçün təxminən 80 milyard avro maliyyə vəsaiti ayrılmışdır. Proqram çərçivəsində mükəmməl elm, sənayenin idarə olunması və sosial problemlər də daxil olmaqla “big data” məsələlərinin uğurla həyata keçirilməsi üçün BV-yə €120 milyon vəsait ayrılmışdır [6]. Dünya Bankı da böyük verilənlərin imkanlarını geniş tədqiq edir, yardım və koordinasiyanın səmərəliliyinin yüksəldilməsi, şəffaflığın və sosial məsuliyyətin artırılması kimi bir çox məsələlərdə BV-dən istifadəni nəzərdə tutmuşdur [7].

2013-cü ilin noyabrında Beynəlxalq Telekommunikasiya İttifaqının “Böyük həcmli verilənlər: bu gün böyükdür, sabah normal” (Big data - big today, normal tomorrow) adlı hesabatı nəşr olunmuşdur. Hesabatda BV-nin səhiyyə, elmi tədqiqat və nəqliyyat sahələrində tətbiqi və verilənlərin tətbiqiylə bağlı təhlükəsizlik məsələləri yer almışdır [8]. 2009-cu ilin sentyabrında Avropa komissiyası tərəfindən qəbul olunmuş iKnow (Interconnect Knowledge) layihəsi Avropada elmi-tədqiqatların dərinləşməsinə istiqamətlənmişdir. Avropa Birliyinin bu layihə çərçivəsində zəlzələ, sunami, terrorizm, şəbəkə və qlobal böhranlarla bağlı apardığı tədqiqatlar BV-nin imkanlarından istifadəyə əsaslanmışdır [9].

III. BİG DATA DÜNYA ÖLKƏLƏRİNİN SİYASƏTİNDƏ

BV beynəlxalq təşkilatlar kimi dünyanın bir çox qabaqcıl ölkələrinin də diqqətini cəlb etmişdir. Bu istiqamətdə ilk təşəbbüs Amerika Birləşmiş Ştatları tərəfindən irəli sürülmüşdür. Sonralar bir sıra qərb dövlətləri, Avstraliya, Çin, Yaponiya və s. dövlət və hökumət strukturları tərəfindən BV neft qədər strateji resurs kimi dəyərləndirilmiş, bu sahədəki problemlərə çox böyük önəm verilmişdir.

A. ABŞ-ın Big Data təşəbbüsü.

Amerikada BV-nin tədqiq və inkişaf səviyyəsindən artıq onun tətbiq mərhələsinə keçilmişdir. Onu demək kifayətdir ki,

ABŞ prezident administrasiyası 2012-ci ilin martında “Böyük verilənlərin tədqiqi və inkişafı” (Big data Research and Development İnitiative) təşəbbüsünü elan etmişdir [10]. Təşəbbüs çərçivəsində BV texnologiyalarının ABŞ dövlət siyasətinin aparıcı istiqamətlərində istifadəsi üçün kompleks tədbirlərin keçirilməsi (konfransların, forumların və s.), layihələrin işlənməsi nəzərdə tutulmuş, böyük həcmli rəqəmsal verilənlərin təşkili və analizi üçün aidiyyatı dövlət agentliklərinə (NSF - National Science Foundation, NIH - National Institutes of Health, DoD - Department of Defense, DoE - Department of Energy, DARPA - Defense Advanced Research Projects Agency və USGS- US Geological Survey) 200 milyon dollar həcmində vəsait ayrılmışdır. Bu təşəbbüs çərçivəsində 84 layihənin işlənməsi nəzərdə tutulmuşdur [10].

Təşəbbüs BV-nin yeni infrastrukturunun və tədqiqat metodologiyasının, eləcə də elmi kəşflər üçün onlardan istifadə bacarığının yüksəldilməsinə yönəlmişdir.

Ağ Ev BV-nin aşağıdakı məqsədlər üçün istifadəsini nəzərdə tutmuşdur:

- Böyük həcmli verilənlərin yığılması, saxlanması, mühafizəsi, idarə olunması, analizi və mübadiləsi üçün lazım olan texnologiyaların inkişafı;

- Elm və mühəndislik (texnologiya) sahəsində elmi kəşflərin tempinin sürətləndirilməsi, milli təhlükəsizliyin möhkəmləndirilməsi, təhsilin və təlimin kökündən dəyişdirilməsi üçün bu texnologiyalara yiyələnmək;

- BV texnologiyalarının inkişaf etdirilməsi və istifadəsi sahəsində yeni istedadların, mütəxəssislərin hazırlanmasını sürətləndirmək. Layihədə verilənlər alimi və mühəndislərinin, xüsusilə də, istənilən dildə olan mətnlərdən çıxarışların edilməsi sahəsində yüksək qabiliyyətə malik analitiklərin yetişdirilməsi nəzərdə tutulmuşdur.

Xüsusi halda, onun aşagıdakı sahələrdə tətbiqi nəzərdə tutulmuşdur: səhiyyə və əhalinin sosial müdafiəsi, ətraf mühit və dayanıqlı inkişaf, fövqəladə vəziyyətə və təbii fəlakətlərə reaksiya vermək, istehsal, robot texnikası və intellektual sistemlər, kiberfəzanın təhlükəsizliyi, nəqliyyat və energetika, təhsil və əmək resurslarının inkişafı sahələri.

BV-nin yaratdığı imkanlardan tamamilə yararlanmaq, onun potensialından səmərəli istifadə etmək məqsədilə bu təşəbbüsün həyata keçirilməsində dövlət, sənaye, elmi dairələr və qeyri-kommersiya təşkilatları birgə fəaliyyət göstərirlər.

B. Avstraliyanın Big Data strategiyası

Son illər Avstraliya hökumət idarələrində verilənlər milli dəyər kimi tanınmışdır. Yeni texnologiyaların tətbiqi ilə verilənlərin analizinin həm hökumətə, həm də avstraliya xalqına böyük töhfələr verəcəyinə inam yaranmışdır. Belə ki, BV hökumətin siyasətinin işlənməsində, və xidmətlərin təqdim olunmasında əvəzolunmaz amilə çevrilmişdir.

2013-cü ildə Avstraliya hökuməti tərəfindən “Avstraliya dövlət xidmətinin böyük verilənlər strategiyası” (Australian Public Service Big Data strategy) qəbul edilmişdir [11]. Bu strategiyada BV-nin imkanlarından istifadə etməklə dövlət sektorunun xidmət sahəsində islahatların aparılması, daha effektiv dövlət siyasətinin və vətəndaşların informasiya təhlükəsizliyinin təmini sahəsində layihələr nəzərdə tutulmuşdur. Burada əsas məqsəd, dövlət sektorunda göstərilən xidmətlərin səmərəliliyinin yüksəldilməsi, əməkdaşlıq və innovasiya sahələrində böyük verilənlərin analizindən istifadə edilməsində Avstraliyanın dünya ölkələri arasında lider mövqe tutmasından ibarətdir [11].

Bu konsepsiya xidmətlərin genişləndirilməsinə, yeni xidmət və iş ortaqlığı imkanları, təkmilləşdirilmiş siyasi inkişaf və şəxsi məlumatların məxfiliyinin qorunması, həmçinin İKT (İnformasiya kommunikasiya texnologiyaları) sektoruna investisiya qoyuluşunda dövlət dəstəyindən yararlanmaq məqsədi daşıyır.

Strategiyanın inkişafı ilkin olaraq “2012-2015-ci illər üçün Avstraliya dövlət xidmətində informasiya kommunikasiya texnologiyaları strategiyası”-da (Australian Public Service Information Communication Technology Strategy 2012-2015) təsbit edilmişdir. Strategiyanın məqsədi ümumi şəkildə aşağıdakı kimi müəyyən edilmişdir:

• xidmətlərin təkmilləşdirilməsi və keyfiyyətinin yüksəldilməsi vasitəsilə daha yaxşı xidmətlərin göstərilməsi;

• optimal investisiya qoyuluşu və innovasiyaların genişləndirilməsi vasitəsilə hökumətin fəaliyyətinin səmərəliliyinin artırılması;

• biliyin və səmərəli əməkdaşlığın yaradılmasında birbaşa iştirak.

Bundan əlavə, strategiyada daha iki layihə təklif edilmişdir: 1) informasiyanın aktiv registrlərinin inkişaf etdirilməsi (Develop information asset registers); 2) BV-nin analizində texniki inkişafın monitorinqi.

C. Böyük Britaniyanın Big Data siyasəti

2013-cü ilin yanvar ayında Britaniya hökuməti tərəfindən BV sahəsində plan qəbul edilmişdir. Bu planın yerinə yetirilməsinə 189 milyon funt sterlinq ayrılmışdır. Planda elmi-tədqiqat institutlarında BV-nin imkanlarından istifadə öz əksini tapmışdır. Planda həmçinin səhiyyə, kənd təsərrüfatı, kommersiya, elmi tədqiqatlarda və digər sahələrdə də BV-nin tədqiqi və tətbiqi nəzərdə tutulmuşdur [12].

Hazırda Böyük Britaniyada piylənmə (çəki artımı xəstəliyi), sahil eroziyası, iqlim dəyişiklikləri kimi risklərin idarə edilməsində və epidemiyalara qarşı əks tədbirlərin görülməsində BV-dən istifadə edilir [13].

D. Fransanın Big Data planı

2013-cü ilin fevral ayında Fransa hökuməti tərəfindən “Rəqəmsal yol xəritəsi” (“Digital Roadmap”) planı qəbul edilmişdir. BV də daxil olmaqla, gələcəkdə yeddi layihəninin yerinə yetirilməsini nəzərdə tutan bu plana Gələcək üçün investisiyalar proqramı (Programme of Investments for the Future) çərçivəsində 11,5 milyon avro investisiya ayrılmışdır [14]. Plan üç əsas istiqaməti özündə birləşdirir:

• Gənc nəsil üçün rəqəmsal iqtisadiyyatın imkanlarına nail olunması.

• Rəqəmsal iqtisadiyyat hesabına fransız müəssisələrinin rəqabətqabiliyyətliliyinin gücləndirilməsi.

• Rəqəmsal cəmiyyətdə və iqtisadiyyatda dəyərlərin dəstəklənməsi.

Fransa hökuməti yalnız son illər Big Data sahəsində siyasətin inkişaf etdirilməsinə diqqət ayırmışdır. Belə ki, 2013-cü ildə “New Industrial France” strategiyasında BV Fransa sənayesinin yenidən qurulmasının 34 əsas maddəsindən biri kimi yer almışdır [12,14]. “New Industrial France” strategiyası çərçivəsində 2014-cü ilin iyulunda “Big Data plan” qəbul edilmişdir. Bu plan əsasən üç fəaliyyət istiqamətini özündə birləşdirmişdir [12, 14]:

• Fransada BV ekosisteminin inkişafı.

• Böyük verilənlər üzrə sahə təşəbbüsləri. Buraya dövlət və özəl sektorlarına dair layihələr daxildir.

• Tənzimlənmənin qiymətləndirilməsi (Evaluation of regulation). Buraya gizlilik qaydaları daxildir.

E. Yaponiyanın Big Data strategiyası

BV Yaponiya hökuməti üçün də əsas iqtisadi mövzu olmuş, bu məqsədlə hökumət tərəfindən bir sıra strategiyalar qəbul olunmuşdur.

2012-ci ildə Yaponiyada şəffaflığın təmin olunmasına yardım məqsədilə “Open Government Data Strategy” qəbul edilmişdir. İslahatların Tənzimləməsi Şurası (Council for Regulatory Reform) dövlət qarşısında yerli şirkətlərin məxfilik qanunlarını pozmadan BV-dən istifadə etməsi üçün əsas prinsiplərin yaradılması təşəbbüsünü irəli sürmüşdür. Bu strategiyanın həyata keçirilməsi üçün dövlət tərəfindən 13,2 milyard yapon yeni ayrılmışdır [15].

2012 və 2013-cü illərdə Yaponiya hökuməti tərəfindən uygun olaraq “The Integrated ICT Strategy for 2020” və “Declaration to be the World's Most Advanced IT Nation” milli Big data strategiyaları qəbul edilmişdir.

2012 və 2013-cü illərdə Yaponiya hökuməti tərəfindən uygun olaraq “The Integrated ICT Strategy for 2020” və “Declaration to be the World's Most Advanced IT Nation” milli Big data strategiyaları qəbul edilmişdir. Bu strategiyalar 2013-2020-ci illər ərzində açıq dövlət verilənləri və big data ilə Yaponiyanın yeni informasiya texnologiyaları (İT) strategiyasını inkişaf etdirməyə istiqamətlənmişdir. Burada məqsəd Yaponiyanın informasiya texnologiyaları sənayesində BV-nin geniş istifadəsi üzrə yüksək standartlara malik ölkə statusu qazanmasından ibarətdir [16].

F. Çinin Big Data təşəbbüsü

Big Data ilə bağlı bir sıra tədbirlərin, o cümlədən “Big Data Technology Conference”, “Big Data & Analytics Innovation Summit”, “China Legal Big Data Symposium”, “Big Data Asia Showcase”, və “Big Data World Forum” kimi böyük konfransların, sammitlərin, forumların Çində keçirilməsi onun bu sahədə əldə etdiyi texnoloji nailiyyətlərlə izah olunur. [17]. Belə ki, 1,2 milyard mobil rabitə abunəçisinə malik olan Çin dünyanın ən böyük mobil telefon bazarına malikdir. Həmçinin, o, dünyada ən böyük, yəni 600 milyondan çox İnternet istifadəçisi vardır. Çində sosial medialar daha fəal mövqe tutur. Dövlət hesablamalarına görə bloqlar, sosial şəbəkələr, mikrobloqlar və s. daxil olmaqla social mediadan 250 milyondan çox insan istifadə edir. Qeyri-dövlət təşkilatlarının qiymətləndirməsinə görə isə bu rəqəm 590 milyona yaxındır [17].

Çində real-vaxt rejiminində informasiyanın və mobil telefon, sosial media və İnternet istifadəçilərinin sayının sürətlə artması ölkədə yaranmış problemlərin aşkarlanmasında və ölkənin inkişafı üçün BV-nin istifadəsinə əlverişli şərait yaratmışdır. Artıq dövlət sektorunda BV-nin imkanlarının tədqiqinə başlanmışdır. 2014-cü ilin iyun ayında ölkənin siyasi konsultativ (məşvərət) qanunvericilik orqanı olan Çin Xalq Siyasi Məşvərət Şurası (Chinese People’s Political Consultative Conference – CPPCC) tərəfindən keçirilən forumda idarəetmə imkanlarının yüksəldilməsində Big Data texnologiyalarının potensial istifadəsi təşəbbüsü irəli sürülmüşdür [17]. Bu təşəbbüsdə həmçinin aşağıdakı sahələrdə problemlərin həll edilməsi məqsədilə BV-nin istifadəsiylə bağlı konkret məsələlər nəzərdə tutulmuşdur:

- Dövlət sektorunun məhsuldarlığının yüksəldilməsi;
- Şəhər nəqliyyatının planlaşdırılmasının yaxşılaşdırılması;
- Sosial-iqtisadi tendensiyaların dərk edilməsi;
- Yoxsulluğun vəziyyətinin qiymətləndirilməsi;
- Davamlı e-tullantıların praktiki utilizasiyasına yardım edilməsi;
- Şəhərlərdə çirklənmənin qaynar nöqtələrinin müəyyən edilməsi və s.

BV-nin potensial imkanlarından istifadənin həyata keçirilməsi üçün ilk növbədə bir çox problemlərin (verilənlərin analitikası, operativ/system problemləri) həlli üçün dövlət və özəl sektorların birgə işinin aparılması üçün əlverişli şəraitin yaradılması nəzərdə tutulmuşdur [17].

Qeyd etmək lazımdır ki, 2013-ci ilin noyabr ayında BV-nin tətbiqindən əldə olunan səmərənin maksimallaşdırılması üçün Çinin Milli Statistika Bürosu (China’s National Bureau of Statistics) BV-nin istifadəsi üzrə uzunmüddətli əməkdaşlıq münasibətlərinin qurulması məqsədilə 11 iri Çin müəssisəsi, o cümlədən Baidu, Alibaba, China Unicom və s. kimi iri müəssisələr ilə verilənlərin mübadiləsi sahəsində bir sıra müqavilələr imzalamışdır [17].

G. Koreyanın Big Data layihələri

Koreya ən müasir texnologiyaların istifadəsi sahəsində böyük təcrübəyə malik dövlətdir. Ölkənin iqtisadi inkişafı, innovasiya və rəqabət üstünlüyü məhz bu texnologiyaların tətbiqi sayəsində əldə olunmuşdur. BV-nin istifadəsi sahəsində də bu tendensiya saxlanılmışdır. Digər texnologiyalar kimi artıq “big data” da Koreyanın inkişafında mühüm rol oynamağa başlamışdır. BV-nin istifadəsi elmi dairələrin, dövlət müəssisələrinin maraq dairəsindədir. Ölkənin dövlət siyasətinin hazırlanmasında, dövlətin xidmət sektorunda, kiçik və orta müəssisələrin rəqabətqabiliyyətliliyinin saxlanılmasında, elmi

verilənlər sahəsində vərdişlərin inkişafında və yeni infrastrukturun yaradılmasında BV-nin istifadəsi nəzərdə tutulmuşdur. Bu yanaşmalar dövlət və özəl sektorlarının, Milli İnformasiya Cəmiyyəti Agentliyinə (National Information Society Agency) daxil olan Böyük Verilənlər Strateji Mərkəzi (Big Data Strategy Center), Seul Milli Universitetinin (Seoul National University) nəzdində fəaliyyət göstərən Big Data İnstitutunun (Big Data Institute) əməkdaşlığı ilə yaradılmışdır [12].

2012-ci ilin noyabrında Koreyada fəaliyyət göstərən Milli Elmi-Texnoloji komissiya (National Science & Technology Commission) tərəfindən “Big Data Master Plan” işlənib hazırlanmışdır. Bu planın hazırlanmasında hökumətin məqsədi “Smart Nation” ideyasını reallaşdırmaqdan ibarətdir. Bu planın tərkib hissəsi kimi dövlət tərəfindən “Pilot Project for Using Big Data in Official Statistics” layihəsi təsis etmişdir [12].

Koreyada müxtəlif mənbələrdən əldə olunmuş informasiyalar əsasında təbii fəlakətlərin proqnozlaşdırılması məqsədilə BV-nin informasiya sistemi inkişaf etdirilir [18]. Milli hökumət əhalinin siyahiya alınması işlərində BV-dən istifadə etməyi planlaşdırır. Bu isə ənənəvi siyahıya alınma ilə müqayisədə təxminən 140 milyon dollara qənaət etməyə imkan verəcəkdir [18].

NƏTİCƏ

Mövzu üzrə aparılmış araşdırmalar onu deməyə əsas verir ki, dövlətlər big data strategiyalarını qəbul edərkən müxtəlif aspektlər (elm, sosial, iqtisadi, səhiyyə, təhlükəsizlik və s.) nəzərə alınmışdır. Ona görə də qəbul olunmuş bu sənədlər bir birindən fərqlənirlər.

Böyük həcmli verilənlərin dərin analizi dövlət səviyyəsində ən yaxşı qərarların qəbulunda həlledici rol oynaya bilər. Dövlətlərin yoluxucu xəstəliklər, terrorizm, fəlakət və qlobal risklərin qarşısının alınması kimi səhiyyə və sosial təhlükəsizlik məsələlərinin həllində böyük verilənlərin imkanlarından yararlanması mümkündür [1, 2].

İKT sektorunun yüksək inkişaf səviyyəsinə, İnternet və mobil istifadəçilərin sayına, rəqabətqabiliyyətliliyinə görə MDB məkanında ilkin səviyyəni tutan və iqtisadi baxımdan inkişaf etməkdə olan ölkələr statusuna malik olan ölkəmiz də müxtəlif sahələrdə BV-nin imkanlarından istifadəsinə start verilmişdir. AMEA-nın İnformasiya Texnologiyaları İnstitutunda BV üçün əsas infrastruktur elementlərindən hesab olunan bulud hesablamaları mərkəzinin olması bunu deməyə əsas verir.

ƏDƏBİYYAT

[1] Alguliyev R. M., Hajirahimova M. S. "Big Data" phenomenon: Challenges and Opportunities // Problems of Information Technology, 2014, no.2, pp. 3-16.

[2] Jina X., Benjamin W. Waha, Chenga X., Wanga Y. Significance and Challenges of Big Data Research// Big Data Research, Volume 2, Issue 2, June 2015, pp. 59–64.

[3] Big Data, Big Impact: New Possibilities for International Development, available at: http://www.weforum.org/reports/ big-data-big-impact-new-possibilities-international development, 2012.

[4] Big data for development: challenges & opportunities, available at: http://www.unglobalpulse.org/projects/, May 2012.

[5] Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by 'big data': http://www.oecd-ilibrary.org/science-and-technology/exploring-data-driven-innovation-as-a-new-source-of-growth_5k47zw3fcp43-en, 2013.

[6] Horizon 2020 - The EU Framework Programme for Research and Innovation 2014-2020: ec.europa.eu/programmes/horizon2020/, 2014.

[7] “Big Data in Action for Developmenthttp, available at: http://data.worldbank.org/news/big-data-in-action-for-development, 2014.

[8] Big data - big today, normal tomorrow: http://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000220001PDFE.pdf, november 2013.

[9] KNOW (Interconnecting knowledge for the early identification of issues, events and developments) project: http://cordis.europa.eu/result/rcn/54661_en.html, 2011

[10] Big data is a big deal, available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal, 2012.

[11] Australian Pubic Service Big Data Strategy: http://www.finance.gov.au/sites/default/files/Big-Data-Strategy_0.pdf, 2013.

[12] Hans Lammerant, Antonella Galetta, Paul De Hert (VUB), LorenzoBigagli, Paolo Mazzetti (CNR), Stéphane Grumbach (INRIA) Big Data Policies BYTE project - www.byte-project.eu, 2014 avqust

[13] UK data servis The Horizon Scanning Programme Team [Internet] London: https://www.gov.uk/government/groups/horizon-scanning-centre, August, 2014.

[14] The new face of industry in France: http://www.entreprises.gouv.fr/files/files/directions_services/secteurs-professionnels/industrie/nfi/NFI-anglais.pdf

[15] Open Government Data Strategy, available at: http://japan.kantei.go.jp/policy/it/20120704/text.pdf, 2012.

[16] Declaration to be the world’s most advanced IT nationt: http://japan.kantei.go.jp/policy/it/2013/0614_declaration.pdf, June 2013.

[17] Jackie Hoi-Wai CHENG Big Data for Development in China : UNDP China Working Paper, available at:: http://www.cn.undp.org/content/dam/china/docs/Publications/UNDP Working%20Paper_Big%20Data%20for%20Development%20in%20China_Nov%202014.pdf , 2014.

[18] Adarsh Desai , Bruno Sánchez, Rade Nuño Korea shows how to use Big Data for development - http://blogs.worldbank.org/voices/korea-shows-how-use-big-data-development, iyun, 2015.

Təbii Ehtiyatlardan Səmərəli İstifadə Edilməsi Proseslərinin Böyük Verilənlər Əsasında Təhlili Məsələləri

Ələvsət Əliyev[1], Leyla Əkbərova[2], Məleykə İbişova[3]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[3]Azərbaycan Texniki Universiteti, Bakı, Azərbaycan
[1]*alovsat_qaraca@mail.ru*, [2]*akberovaleyla@rambler.ru*, [3]*ibishova96@mail.ru*

***Xülasə* – Məqalə iqtisadiyyatın “yaşıllaşdırılması” kontekstində təbii ehtiyatlardan səmərəli istifadə olunması proseslərinin təkmilləşdirilməsi probleminə həsr olunmuşdur. “Yaşıl iqtisadiyyatın” mahiyyəti açıqlanmış və onun əsaslandığı prinsiplər aydınlaşdırılmışdır. “Yaşıllaşdırma” prosesi cəmiyyətin və iqtisadi inkişafın yeni mərhələsi kimi təqdim olunmuşdur. “Yaşıl iqtisadiyyatın” formalaşmasında ilk növbədə təbii ehtiyatlardan səmərəli istifadə məsələlərinə diqqət yetirilmişdir. Müasir dövrdə formalaşmaqda olan informasiya iqtisadiyyatının “yaşıllaşdırılması” istiqamətlərində yeni İKT texnologiyalarının tətbiqi təklif olunmuşdur. Təbii ehtiyatlardan səmərəli istifadə edilməsi proseslərinin Big Data texnologiyaları vasitəsilə təhlil olunması əsaslandırılmış və onların xüsusiyyətləri araşdırılmışdır**

Açar sözlər* – *yaşıl iqtisadiyyat, “yaşıllaşdırma” texnologiyaları, təbii ehtiyatlar, Big Data texnologiyaları, iqtisadi təhlil, qiymətləndirmə

I. GİRİŞ

Azərbaycanda iqtisadiyyatın modernləşdirilməsi, innovativ inkişafı, rəqabət qabiliyyətinin artırılması, enerji səmərəliliyinin təmini ilk növbədə iqtisadi artıma istiqamətləndirilsə də, əslində əsas məqsəd hər bir insanın həyat şəraitinin yaxşılaşdırılması və gələcək nəsillərin normal rifah şəraitinin təmin edilməsidir. Bu məsələ hazırda formalaşmaqda olan İnformasiya Cəmiyyəti iqtisadiyyatının prinsipləri əsasında davamlı və dayanıqlı inkişafın təmini kimi müəyyənləşdirilmişdir. Bu isə dünya və ölkə miqyasında istehsalın və istehlakın ekolojiləşdirilməsini zəruri etməklə təbii ehtiyatlardan, yaşıl resurslardan daha da səmərəli istifadəni tələb edir. 2012-ci ildə qəbul edilən “Azərbaycan 2020: Gələcəyə baxış” İnkişaf Konsepsiyasının [1], həmçinin Azərbaycan Respublikasında İnformasiya Cəmiyyətinin inkişafına dair 2014-2020-ci illər üçün Milli Strategiyanın [2] əsas hədəflərindən biri ekoloji cəhətdən dayanıqlı sosial-iqtisadi inkişafa nail olmaqdır. Konsepsiyada göstərilib ki, biomüxtəlifliyin qorunması, yanacaq-enerji kompleksinin ətraf mühitə mənfi təsirinin neytrallaşdırılması, dənizin və onun akvatoriyasının çirklənməsinin aradan qaldırılması və qorunması, yaşıl ərazilərin bərpası və mövcud resursların səmərəli mühafizəsi istiqamətində zəruri tədbirlər həyata keçirilməlidir. Konsepsiyanın əhatə etdiyi dövrdə Azərbaycanda orta hesabla bir vahid Ümumi Daxili Məhsulun (ÜDM) istehsalı üçün istifadə edilən enerjinin və emissiya olunan zərərli qazların miqdarı İqtisadi Əməkdaşlıq və İnkişaf Təşkilatı (İƏİT) ölkələri üzrə müvafiq göstəriciyə yaxınlaşmalıdır. Azərbaycanda iqtisadiyyatın yaşıllaşdırılması, alternativ (bərpa olunan) enerji mənbələrindən istifadə səviyyəsinin yüksəldilməsinin stimullaşdırılması, bu sahədə həyata keçiriləcək təhlil, qərarqəbuletmə və proqnozlaşdırma işlərində yeni İKT texnologiyalarından, xüsusilə Böyük Verilənlərdən istifadə olunması ciddi məsələlərdəndir.

II. YAŞIL İQTİSADİYYATIN MAHİYYƏTİ VƏ XÜSUSİYYƏTLƏRİ

Yaşıl iqtisadiyyat (green economy) iqtisad elmində son onilliklərdə meydana gəlmiş yeni istiqamətdir. Bu təbii mühitin bir hissəsi olan, həmçinin ondan asılı olan bir iqtisadiyyatdır. Yaşıl iqtisadiyyatın konsepsiyası iqtisad elmində mövcud olan resursyönümlü iqtisadiyyat, ekoloji iqtisadiyyat, ətraf mühitin iqtisadiyyatı, yaşıl siyasət və s. kimi bir çox istiqamətlərin ideyalarını özündə birləşdirir.

Yaşıl iqtisadiyyat paradiqmasının əksi olan qəhvəyi iqtisadiyyat (brown economy) termini, resursların daha çox istifadəsi, aşağı enerji səmərəliliyi, dayanıqlı olmayan enerji qaynaqlarından asılılıq və yüksək dərəcədə iqlim riskinə əsaslanır. Belə ki, havanın, suyun çirkləndiyi, yaşıllıqların azaldığı indiki vaxtda get-gedə mühüm əhəmiyyət qazanan yaşıl iqtisadiyyatın mahiyyəti müntəzəm olaraq “davamlılıq” şüuruyla, həm ekosistemin, həm də gələcək nəsillərin qeydinə qalmaq, təbiətdəki bütün canlıların əsas həyat haqqlarının qorunmasına nail olmaqdır [3].

İqtisadi Əməkdaşlıq və İnkişaf Təşkilatı (Organisation for Economic Co-operation and Development), Birləşmiş Millətlər Təşkilatının Ətraf Mühit Proqramı (United Nations Environment Programme) kimi beynəlxalq təşkilatlar “yaşıl inkişaf” və ya “yaşıl iqtisadiyyat”ı ekoloji təhlükəsizliyi təmin edən mal və xidmətlərin istehsal və istehlakını, həmçinin yaşıl inkişafa yatırılan investisiyaları nəzərdə tutan bir anlayış kimi qəbul edirlər. Bu baxışda ekoloji davamlılıq təmin edilərkən iqtisadi inkişaf, gəlir artımı, məşğulluq və yoxsulluğun azaldılması amilləri ətraflı şəkildə düşünülür. Yaşıl

iqtisadiyyat, əslində yeni iqtisadiyyat anlayışı kimi də ifadə edilə bilər və ya klassik iqtisadiyyatın təməllərinə insan və təbiət faktorlarının əsaslı bir şəkildə əlavə olunması kimi də təyin oluna bilər. İnsanın və təbiətin dəyərinin dərk edilməsi, insanların rifahını yüksəlməklə yoxsulluğun azaldılması, məhdud təbii ehtiyatlardan effektiv və optimal şəkildə istifadə edilməsi yaşıl iqtisadiyyatın təməl prinsiplərini meydana gətirir [4].

Hazırda yaşıl iqtisadiyyat anlayışının nəyi əhatə etdiyini dəqiq olaraq müəyyənləşmiş və hamı tərəfindən qəbul edilən tərifi yoxdur. Bu səbəblə, ayrı-ayrı ölkələr özlərinin xüsusi şərtləri çərçivəsində fərqli təriflər ortaya qoyurlar. Əslində bu günkü dövrdə yaşıl iqtisadiyyatın anlaşılması və ona müvafiq addımların atılmasının dövlətdən-dövlətə fərqlənməsi normaldır. Çünki iqtisadi cəhətdən inkişaf etmiş ölkələrin (İEÖ) insan kapitalı ilə inkişaf etməkdə olan (İEOÖ) və ya yeni sənayeləşmiş ölkələrin (YSÖ) insan kapitalı arasında keyfiyyət baxımından böyük fərq vardır və bu fərq müvafiq ölkələrin yaşıl iqtisadiyyatla bağlı atdıqları addımlarda da özünü göstərir. Qeyd olunanları ümumiləşdirəndə aydın olur ki, davamlı inkişaf iqtisadi, ekoloji və ictimai inkişaf sütunlarının təməli olan yaşıl iqtisadiyyata əsaslanır [5, 6]. Müşahidələr göstərir ki, yaşıl iqtisadiyyatın səmərəli formalaşma istiqamətlərinin müəyyənləşdirilməsi kifayət qədər böyük həcmli informasiya resurslarının emalı əsasında həyata keçirilə bilər. Ona görə də iqtisadiyyatın yaşıllaşdırılmasının istiqamətlərindən biri də big data texnologiyalarının tətbiqi hesab olunur.

III. YAŞIL İQTİSADİYYATDA TƏBİİ EHTİYYATLARDAN SƏMƏRƏLİ İSTİFADƏNİN TƏHLİLİNDƏ BİG DATA TEXNOLOGİYALARININ TƏTBİQİ

Yaşıl iqtisadiyyat sahəsində fəaliyyətin, bilik və bacarığın əhəmiyyəti, təbii ehtiyatlardan səmərəli istifadə edilməsinin təhlilindən aydın ola bilər. Hər hansı bir ölkənin iqtisadiyyatının vəziyyəti istehsal olunan məhsul və xidmətlərin dəyəri ilə, milli məhsulun əldə edilməsi üçün sərf edilən enerji ilə qiymətləndirilə bilər. Beynəlxalq statistikada istifadə edilən bu göstərici istehsalın nə dərəcədə qənaətli olmasını, müasir texnoloji nailiyyətlər əsasında qurulması və bu səviyyədə çalışan işçi quvvəsinin peşəkarlığını göstərir. Belə müqayisələr zamanı adətən vahid enerjinin istifadəsi nəticəsində istehsal olunan məhsulun dəyəri qiymətləndirilir [7]. Müqayisələr təsdiq edir ki, bu göstərici müxtəlif ölkələr arasında çox fərqlənir. Məsələn, Yaponiyada 1 vahid enerji hesabına istehsal olunan məhsulun orta qiyməti 7 ABŞ dolları, ABŞ-da 4.6, Kanadada – 4.3-ə bərabərdir. Norveçin neft və qaz ixracatçısı olmasına baxmayaraq 1 vahid həcmdə enerji üçün ekvivalent olaraq istehsal olunan məhsulun dəyəri 5.6 ABŞ dolları təşkil edir. Bunlar iqtisadiyyatın qeyri-neft sektorunun inkişafı nəticəsində əldə edilmişdir. Qeyd etmək lazımdır ki, Azərbaycanda da qeyri-neft sektorunun inkişafı dövlət siyasətinin əsas prioritetlərindəndir. Son statistik göstəricilərə görə, ölkədə qeyri-neft sektorunun artımı neft sektorunda müşahidə olunan artımdan yüksəkdir. Beynəlxalq təşkilatların statistik məlumatlarına görə, ölkədə vahid enerji hesabına istehsal olunan məhsulun qiyməti orta hesabla 2.5 ABŞ dolları təşkil edir. Bu, bir sıra digər neft ixrac edən ölkələrin göstəricilərindən yüksəkdir. Məsələn, Küveytdə bu rəqəm 1.9, Ərəb Əmirliklərində 2.2, Rusiyada 2.0 ABŞ dolları təşkil edir. Eyni zamanda dünyada bərpa olunan enerjidən də istifadə hər il artmaqdadır. Ölkədə alternativ enerji istehsalının inkişafı üçün geniş imkanlar yaradılmışdır. Günəş, külək, geotermal və biokütlənin enerjisindən istifadə etmək üçün böyük potensial vardır. Göründüyü kimi təbii ehtiyatlardan səmərəli istifadə məsələsi birinci növbədə düzgün iqtisadi təhlil və qiymətləndirmələr əsasında həyata keçirilə bilər. Bu isə qeyd olunduğu kimi Big Data texnologiyalarının tətbiqini qaçılmaz edir.

IV. TƏBİİ KAPİTALIN ƏSAS FUNKSİYALARI ÜZRƏ QİYMƏTLƏNDİRMƏLƏRDƏ BÖYÜK HƏCMLİ VERİLƏNLƏR

Son onilliklərdə iqtisadi artım daha çox təbii kapitalların - torpaq, su, meşə, enerji resursları, müxtəlif filizlər və s. tükənməsinə əsaslanmışdır. Təbii kapitalın məhdudlaşdırılmış ehtiyatları müasir iqtisadi tendensiyaların saxlanılması ilə planetin əhalisinin rifahının davamlı inkişafını təmin edə bilmir. BMT-nin ekspertlərinin qeyd etdiyi kimi ekoloji defisit və sosial ədalətsizlik kimi xüsusiyyətlərə malik iqtisadi model “yaşıl iqtisadiyyat”dan olduqca uzaqdır. “Yaşıl iqtisadiyyat”a keçid üçün prinsipcə ən əhəmiyyətlisi iqtisadi aktiv və ictimai fayda mənbəyi kimi təbii kapitalın qorunması və artırılması, onun istifadəsinin səmərəliliyinin yüksəldilməsi, funksiyalarının dəstəklənməsi və gələcək nəsillər üçün saxlanılması xüsusi əhəmiyyət daşıyır.

Yaşıl iqtisadiyyatın əsas xüsusiyyəti “təbii kapitalın” iqtisadiyyatın bir hissəsi kimi bazar prosesinə daxil edilməsidir [8]. Yaşıl iqtisadiyyat təbii kapitala münasibəti transformasiya edir, təbiət və cəmiyyətin münasibətləri ilə bağlı mühüm məsələlərə yeni aspektdən baxır. Yaşıl iqtisadiyyat təbiəti kapitalın və istehsal qüvvələrinin strukturuna cəlb etmək yolu ilə iqtisadi inkişafı təmin edir.

Son zamanlar dünyada təbii kapital daha çox təbii resurslar kimi izah olunur. Uğurlu iqtisadi artım üçün onun bütün funksiyalarını da nəzərə almaq lazımdır.

Təbii kapital davamlı inkişafın əsası kimi bütün ölkələrin iqtisadiyyatında mühüm rol oynayır. Dünya Bankının qiymətləndirməsinə görə Rusiyanın milli sərvətlərinin strukturunda təbii kapitalın payı 70%-ə yaxın, insan kapitalı 20%, fiziki kapital isə 10% təşkil edir. İƏİT-nin inkişaf etmiş ölkələrində isə təbii kapital təxminən 5%, insan kapitalı 85%, fiziki kapital isə 10% təşkil edir. Bu kimi qiymətləndirmələr ölkə və regionlar üzrə həddən artıq böyük həcmli verilənlərin operativ sistemləşdirilməsini və təhlilini tələb edir.

V. EKOSİSTEM XİDMƏTLƏRİNİN İQTİSADİ QİYMƏTLƏNDİRİLMƏSİNDƏ BİG DATA TEXNOLOGİYALARINDAN İSTİFADƏ

Ekosistem xidməti – insanların ekosistemlərdən aldıqları faydalardır. Ekosistem xidmətlərinin və onun qiymətləndirilməsi zərurəti aşağıdakı bir sıra şərtlərlə bağlıdır: 1)Ekosistemlərin qlobal deqradasiyasının artan templəri, 2)Gələcək deqradasiyasının qarşısının alınması üçün ekosistem xidmətlərin iqtisadi kontekstə daxil edilməsi zəruriliyi, 3)Ekosistem xidmətlərinin iqtisadi dəyəri və onlar üçün ödəniş mexanizmlərinin tətbiqi və s. Ayrı-ayrı ekosistemlər və biosfera bütovlükdə malların və xidmətlərin böyük müxtəlifliyini təmin edir. Bəşəriyyətin mövcudluğu və rifahı ekosistem xidmətlərindən asılıdır [9].

Hal-hazırda dünyada ekosistem xidmətləri, onların qiymətləndirilməsi, potensial satıcı və alıcılarının, kompensasiya mexanizmlərinin təyini, bu xidmətlər bazarının formalaşması ilə bağlı məsələlərin geniş dairədə təhlilinə və fəal işlənilməsinə başlanılmışdır. Nəzəri işləmələrlə yanaşı, dünya təcrübəsində və ayrı-ayrı ölkələrdə iqtisadi qiymətləndirilmə və ekosistem xidmətlərin kompensasiyaları nümunələri meydana gəlmişdir. Ayrı-ayrı ölkələr daxilində ekosistem xidmətləri üçün ödəniş mexanizmlərini reallaşdırmaq cəhdləri genişlənməkdədir [10, 11].

Ekosistemlərin və onlar tərəfindən verilən xidmətlərin qorunması "yaşıl iqtisadiyyat"ın prioritetli istiqamətidir. Beynəlxalq iqtisadiyyatda bu xidmətlər tez-tez dünya üçün yeni olan "ekosistem xidmətləri üçün ödənişlər", "ekoloji donor", "kompensasiya mexanizmi", "təbiətə mübadiləyə görə borclar" kimi terminlərlə əlaqələndirilir. Bir sıra fundamental beynəlxalq tədqiqatlar da ekoxidmətlər iqtisadiyyatına həsr edilmişdir. Ekosistem xidmətlərinin iqtisadi praktikaya, o cümlədən maliyyə sektoruna daxil edilməsi üzrə tətbiqi işləmələr də meydana çıxmışdır. Belə işləmələr, təhlillər və müvafiq idarəedici qərarların qəbulu ilkin toplanmış iri həcmli verilənlərin müasir İKT üsulları ilə analitik təhlilinə əsaslanır.

VI. ENERJİ VƏ SU RESURSLARINDAN SƏMƏRƏLİ İSTİFADƏ GÖSTƏRİCİLƏRİNİN TƏHLİLİNDƏ BİG DATA TEXNOLOGİYALARININ TƏTBİQİ

Təhlillər göstərir ki, sənayenin inkişafı enerji və su resurslarından səmərəli istifadədən olduqca asılıdır. Su və enerji resurslarından daha səmərəli istifadə, onların effektivliyinin artırılması iqtisadi əhəmiyyət daşıdığı kimi, ekoloji baxımdan da çox vacibdir [12]. Enerji və su resurslarından istifadə səmərəliliyinin bir neçə əsas göstəricisi məlumdur. Onlardan birincisi və beynəlxalq təşkilatlar tərəfindən daha çox istifadə ediləni hər hansı bir ölkədə və yaxud iqtisadi fəaliyyət sahəsində bir vahid enerjinin istifadəsi nəticəsində sənayedə, kənd təsərrüfatında və yaxud xidmət sahələrində yaranmış (ÜDM) dəyəridir. Azərbaycanda son on ildə ÜDM-un hər bir vahidinin istehsalı üçün enerji sərfi sürətlə azalıb. Göstərilən fakt davamlı inkişaf və yaşıl iqtisadiyyatın ən önəmli göstəricilərindən sayılan bərpa olunan və bərpa olunmayan ehtiyatlardan səmərəli istifadənin bariz təsdiqidir. Hesablamalara görə, ölkədə hər bir kq neftin istifadəsi nəticəsində 7 ABŞ dollarından çox məhsul istehsal edilib və yaxud xidmət göstərilib. Dünya Bankı tərəfindən həyata keçirilmiş təhlillər və onların əsasında dərc edilmiş rəsmi hesabatlar göstərir ki, Azərbaycanda bu rəqəm əksər neft və qaz ixrac edən ölkələrin göstəricilərindən daha yüksəkdir [4].

Son on ildə Azərbaycanda ÜDM-n hər bir vahidinin istehsalı üçün sərf edilən enerjinin səmərəliliyi 300%-dən çox artıb. Hazırda ölkədə enerji istifadəsinin səmərəliliyi dünya üçün hesablanmış göstəricilərdən daha yüksəkdir. Dünya Bankının 2013-cü ildə təqdim etdiyi beynəlxalq statistikaya görə, 2003–2011-ci illər arasında dünyada orta hesabla hər bir kq neftdə olan enerji ekvivalentinin istifadəsi nəticəsində istehsal olunmuş ÜDM-nin dəyəri 4,93 ABŞ dollarından 6,19 ABŞ dollarına qədər yüksəlib. Azərbaycanda isə eyni mənbələrə görə, həmin dövrdə bu rəqəmlər müvafiq olaraq 2,26 (2003 il) və 7,22 (2011 il) ABŞ dolları təşkil edib. Göstərilən dinamika Azərbaycanda enerji mənbələrindən səmərəli istifadə sahəsində ardıcıl siyasətin həyata keçirilməsindən xəbər verir. Azərbaycan ərazisinin 60 faizinin arid ərazi olması proqnozlaşdırmağa əsas verir ki, onsuz da su ehtiyatlarından elə də zəngin olmayan respublikamızda yağıntının miqdarı azalmağa meyllidir. Belə ki, iqlimin istiləşməsi prosesində quraqlıq olan ərazilərdə yağıntının miqdarı bir qədər də azalır, rütubətli ərazilər üçün isə əksinə, yağıntının miqdarı artır. Hesablamalara görə Azərbaycanda gələcəkdə temperaturun yüksəlməsi və yağıntının azalması proqnozlaşdırırlır [4].

Göründüyü kimi enerji və su resurslarından istifadənin səmərəliyinin daha da artırılması milli və qlobal prioritet təşkil edir. Onların səmərəli istifadəsinə aid iki ən əsas istiqamət göstərilir: mövcud olan enerji mənbələrindən daha səmərəli istifadə və yeni alternativ enerji mənbələrinin səfərbər edilməsi yolları. Hər iki istiqamətdə həyata keçirilən tədbirlər və işlənilməkdə olan layihələr ölkə həyatının bütün sahələrini və təsərrüfat subyektlərini əhatə etdiyindən bu proseslərdə yeni İKT, o cümlədən Big Data texnologiyalarının tətbiqi mühüm yeni nəticələrin əldə olmasına şərait yaradır.

NƏTİCƏ

Aparılmış tədqiqatların nəticələrinə görə, axırıncı yüzillikdə ölkəmizin ərazisində temperaturun dəyişməsi müxtəlif regionlarda 0,2-dən 1,5 °C-yə qədər təşkil edir. Son 50 il isə ən isti dövr kimi qiymətləndirilir. Belə bir ekoloji şəraitdə kənd təsərrüfatında istiliyə davamlı bitkiçiliyin inkişafı, enerji, su və digər təbii resurslardan daha səmərəli istifadəsi, meliorasiya və irriqasiya infrastrukturunun inkişafı, Xəzərin təbii ehtiyatlarından təsərrüfat məqsədləri ilə istifadə və s. istiqamətlər dövlətin və əhalinin əsas diqqət mərkəzində olmalıdır. Göstərilən təhlükələrin vaxtında qarşısının alınması və fəsadların aradan qaldırılması üçün zamanında müvafiq tədbirlər həyata keçirilməlidir. Belə tədbirlərin görülməsi isə ilk növbədə düzgün informasiyalara əsaslanan iqtisadi təhlillərin, qiymətləndirmələrin, proqnozların aparılmasını

tələb edir. Ona görə də fikrimizcə göstərilən sahələrdə Big Data texnologiyalarının tətbiqinə müvafiq şərait və mexanizmlər yaradılmalıdır.

ƏDƏBİYYAT

[1] “Azərbaycan - 2020: Gələcəyə Baxiş” İnkişaf Konsepsiyası. Bakı, 29 dekabr 2012-ci il. http://www.president.az

[2] “Azərbaycan Respublikasında informasiya cəmiyyətinin inkişafına dair 2014-2020-ci illər üçün Milli Strategiya”nın təsdiq edilməsi haqqında Azərbaycan Respublikası Prezidentinin Sərəncamı, 02.04.2014-cü il, http://www.president.az

[3] Dinda S.. A theoretical basis for green growth. International Journal of Green Economics, 2014, vol. 8, issue 2, pages 177-189.

[4] Qasımlı V., Vəliyev Z., Hümbətov M. və s. Yaşil inkişaf: enerji səmərəliliyi və alternativ mənbələr, Bakı, 2014, http://sia.az/az/news/economy

[5] Bruyère S., Filiberto D. The green economy and job creation: inclusion of people with disabilities in the USA. International Journal of Green Economics 2013 - Vol. 7, №3, pp. 257 - 275.

[6] Adrian C. Newton. The green economy and the knowledge economy: exploring the interface. International Journal of Green Economics. 2011 - Vol. 5, No.3 pp. 231 - 247.

[7] By: Xu Ming, Cai Hua, Liang Sai. Big Data and Industrial Ecology. Journal of Industrial Ecology. 2015, Vol. 19 Issue 2, p. 205-210.

[8] Tomic, S.D.K.; Fensel, A. A platform for data economy for energy efficiency data. Big Data. IEEE International Conference. 2013. pp. 43-47.

[9] Xiwei Wang , Yu Guo, Mengqing Yang, Yong Chen, Wenxiao Zhang. Information ecology research: past, present, and future. J. Information Technology and Management. 2015. pp.1-13. http://link.springer.com

[10] David Schimel , Michael Keller. Big questions, big science: meeting the challenges of global ecology. Oecologia. 2015, Vol. 177, Issue 4, pp. 925-934. http://link.springer.com

[11] Stefana Broadbent , Claire Lobet-Maris. Big Data – ecosystem – Business ecosystem – Big Data components – Industry. Towards a Grey Ecology. 2014. pp 111-124. http://link.springer.com

[12] Hampton, S.E., et al.: Big data and the future of ecology. Front. Ecol. Environ. 11(3), 156–162 (2013)

Verilmiş Predmet Sahəsində Terminoloji İnformasiya Sisteminin İşlənməsi

Əfruz Qurbanova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
afruz@iit.ab.az, afruz1961@gmail.com

***Xülasə* — Məqalədə terminoloji fəaliyyət sahəsinin geniş mütəxəssis dairəsi üçün müəyyən predmet sahəsində qərarların qəbulunu dəstəkləyən informasiya sisteminin yaradılması istiqamətində sistemin arxitekturu və konseptul sxemi işlənib. Terminoloji informasiya sisteminin reallaşdırılmasının terminlərin təhlilinə yönəlmiş xüsusi biliklər bazasının qurulmasına yaratdığı imkanlar müəyyənləşdirilib.**

Açar sözlər — terminoloji lüğət, informasiya sistemi, ontologiya, semantik şəbəkə

I. GİRİŞ

Müasir dövrdə elm və təhsil sistemində yüksək ixtisaslı kadrların hazırlanması səviyyəsinin yüksəldilməsi üsullarından biri müasir informasiya texnologiyalarının, müxtəlif predmet sahələrində bilik və verilənlərin intellektual emalı sistemlərinin yaradılması, tətbiqi və istifadəsidir. Son illər elm və texnikanın müxtəlif sahələrində baş verən əhəmiyyətli dəyişikliklər və inkişaf ayrı-ayrı predmet sahələrinin terminoloji bazalarının zənginləşməsinə və dilimizdə yeni terminlərin yaranmasına səbəb olmuşdur. Dünyada və Azərbaycanda terminologiya fəaliyyəti ilə məşğul olan elm və tədris mərkəzlərinin çoxalması, yayılması, elmin və texnikanın müxtəlif sahələrinin inkişaf dinamikası, terminlərlə bağlı beynəlxalq standartların daim artan tələblərinə uyğunluq və s. kimi amillər müxtəlif predmet sahələrinin terminlərinin strukturlaşdırılması, vahid bir mərkəzə toplanması, işlənməsi, istifadəsini təmin edən terminoloji informasiya sisteminin yaradılması və müşayiət edilməsini şərtləndirdi.

II. MƏSƏLƏNİN QOYULUŞU

Ayrı-ayrı predmet sahələrinin ontologiyasının yaradılması həmin sahələr üzrə mütəxəssislərin həm tədris prosesində, həm də praktiki işdə öyrənməsi və istifadəsini təmin etməklə yanaşı, biliyinin qeyri-müəyyənliyinin aşağı salınması və ya bilik səviyyəsinin yüksəldilməsinə xidmət edəcəkdir.

Verilmiş predmet sahəsinin terminoloji bazası müəyyən məqsədlərə yönəldilmiş sistem şəklində təqdim edilə bilər [1].

$$S=<M,R,P> \quad (1)$$

M – sistemin elementləri çoxluğudur: predmet sahəsinə aid terminlər və onların tərifləri, R – predmet sahəsinin terminləri arasında əlaqə çoxluğu, P – sistemin verilmiş məqsədə çatmaq imkanı verən xassələr çoxluğunu göstərir.

Terminoloji sistemin elementləri kimi terminlər aşağıdakı xaccələrə malik olmalıdır:

- Həyat təcrübəsinin subyektivliyindən azad olmalıdır, yəni müxtəlif tədqiqatçıların eyni bir termini söyləyərkən müxtəlif obyektləri nəzərdə tutması yolverilməzdir.
- Birmənalı olmalıdır, yəni eyni bir elm sahəsinin eyni bir termininin müxtəlif hallarda müxtəlif obyektləri təsvir etməsi yolverilməzdir.
- Dəqiq müəyyən olunmuş məna sahəsinə malik olmalıdır, yəni bu terminlərlə təsvir olunan dəqiq müəyyən olunmuş obyektlər çoxluğu olmalıdır.

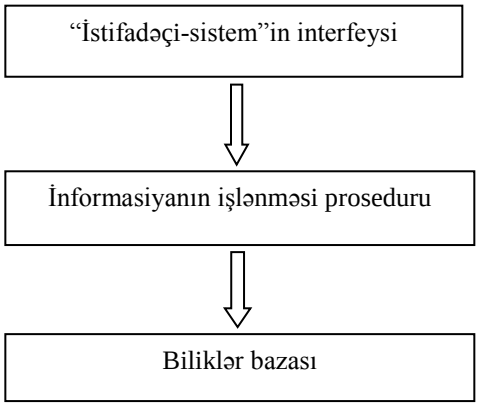
Bu sistemin elementlərinin mühüm xüsusiyyəti ondan ibarətdir ki, hər terminin dəqiq izahı var, termini anlamaq üçün onun izahını bilməklə yanaşı, izahında işlənən terminlərin də, öz növbəsində, izahını bilmək vacibdir. Terminoloji sistemin elementləri arasındakı əlaqə bir-biri ilə qarşılıqlı münasibətdə olan anlayışların iyerarxiyasını əks etdirir. Terminlərin özü kimi, terminlər arasındakı əlaqə də sahə mütəxəssisləri (ekspertlər), logistlər və dilçilər tərəfindən təyin olunur. Qurulmuş əlaqələrin xarakteri isə müxtəlif ola bilər. Qeyd etmək lazımdır ki, terminoloji sistem açıq sistemdir: bir tərəfdən sistemin elementlərinin köhnələrək tamamilə yox olması və ya ümumişlək sözə çevrilməsi – determinləşmə, digər tərəfdən isə sistemə daim yeni elementlərin – terminlərin axını (daxil olması) baş verir.

$P=(p_1, p_2, p_3, \ldots p_n)$ çoxluğunun xassələrinin hər biri ayrılıqda lokal funksional keyfiyyəti xarakterizə edir (məs. p_1 – tamlıq, p_2 – açıqlıq, p_3 – dəqiqlik və s.), birlikdə isə sistemi tam xarakterizə edirlər.

Burada informasiya texnologiyalarının əsas instrumental vasitəsi kimi çıxış edən informasiya sisteminə proqram və texniki vasitələr sistemi, həmçinin reallaşdırılmış bank verilənləri (biliklər bankı) kimi baxılır. Bu sistem müəyyən metodlar, metodik və normativ sənədlər (standartlar) çərçivəsində xüsusi işlənmiş informasiyanın istifadəçilərin sorğularına uyğun yığılması, axtarışı, tanınması, əldə edilməsi, saxlanması, qorunması, işlənməsi və ötürülməsini avtomatik və birqiymətli dəstəkləyir.

III. İNFORMASİYA SİSTEMİNİN ARXİTEKTURUNUN İŞLƏNMƏSİ

Sistemin arxitekturu bu sistemin bütün komponentlərinin qarşılıqlı əlaqəsini və yerinə yetirilən funksiyalarını, strukturunu, modelini, həmçinin sistemin özünü onu əhatə edən informasiya mühiti ilə birlikdə əks etdirir və reallaşdırır. O üç səviyyədə təyin olunur (Şəkil 1).



Şəkil 1. İnformasiya sisteminin arxitekturu

Sistemdə ilkin struktur vahidləri bütöv informasiya obyekti, daha dəqiq lüğətlərin məqalələridir. Burada obyektin hissəsi, məsələn, məqalələrin başlıqları fiziki olaraq aşağı səviyyənin bir verilənlər bazasında saxlana bilər, lakin məqalənin mətni və ya bu mətnin bir fraqmenti digər bazada yerləşə və tapıla bilər, lakin xarici interfeysin sorğusuna bütün məqalə başlığı ilə birlikdə bütövlükdə və eyni zamanda onun ilkin formasında verilir.

Sistemin arxitekturunun 3-cü səviyyəsinin reallaşdırılması aşağıdakı kimi həyata keçirilir:

- İnformasiya sisteminin biliklər bazasının konseptual modeli (1) asılılığının təqdimatı əsasında qurulur.
- Vahid biliklər sahəsinin təmin olunması üçün predmet sahələrinin ümumi anlayışlarının (M – terminlər, anlayışlar çoxluğu) tabeçiliyi formasında ontologiyaları qurulur.

Predmet sahəsinin ontologiyası - predmet sahəsinin formal təsviri deməkdir, adətən predmet sahəsinin ümumi terminoloji bazasını təyin etmək üçün tətbiq olunur. İnformasiya texnologiyaları və kompüter elmlərində ontologiya dedikdə obyektlər çoxluğu və onlar arasındakı əlaqələrin təsviri nəzərdə tutulur. Formal olaraq ontologiya terminlər, onların təsviri və çıxış qaydalarından ibarətdir. Beləliklə, ontologiya predmet sahəsinin modelidir və əsasən, biliklər bazası və biliklər bankının əsasını təşkil edir.

Ontologiyanın qurulmasında əsas məqsəd bilik modelini standartlaşdırmaq və maksimal (riyazi) təsvir etməkdir [2].

Ontologiyanın qurulmasının üstünlükləri aşağıdakılardır:

- Predmet sahəsi haqqında informasiyanın strukturundan qarşılıqlı istifadəni asanlaşdırır;
- Predmet sahəsinin modellərinin müxtəlif hissələrinin uyğunluğunu asanlaşdırır;
- Yeni istifadəçilər üçün modellərin əlyetərliyi və dəyişdirilməsi;
- Predmet sahəsinin biliklər strukturunun konkret verilənlərdən ayrılması.

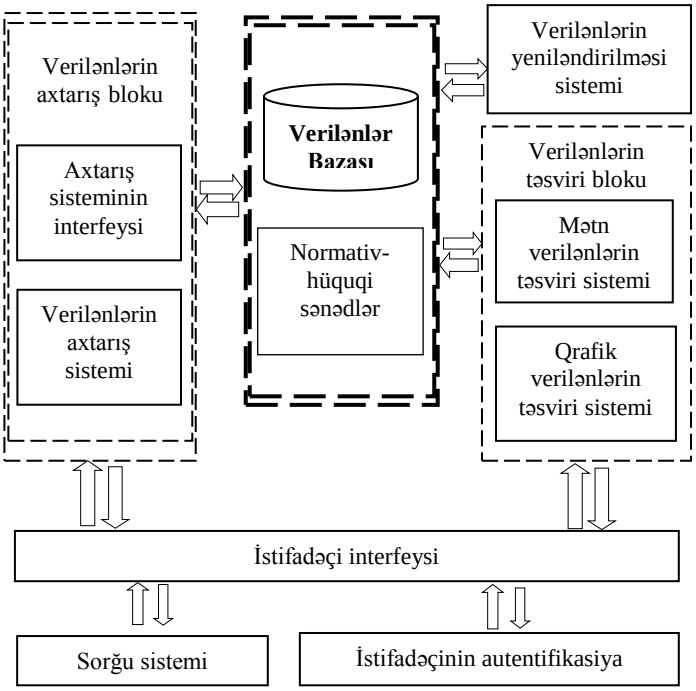
Ontologiyanın işlənilməsinin bir neçə mərhələsini ayırmaq olar. Predmet sahəsini müəyyən etdikdən sonra mövcud ontologiyaların axtarışını həyata keçirmək və onların təkrar istifadəsi imkanına baxmaq lazımdır. Sonra predmet sahəsinin terminlərinin tam siyahısını əldə etmək, siniflərin iyerarxiyasını işləmək və onların xüsusiyyətlərini təsvir etmək vacibdir. Sonrakı addım ontologiyanın elementləri arasındakı münasibətin təsvir edilməsidir. Bu mərhələdə anlayışın iyerarxiyanın hansı mərhələsində olması, obyektin sinfi olub-olmaması və s. müəyyənləşdirilir. Son mərhələdə siniflərin ayrı-ayrı nümunələri yaradılır. Predmet sahəsinin modelləşdirilməsinə ontoloji yanaşma yeni informasiya sistemlərinin işlənilməsinə və ilkin informasiya sistemlərinin interaperabelliyinin təmin olunmasına imkan yaradır. Hazırkı dövrdə ikisəviyyəli informasiya sistemləri paradiqması inkişaf edir. İnformatikada ontologiya müəyyən bilik sahələrinin konseptual sxemin (anlayış və konsepsiyaların müəyyən qaydalar üzrə qarşılıqlı əlaqələrindən ibarət semantik şəbəkə) köməyilə müfəssəl təsviridir. Adətən bu sxem verilənlərin strukturunu əks etdirir, bura obyektin bütün uyğun sinifləri, onlar arasında əlaqə və bu sahədə qəbul edilən qaydalar (teoremlər, məhdudiyyətlər) daxildir [3]. Ontologiya proqramlaşdırma prosesində real dünya və onun hissələri haqqında biliyin təqdimat forması kimi istifadə olunur, əsas tətbiq sahələri biznes proseslərinin modelləşdirilməsi, semantik veb və süni intellektdir.

Ontologiyanın təsviri dili formal dildir, bu dildən onun kodlaşdırılması üçün istifadə olunur, bir neçə oxşar formal dillər (OWL — Web Ontology Language, KIF - Knowledge Interchange Format (biliklə mübadilə formatı) və s.) mövcuddur. Verilənlərin təqdimatının müxtəlif modellərinin analizi ontologiyanın təsviri üçün instrumental vasitə kimi müxtəlif tərkibli semantik şəbəkələrin seçilməsini əsaslandırmağa imkan verdi. Semantik şəbəkə istiqamətlənmiş qrafdır, onun təpələri anlayışları, tilləri isə anlayışlar arasındakı münasibətləri ifadə edir. Anlayış hər hansı abstrakt və ya konkret obyektləri, münasibət isə bu obyektlər arasındakı əlaqələri ifadə edir. Semantik şəbəkədə münasibətlər "hissə-tam" tipli (sinif-qrup, element-çoxluq), funksional ("baş verir", "təsir edir" və s.), kəmiyyət xarakterli (çox, az, bərabər və s.), məkan xarakterli (uzaq, yaxın, altında, üstündə, içində və s.), məntiqi (və, və ya, yox), linqvistik və s. ola bilər [4].

Semantik şəbəkə tipli biliklər bazasında həllin axtarışı bazaya verilmiş hər hansı altşəbəkəyə uyğun sorğunu əks etdirən şəbəkə fraqmentinin axtarışı məsələsinə gətirilir. Bu model amerikalı psixoloq Quillian tərəfindən təklif olunub [5]. Semantik şəbəkənin reallaşdırılması anlayışlar iyerarxiyasının əks olunduğu tezaurusun (ayrı-ayrı predmet sahələrinin təsviri üçün aktiv alət) qurulması əsasında həyata keçirilir.

İzahlı lüğətdən fərqli olaraq tezaurus mənanın yalnız tərif vasitəsilə deyil, həm də sözün digər anlayış və onların qrupları ilə əlaqələndirilməsi vasitəsilə aşkar edilməsinə imkan verir. Tezaurusa uyğun olaraq münasibətlərin 5 tipinə baxılır: daha geniş istifadəli termin – yuxarıda; daha az istifadə olunan xüsusi termin – aşağıda; əlaqələndirilmiş termin – assosiasiya; termin üçün tam – hissə; termin üçün hissə - tam. Qeyd etmək lazımdır ki, ümumi halda biliklər sahəsi və professional fəaliyyət sahəsi üzrə ontologiyanın qurulması sahə standartları

səviyyəsinə və texniki materiallara qədər inkişaf tendensiyasına malikdir. Sistemin arxitekturunun 2-ci səviyyəsinin qurulması mətn informasiyasının axtarışı modelinə əsaslanır. Bu modelə sənəd və sorğuların təqdimatı, məna uyğunluğu kriteriyası, sorğunun nəticəsinin sıralanma (əhəmiyyətinə görə) mexanizmi və sənədin relevantlığının (sorğuya uyğunluq dərəcəsinin) qiymətləndirilməsi üçün əks-əlaqə mexanizmi aiddir. Bu modelin həyata keçirilməsi, reallaşdırılması bütün sənədlər toplusu üzrə tam mətnin axtarışı ilə əlaqəlidir. Axtarışın effektivliyi kriteriyaları qismində istifadəçinin sorğusuna uyğunluq dərəcəsi (relevantlıq), axtarış indekslərinin aktuallığı, etibarlılığı, istifadəçiyə xidmətin tezhərəkətliliyi müəyyən olunur.



Şəkil 2. İnformasiya sisteminin arxitekturunun konseptual modeli

Sistemin arxitekturunun 1-ci səviyyəsinin qurulması HTML dilinin teqlərindən istifadə etməklə hipetmətn formasında informasiyanın təqdimatına əsaslanan verilənlərə əlyetərliliyin interaktiv vasitəsinin reallaşdırılması, həyata keçirilməsini daxil edir. Qeyd edək ki, bu tədqiqatın tapşırıqlarının yerinə yetirilməsi prosesində İS-nin proqram təminatında texniki tapşırıq yerinə yetirilib və əsas tələblər müəyyən olunub: istifadənin rahatlığı, dizaynın uyğunluğu, əhatəlilik, açıqlıq, axtarış sisteminin olması, sorğu sisteminin olması, informasiyanın qorunma vasitələrinin olması, qoşulma və baxış, normativ sənədlər, materialların (terminoloji fəaliyyətlə bağlı kitablar, terminoloji lüğətlər və s.) kitabxanası. Bu sistemin yaradılması üzrə görülən iş Beynəlxalq standartlaşdırma təşkilatının müvafiq standartlarına uyğun aparılıb [6,7]. Beləliklə, hər hansı bir predmet sahəsi üzrə terminoloji informasiya sisteminin arxitekturunun qurulması onun konseptual modelini qurmağa imkan verir. İS-nin arxitekturunun konseptual modelinin həyata keçirilməsi, reallaşdırılması üçün proqram təminatının işlənilməsi zəruridir. Beləliklə, terminoloji informasiya sistemi bütün terminlər çoxluğu üzərində işlənən semantik şəbəkədir. O istifadəçilərə müxtəlif interaktiv xidmətlər təqdim edən tematik İnternet portalının köməyi ilə həyata keçirilir, burada əsas güc portalın axtarış sisteminin üzərinə düşür.

NƏTİCƏ

Azərbaycanda terminoloji fəaliyyət sahəsinin geniş mütəxəssis dairəsi üçün qərarların qəbulunu dəstəkləyən belə bir sistemin və onun proqram təminatının işlənməsi ilk addımdır. Qeyd edək ki, terminoloji informasiya sisteminin proqram təminatı üzrə müəyyən işlər görülüb. Proqram təminatı Linux əməliyyat sisteminin idarəçiliyi altında işləyən PHP proqram dilində işlənib. Burada hər hansı predmet sahəsinin axtarılan termini, həmin terminin mənası (izahı), mümkün dillərə tərcüməsi və s. metaməlumatlar öz əksini tapıb. Terminoloji informasiya sisteminin reallaşdırılması terminlərin təhlilinə yönəlmiş xüsusi biliklər bazasının yaradılmasına imkan verir. Bu sistem kompüter şəbəkələrində süni intellekt texnologiyasının tətbiqi üçün 2 qarşılıqlı əlaqəli aspektdən - şəbəkədə müəyyən predmet sahəsi üzrə biliklərin mübadiləsi və şəbəkə üzrə paylanmış biliyin istifadəsi ilə məsələnin qarşılıqlı həlli baxımından baza rolunu oynaya bilər.

TƏŞƏKKÜRNAMƏ

Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişafı Fondunun maliyyə yardımı ilə yerinə yetirilmişdir – **QRANT № EİF-2014-9(24)-KETPL-14/02/1.**

ƏDƏBİYYAT

[1] М.В.Новожилова, А.В.Ушеров-Маршак, Е.В.Латорец, И.А.Михеев. Разработка информационной терминологической системы в области бетоноведения. Харьковский государственный технический университет строительства и архитектуры, Тематический сборник "Системы обработки информации", 2010, випуск 6 (87)

[2] Н.С. Константинова, О.А. Митрофанова, Онтологии как системы хранения знаний, Санкт-Петербургский государственный университет, г. Санкт-Петербург,
http://www.ict.edu.ru/ft/005706/68352e2-st08.pdf

[3] https://ru.wikipedia.org/wiki/Онтология_(информатика)

[4] https://ru.wikipedia.org/wiki/Семантическая_сеть

[5] Quillian, M. R. Semantic memory. Semantic information processing, MIT Press; reprinted in Collins & Smith (eds.), Readings in Cognitive Science, section 2.1, 1968, pp. 227—270.

[6] ISO/TR 12618:1994. Средства автоматизированной обработки терминологических данных. Создание и использование терминологических баз данных и совокупности текстов. www.iso.org

[7] ISO 12207. «Процессы жизненного цикла программных средств». www.klubok.net/pageid313.html

Proqram Mühəndisliyində Big Data Problemləri

Tamilla Bayramova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
tamilla@iit.ab.az

Xülasə **— Məqalədə “Big Data” konsepsiyasının rellaşdırılmasında istifadə edilən əsas texnologiyalar və onların yaradılması üçün istifadə edilən proqramlaşdırma dilləri araşdırılmışdır. Böyük verilənlərin emalı üçün proqram vasitələrini işləyərkən meydana gələn problemlər təhlil olunmuşdur.**

Açar sözlər — Big Data, Hadoop, MapReduce, SCOPE, JAQL, NoSQL

I. GİRİŞ

Müasir dünyada informasiyanın həcmi eksponensial şəkildə artır, onun saxlanılması və emalı çətinləşir. Bunlar sosial şəbəkələrdən daxil olan verilənlər, fərdi verilənlər, sensor qurğularının göstəriciləri, tranzaksiyalar, administrativ verilənlər və s. ola bilər. Bu informasiyaların böyük hissəsi strukturlaşdırılmamış verilənlərin payına düşür. Böyük verilənlərdən istifadə edərkən bir sıra çətinliklər və suallar yaranır (məsələn, verilənlərə əlyetərlilik, etik məsələlər, verilənlərin yadda saxlanılması və metodoloji problemlər). Bu verilənlər cəmiyyətə, siyasətə və iqtisadiyyata əhəmiyyətli dərəcədə xeyir gətirə bilər. İnsanlar haqqında ətraflı verilənlərin olması cinayətkarlığın azaldılması, tibbi xidmətin keyfiyyətinin artması, idarəetmənin səmərəliliyinin yüksəldilməsi və s. bu kimi sahələrdə istifadə edilə bilər. Böyük həcmdə informasiyanın emalı üçün 2000-ci illərdə bir bir qrup konsepsiya, metod və instrumental vasitələr yaradıldı və bunlar “Big Data” texnologiyaları adı altında ümumiləşdirildi [1].

Big Data termini çoxsaylı mübahisələr yaradır, əksəriyyət belə hesab edir ki, bu ənənəvi üsullarla emal edilə bilməyən, strukturlaşdırılmış və strukturlaşdırılmamış böyük informasiya yığınıdır. Ümumilikdə isə Big Data aşağıda göstərilən əlamətlərlə xarakterizə olunur [2]:

- ***Həcm*** (volume) – verilənlər bazası emalı və saxlanılması ənənəvi üsullarla çox çətin olan böyük həcmli informasiyadan ibarətdir və onlarıla işləmək üçün yeni yanaşma və təkmilləşdirilmiş vasitələr tələb olunur;
- ***Sürət*** (velocity) – emalın nəticələrinin vaxtında alınması çox vacib amillərdəndir, həcm artdıqca emal üçün çox böyük sürət tələb olunur;
- ***Müxtəliflik*** (variety) – eyni zamanda müxtəlif formatlı strukturlaşdırılmış və strukturlaşdırılmamış informasiyanı emal etmək imkanı olmalıdır;
- ***Həqiqilik*** (veracity) – toplanan məlumatların düzgünlüyünə diqqət yetirilməlidir;
- ***Dəyər*** (value) – verilənlər şirkət üçün faydalı olmalıdır və ona xeyir gətirməlidir.

Məhdud zaman ərzində və ya verilən vaxt müddətində nəticələrin alınmasının vacibliyi böyük verilənlərin təqribi emalı metodlarının tətbiqinə gətirir. Big Data konsepsiyasının rellaşdırılmasında istifadə edilən əsas texnologiyalar və onların yaradılması üçün istifadə edilən proqramlaşdırma dilləri aşağıda araşdırılmışdır.

II. BİG DATA TEXNOLOGİYALARINDA İSTİFADƏ EDİLƏN PROQRAM VASİTƏLƏRİNİN BƏZİ PROBLEMLƏRİ

Böyük verilənlər (BV) müxtəlif mənbələrdən daxil olduğundan onları öz aralarında əlaqələndirmək, çevirmək və sənədləşdirmək lazımdır. Verilənlərin təhlili zamanı xətaların mənbəyinin və təbiətinin əvvəlcədən müəyyən olunması BV-də yaranan riskləri azaldır. Bunun üçün də verilənləri generasiya edən, massivlərin yaradılmasında və emalında istifadə edilən proqram vasitələrinə diqqət yetirilməlidir. İnformasiya sistemlərinin məhsuldarlığının və yaddaş qurğularının həcminin artması avadanlıqların imkanlarının qeyri məhdud olması haqda fikirlər yaradır. Hal-hazırda emal ediləcək verilənlər ilə onların saxlanılma imkanları arasındakı fərq sürətlə artır və verilənlərin səmərəli şəkildə idarə edilmə problemləri daha da aktuallaşır. Getdikcə artan verilənlərin təhlili və emali üçün keyfiyyətsiz proqram kodunun yazılması isə sistemlərin istismarı zamanı həlli mümkün olmayan problemlər yarada bilər [3].

Verilənləri idarə etmək üçün proqramları yazarkən düşünmədən obyekt-yönümlü metodologiyadan istifadə edilməsi çox pis nəticələrə gətirə bilər. Belə ki, obyektlər şəklində yaradılmış kiçik sorğuların sayı artdıqca onlar böyük informasiya axınına çevrilir və ən müasir çoxprosessorlu serverlər belə bu axını emal edə bilmir. Daxil olan sorğuların əksəriyyəti təkrarlandığı üçün onların və proqram kodunun optimallaşdırılması lazım gəlir. Verilənlərin analitik emalı məsələləri Big Data problemlərini həll etməyə yönəlib. Sorğuların cavablarını sürətləndirmək və istifadəçilərə daha keyfiyyətli xidmət göstərmək üçün müəssisələr Big Data texnologiyalarından istifadə etməli olurlar. Böyük həcmli verilənlərin emalı və analizini həyata keçirmək məqsədi ilə informasiya texnologiyaları (İT) sahəsinin nəhəngləri tərəfindən proqram-aparat platformaları yaradılmışdır.

Onlardan **Apache Hadoop** böyük həcmli (terabayt, petabayt) verilənləri yadda saxlamağa və BV-lə əlaqəli alqoritmləri paralel emal etməyə imkan verən paylanmış fayl sisteminin işlənilməsi və istifadə edilməsi üçün proqram platformasıdır (Software Framework) [4]. Hazırda Hadoop ən böyük İnternet-servis və İT şirkətlərində uğurla tətbiq olunur, bunlardan Yahoo, Facebook, Twitter, Amazon, Apple, Ebay və s. göstərmək olar. Hadoop platformasında əsasən 2 komponentdən istifadə edilirr:

- **Hadoop Distributed File System** (HDFS) verilənləri sürətli əlyetərliliyi təmin edən paylanmış fayl sistemidir;
- **MapReduce** hesablama klasterində böyük həcmli verilənlərin paylanmış emalı üçün proqram platformasıdır.

Bəzi hallarda əvvəl paralel sonra ardıcıl və yenidən paralel hesablamalar aparmaq lazım gəlir. MapReduce elə layihələndirilib ki, həm aralıq, həm də son nəticələr diskə yazılır. Ona görə də yuxarıda qeyd etdiyimiz tip məsələlərin həllində diskə yazma və oxuma vaxtlarını nəzərə alsaq hesablama müddəti bir neçə dəfə artacaq. Belə hallarda Spark proqram platforması tətbiq edilə bilər, burada əksər aralıq hesablamalar diskə deyil, yaddaşa göndərilir.

Cədvəl formalı verilənlərin təhlili daha asan olduğundan Hadoop infrastrukturunda SQL-yönümlü instrumental vasitələr də var [5]:

- **Hive** – HDFS fayl sistemində yadda saxlanılmış verilənlər üzərində kifayət qədər mürəkkəb sorğuları yerinə yetirməyə imkan verir.
- **Impala** – Cloudera şirkətinin məhsuludur. Sonuncudan fərqli olaraq MapReduce platformasından deyil, C++ dilində yazılmış öz platformasından istifadə edir.
- **Spark SQL** – Spark platformasında işləmək üçün yaradılmışdır.

Hadoop bazasında BV-nin emalı və analizi zamanı SQL – vasitələr bəzi hallarda kifayət etmir. Belə məsələlərin həllində NoSQL (Not only SQL) bazaları daha səmərəli olduğu üçün **HBase** proqramından istifadə edilir.

Geniş tətbiqinə baxmayaraq Hadoop proqram platforması bəzi məhdudiyyətlərə malikdir [6]:

- Hadoop klasterinin miqyaslanma imkanının məhduddur;
- Paylanmış hesablamaları yerinə yetirən alternativ proqram modeli yoxdur (Hadoop 1.0 ancaq Map Reduce hesablamalarını dəstəkləyir);
- Hadoop platformasında tək-tək imtina nöqtələrinin olduğuna görə etibarlılığa yüksək tələblərin qoyulduğu mühitlərdə istifadə edilə bilmir;
- Versiyalarındakı uyğunsuzluq: versiyasını yeniləndikdə bütün hesablama qovşaqlarında yenidən qurulmalıdır;
- Yeniləşən axın tipli verilənlərlə işi dəstəkləyə bilmir.

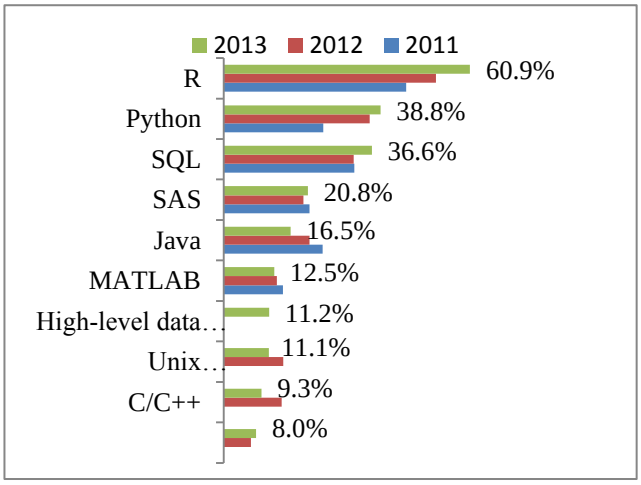
Çatışmazlıqlarına baxmayaraq özündə verilənlərin kütləvi-paralel emalı konsepsiya və texnologoyalarını birləşdirən Hadoop platforması BV-nin analizi üçün əsas platformadır.

Həm tədqiqat mərkəzlərinin, həm də böyük şirkətlərin diqqət mərkəzində olan belə maraqlı ideyalardan biri Hadoop və bulud texnologiyalarının birləşdirilməsidir. 2009-cu ildən başlayaraq IaaS/PaaS (İnfrastructure as a servis/Platform as a service) platformalarını işləyən iri şirkətlər bulud texnologiyaları ilə Hadoop platformasının birgə həllini təqdim etməyə başladılar.

Böyük verilənlərlə işləyən proqram təminatını yaradarkən proqram mühəndisliyinin əsas problemlərini nəzərə almaq lazımdır. Bunlar tələblərin təyin edilməsi, bu tələblərə uyğun proqram məhsulunun işlənilməsi və sınaq zamanı yaranan bir sıra problemlərdən ibarətdir. Belə proqramları işləyərkən aşağıdakı şərtlər nəzərə alınmalıdır:

- Miqyaslana bilmə: proqram getdikcə artan informasiyanı saxlamağa imkan verməlidir;
- Xətalara davamlı olmalıdır;
- İstənilən əməliyyat sistemində icra oluna bilməlidir;
- Məsələnin həllinin avtomatik olaraq paralelləşməsi;
- Çevik (strukturlaşdırılmış və strukturlaşdırılmamış verilənlərin saxlanması və təhlili) işləməlidir;
- Səmərəli olmalıdır;
- Risklər minimal olmalıdır.

Son illərdə Big Data texnologiyalarının yaradılmasında istifadə edilən proqramlaşdırma dillərinin populyarlığının artması bu sahəyə olan marağın artmasını göstərir. KDNuggets sorğu aparmışdır, 700 peşəkar proqramçının cavabları əsasında verilənlərin intellektual analizi, Data mining və elmi araşıdırmalarda ən çox istifadə edilən proqramlaşdırma dillərinin siyahısını vermişdir [7].



Şəkil 1. Proqramlaşdırma dillərinin reytinqi

Şəkil 1-dən göründüyü kimi proqramçıların verilənlərin statik emalı və qrafika ilə işləmək üçün yaradılmış R dilindən istifadəsi stabil olaraq artmaqda davam edir. Ümumilikdə Hadoop platforması ilə bu və ya digər səviyyədə əlaqəsi olan dillərdən istifadə 19% artmışdır. Bu proqramlaşdırma dilləri ilə yanaşı yeni proqramlaşdırma dillərindən də istifadə olunur, bunlara misal olaraq aşağıdakı dilləri göstərmək olar: SCOPE çevik yüksək səviyyəli proqramlaşdirma dilidir. İstifadəsi çox sadədir və analitik emalın müxtəlif klasterlərdə parallel yerinə yetirilməsi üçün nəzərdə tutulub [8].

Çox mürəkkəb və zəif strukturlu verilənlərin (məsələn, mətnlərin təhlili) emalı üçün JAQL (A JavaScript Object Notation Query Language) skript dili işlənilib [9].

NƏTİCƏ

Big Data texnologiyalarının işlənilməsində qazanılmış uğurlara baxmayaraq proqramçılar qarşısında bəzi açıq suallar qalır. Bunlar real zamanda prosesi dayandırmadan verilənlərin emal edən, xaker hücumlarının qarşısını alan məxfiliyi qorumaqla verilənləri emal edən və özü adaptiv proqram platformalarının işlənilməsidir.

ƏDƏBİYYAT

[1] Новиков Б.А., Графеева Н.Г., Михайлова Е.Г.Big Data: Новые задачи и современные подходы //Компьютерные инструменты в образовании, 2014, No4, с.10-18.
[2] Beyer Mark A., Laney Douglas. The Importance of «Big Data»: A Definition. Stamford, CT: Gartner, 2012.
[3] Tauheed Farhan, Nobari Sadegh, Biveinis Laurynas, Heinis Thomas, Ailamaki Anastasia. Computational Neuroscience Breakthroughs through Innovative Data Management, Proceedings of the 7th East-European Conference on Advances in Databases and Information Systems, Italy, 2013, p.14-27.
[4] http://www.codeinstinct.pro/2012/08/hadoop-overview.html
[5] http://www.zdnet.com/article/sql-and-hadoop-its-complicated
[6] http://www.codeinstinct.pro/2012/08/hadoop-design.html
[7] [http://www.kdnuggets.com/2013/08/languages-for-analytics-data-mining- data-science.html
[8] Chaiken Ronnie, Jenkins Bob, Larson Per-Ake, Ramsey Bill, Shakib Darren, Weaver Simon, Zhou Jingren. ˚ SCOPE: easy and efficient parallel processing of massive data sets, Journal proceedings of the VLDB Endowment, 2008. Vol 1, No. 2, p. 1265–1276.
[9] Beyer Kevin S., Ercegovac Vuk, Gemulla Rainer, Balmin Andrey, Eltabakh Mohamed Y., Kanne Carl-Christian, Ozcan Fatma, Shekita Eugene J.. ¨ Jaql: A Scripting Language for Large Scale Semistructured Data Analysis, Proceeding of the 37th International Conference on Very Large Data Bases, Seattle, WA, 2011, Vol. 4, No.2, p.1272-1283.

Böyük Verilənlər Bazasının İnternet-Reklam Problemlərinin Həllində Rolu Nədədir?

Kəmalə Haşımova
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
kama25@mail.ru, hashimovakama@gmail.com

Xülasə — **Məqalədə geniş həcmli məlumatların durmadan artımının gətirdiyi problemlər, onların xarakteri analiz edilərək, web-səhifələrdə olan reklamların effektivliyi nəzərdən keçirilmişdir.**

Açar sözlər — internet-reklam, Web saytlar, Big Data, marketing

I. GİRİŞ

Elm və texnologiyanın sürətli inkişafı nəticəsində internet-reklam məlumatları kəskin artır. İnternet və informasiya texologiyalarının sürətli inkişafı dünyanın bir cox ölkələrində müxtəlif sahələrdə- biznes, iqtisadiyyat, sənaye və menecment sahələrində əsaslı dönüş yaratdı.

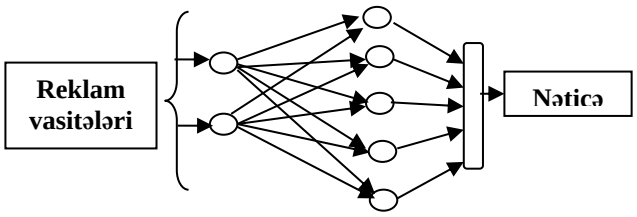
Müasir dövrdə şirkətləri “Mənim üçün əsas müştəri kimdir?” sualı düşündürür. İstehsalçını əsasən məhsul və xidmətə görə maraqlı müştəriyə, münasib yerdə, sərfəli qiymətə və uyğun zamanda təqdim etmək düşündürür. İqtisadiyyatın axsadığı dövrdə artıq məhsula görə müştəri deyil, müştərinin marağına görə məhsul istehsalı dövrüdür. Bu üsul inkişaf etdikdən sonra bazarda canlanma yarandı.

II. BİG DATA İLƏ REKLAMIN QARŞILIQLI ƏLAQƏSİ

İnformasiya texnologiyalarının sürətli inkişafı eyni zamanda informasiya bolluğuna səbəb olur. Bir çox ünsiyyət vasitələri- satış nöqtələrində xidmət, telefon zəngləri, elektron poçt, tədbirlər, görüşlər, veb-saytlarda qeyd formaları, reklam istinadları, çatlar, sosial şəbəkələr və s. toplanmış məlumatlar, müştərilər haqqında sorğular vahid anbara toplanır [1].

Müştəri ilə qarşılıqlı təsir zamanı münasibətlər haqqında bütün lazımlı informasiya qəbul edilməklə bazada saxlanılır.

Müştərilər haqqında yığılmış informasiyaların analizinin nəticələri əsasında müştəriləri seqmentlərə bölərək potensial reaksiya verir, şirkətin bu və ya digər məhsullarına tələbatı proqnozlaşdırır.



Şəkil1. Reklamın analizindən alınan nəticə

Toplanmış informasiyaların həcminin hər il demək olar ki, 2 dəfə artımını nəzərə alsaq bu artıq Big Datadır [7].

Big Data ifadəsinin günümüzdə populyar olmasının əsas səbəbi təbii ki, dayanmadan artan informasiyalardır.

Big Data çoxlu miqdarda məlumatın saxlanması və idarə olunmasıdır.

III. HƏDƏF SEQMENTİNİN SEÇİLMƏSİ.

Şirkət tərəfindən Web-səhifədə yerləşdirilən mal, ideya və ya xidmət haqqında informasiya məqsədinə çatmağı planlaşdırır. Ona görə də şirkətin məqsədinə çatmasına imkan yaradan istifadəçilərin Web-səhifəyə baş çəkməsi şirkət üçün daha maraqlıdır.

Şirkət tərəfindən təqdim olunanWeb-səhifə marketinq strategiyasının istifadəsi həm şirkəti öz hədəfinə nail olmağa, həm də rəqiblərlə müqayisədə istehlakçıların ehtiyaclarını təmin etməyə imkan verir.

IV. WEB-SƏHİFƏNİN KONSEPSİYASININ HAZIRLANMASI VƏ HƏYATA KEÇİRİLMƏSİ

Müştəriyə təsiretmə vasitələrindən biri Web-səhifədə reklam-marketinqin yerləşdirilməsidir. Web-səhifənin konsepsiyasının hazırlanması və həyata keçirilməsi zamanı reklamın düzgün və qüsursuz təqdim edilməsi Web-səhifənin müvəffəqiyyətli və uzunmüddətli olması üçün əsas şərtdir. Ona görə də Web-səhifədə təqdim edilən informasiya aşağıdakı kriteriyalara cavab verməlidir:

- Web-səhifə reklam hədəflərə uyğun olmalıdır;
- istehlakçıların hədəf seqmentinin xüsusiyyətləri nəzərə alınmalıdır;
- müştərilərin diqqətini cəlb etmək üçün müəyyən dərəcədə unikal olmalıdır;
- yerləşdirmə üsuluna uyğun olmalıdır;
- operativ olmalıdır. Web-səhifəyə olan marağın saxlanılması üçün onun daim yenilənməsi və müasirləşdirilməsi zəruridir;
- obyektiv və etibarlı olmalıdır. Əks təqdirdə yanlış və ya qeyri-dəqiq informasiya aşkar edən müştərilərin təkrar bu səhifəyə müraciət etmək ehtimalı aşağı olur;
- həddən artıq bezdirici xarakter daşımamalıdır.

Web-səhifədə şirkət və onun məhsulu, məhsul və xidmətlər kataloqu, dillerlərin siyahısı, ticarət nümayəndələrinin

ünvanları, sorğu məlumatları, yeniliklər, şəbəkə mağazaları, əks əlaqə bölməsi, vakansiyalar və s. haqqında müxtəlif materiallar təqdim etmək olar. Ona görə də Web-səhifədə yalnız istehlakçıların sorğularını təmin etmək imkanı verən informasiyanın yerləşdirilməsi məqsədəuyğundur. Web-səhifələr üçün mətnləri tərtib edərkən əsas aspektlərdən biri şirkətin xüsusiyyətlərini özündə əks etdirən texniki və əyləncəli informasiyalardan istifadə olunmasıdır.

Bir qayda olaraq əhəmiyyətli və maraqlı materialların diqqəti cəlb etməsi üçün qəzetlərin birinci səhifələrinin yuxarı hissəsində yerləşdirilir. Web-səhifələrin yaradılması zamanı belə qaydalar rəhbər tutulmalı və onlara əməl olunmalıdır.

Aparılan müşahidələr göstərir ki, istifadəçilərin əksəriyyəti Web-səhifənin aşağı hissəsində yerləşən linklərə, şəkillərə və mətnlərə çox az diqqət yetirirlər, bəzi istifadəçilər isə ekrana yerləşdirilməmiş reklama baxmağa maraq göstərirlər. Ona görə də ən vacib məlumatlar və Web-səhifənin bütün idarəetmə qurğuları səhifənin yuxarı hissəsində verilməlidir.

V. WEB-SƏHİFƏNİN EFFEKTİVLİYİNİN QİYMƏTLƏNDİRİLMƏSİ

Web-səhifənin effektivliyini iki üsulla qiymətləndirmək olar:

A) Serverin statistikasının köməyi ilə kəmiyyət analizi. Bunun üçün aşağıdakı göstəricilərdən istifadə etmək olar:

- Web-səhifəyə müraciətlərin sayı
- Web-səhifədə istifadəçi tezliyinin bir günlük statistikası
- Web-səhifəni nəzərdən keçirmiş ayrı-ayrı insanların sayı
- Web-səhifədə keçirilmiş vaxt
- İstifadəçilərin giriş və çıxışı üçün istifadə olunmuş səhifələr [2].

B) Aşağıdakı verilənlərin köməyi ilə keyfiyyət analizi:

- Web-səhifənin razılıq səviyyəsinin təyini məqsədi ilə potensial istehlakçıların sorğusu
- Sorğu prosesində istifadəçilərin ehtiyaclarının Web-səhifəyə (məzmun, dizayn, istifadənin rahatlığı və s.) uyğunluğunu təsvir edən komponentlərin əhəmiyyətlilik dərəcəsi və qiymətləndirilməsi
- Web-səhifənin istifadəçiləri tərəfindən daxil olan cavab reaksiyalarının öyrənilməsi

Bu mülahizələr tam qaydalar toplusu hesab edilə bilməz. İnternet şəbəkə sürətlə inkişaf etdiyindən qurulmasının marketinq prinsiplərinin də sürətlə dəyişəcəyini deməyə əsas verir.

Web-server internetdə malların hərəkətinin vacib elementi olduğundan onun iqtisadi effektivliyinin hesablamasının nəzərdən keçirilməsi zəruridir [3].

İnternet mühitdə Web-server əsasında müəssisənin elektron ticarət sistemlərinin qurulması seçilmiş variantın iqtisadi effektivliyi sisteminin tətbiqi onun hazırlanması və istismarı ilə bağlı xərclərə nisbəti kimi təyin oluna bilər.

$$E = R/CC$$

Tam xərclər (CC)

$$CC = K + EC$$

burada K - Sistemin konstruksiyasına kapital qoyuluşunun ümumi həcmidir;

EC – istismar xərcləridir.

Bu halda nəzərə almaq lazımdır ki, Web-serverin tətbiqi və istifadəsi bir anlıq məsələ zamandan asılı olaraq daim genişlənir. Müxtəlif ölkələrdə İnternetin istifadə səviyyəsi çox fərqlidir və ölkələrin əksəriyyətində çox aşağıdır [4].

Şirkətın öz saytını yaratdığı zaman istifadəçiyə şəbəkədə yeni informasiya resursunun mövcudluğu haqqında məlumat çatdırması zəruridr. Bu mərhələdə reklamın səmərəlilik göstəricisi məlumatlandırma əmsalı - məlumatlı istifadəçilərin sayının (S_0) Web-naşir saytının hədəf istifadəçilərin sayına (S) nisbəti ola bilər [5].

$$K_0 = S_0/S$$

Web-serverdə toplanmış məlumatların səmərəliliyinin qiymətləndirilməsi zamanı Big Datanın imkanlarından istifadə nailiyyəti nəzərə alınmalıdır.

NƏTİCƏ

Web-saytlardakı reklam vasitələrinin toplandığı məlumatlar böyük verilənlər bazasına toplanaraq lazım olan sahəyə paylanması reklam marketinqin inkişafında yeni mərhələ açdı.

Məlumatları köhnəltmədən, istifadəçinin davranış və reaksiyalarını real vaxtda faydalı informasiyaya çevirmək və məlumatları qiymətələndirmək rəqabət üçün ciddi bir üstünlükdür [6].

Big Data: Müxtəlif növ məlumatlar toplusudur və yeni tətbiqi üsullarla informasiya kütləsi və məlumat növlərini istismar edir. Tədqiqatlar göstərdi ki, məlumatların 80% strukturu olmayan məlumatlardır [8]. Məsələn şəkil, video, audio, mətn və s.

İqtisadiyyatın inkişafında reklam-marketinqin səmərəliliyinin artırılması, nəzərdə tutulmuş məqsədə çatmaq üçün Big Datanın imkanları və rolu danılmazdır.

ƏDƏBİYYAT

1. Understanding Big Data: analytics for enterprise class hadoop and streaming data, Paul C. Zikopoulos, McGraw Hill, 2012
2. Как создать сайт рекламного агенства, Лаборатория рекламы,маркетинга и Public Relations, № 7/8, 2001.
3. 2015-03-042015-03-04Учебн. пособие -М. : Издательский дом " Вильяме ", 2000 .
4. Курицкий А.Б. Интернет-экономика. Спб., 2000.
5. Основы электронной коммерции: Учебное пособие. Под ред.член-кор. РАН СЮ. Глазьева. -М.: МГУК, 2001
6. Understanding Big Data: analytics for enterprise class hadoop and streaming data, Paul C. Zikopoulos, McGraw Hill, 2012
7. www.mckinsey.com/mgi
8. www.computerworld.com.tr/featured/buyuk-veri-ile-gelen-tasarruf/

Böyük Verilənlərin Emalı Texnologiyaları İnnovativ İnformasiya İqtisadiyyatının İnkişafında Yeni Mərhələ Kimi

Ələvsət Əliyev[1], Roza Şahverdiyeva[2], Elza Musayeva[3]

[1,2,3]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*alovsat_qaraca@mail.ru*, [2]*roza02@rambler.ru*, [3]*elza.musayeva@gmail.com*

***Xülasə* – Məqalə formalaşmaqda olan innovativ informasiya iqtisadiyyatının Big data texnologiyaları ilə əlaqədar yeni inkişaf mərhələsinin elmi şərhinə həsr olunmuşdur. Yeni tipli iqtisadiyyatın bilik və informasiya resursları əsasında formalaşması xüsusiyyətləri təhlil edilmişdir. İKT yönümlü Milli Innovasiya Sisteminin üstünlükləri təsvir edilmişdir. İqtisadi infrastrukturun formalaşmasında İKT amilinin təsiri göstərilmişdir. İnkişaf səviyyəsinin qiymətləndirilməsi göstəriciləri izah edilmişdir. İnformasiya iqtisadiyyatının Big data texnologiyalarının tətbiqi ilə bağlı inkişaf tendensiyaları müəyyənləşdirilmişdir.**

Açar sözlər – informasiya iqtisadiyyatı, innovativ proseslər, Milli Innovasiya Sistemi, böyük verilənlər, verilənlərin emalı, informasiya infrastrukturu, informasiya və bilik resursları

I. GİRİŞ

Müasir dövrdə dünyanın bir çox qabaqcıl ölkələrində informasiya iqtisadiyyatının formalaşması xarakterik hal almışdır. İnformasiyanın əsas əlaqələndirici vasitə kimi çıxış etdiyi yeni tip iqtisadi münasibətlərin məcmusu olan innovativ informasiya iqtisadiyyatı müvafiq informasiya infrastrukturununu və informasiya resurslarından ibarət olan informasiya fəzasını formalaşdırır [1]. Hazırkı dövrdə formalaşmaqda olan yeni iqtisadiyyatda innovativ tədqiqatlar üçün lazım olan ilkin zəruri informasiyanın həcmi yüksək sürətlə çoxalmaqdadır.

Bu, onunla izah olunur ki, informasiyanın ilkin toplanılma və qeydiyyata alınma mənbələrinin sayının durmadan artmasıyla yanaşı, göstəricilərin say tərkibinin və müxtəlif təhlil kəsimlərinin çoxalması, ayrı-ayrı keyfiyyətlilik hallarının qiymətləndirilməsi üçün birbaşa göstəricilərin olmaması və s. kimi hallar da xeyli artmışdır. Bu isə öz növbəsində sürətlə dəyişməkdə olan müxtəlif xüsusiyyətli böyük həcmli informasiyanın emalı və idarə olunması üçün müasir idarəetmə mexanizmlərinin və vasitələrinin işlənilməsini və tətbiqini qaçılmaz edir. İnnovativ iqtisadi proseslərdə ilkin verilənlərin operativ və analitik təhlili inkişafa xidmət edir. Ancaq həmin verilənlərin həcminin kifayət qədər artması onların nizamlanmasını, strukturlaşmasını və operativ təhlilini tələb etdiyindən, müvafiq proseslərə Big data texnologiyaları tətbiq olunmalıdır. Bu isə iqtisadiyyatda yeni inkişaf mərhələsini formalaşdırır və ona təsiredici amillərin və proseslərin tədqiqini şərtləndirir.

II. YENİ TİPLİ İQTİSADİYYATIN BİLİKLƏRƏ VƏ İNFORMASİYAYA ƏSASLANMASI

İnformasiya amilinin nəzərə alınmasının çətinliyi, əhəmiyyəti, onun qeyri-müəyyənlıyi, mürəkkəbliyi, çatışmamazlıqları, toplanması problemləri kimi məsələlərin həlli müasir iqtisadiyyatın inkişaf istiqamətlərinin müəyyənləşdirilməsində və onun həyata keçirilməsində aktual məsələlərdən hesab olunur

İnformasiya və bilik bəşəriyyətin inkişafında mühüm rol oynamışdir. Lakin postsənaye cəmiyyətində informasiyanın və biliyin rolu o qədər artmışdır ki, onlar istehsal prosesində həlledici yer tutmuşlar. Hazırda iqtisadi artım daha çox informasiya və bilik resurslarının istifadəsi hesabına təmin olunur. İnformasiya və bilik istehsal faktorlarının daha səmərəli istifadəsi üçün əlavə şərait yaradır. Bilik yeni iqtisadiyyatda istehsal fakforu kimi ənənəvi istehsal faktorlarından fərqlənən bir sıra xüsusiyyətlərə malikdir. İnformasiya biliyin yaradılması üçün resursdur. İnformasiya iqtisadiyyatında bilik məqsədyönlü istiqamətlənmiş fəaliyyət üçün toplanmış potensialdır [2]. İqtisadi informasiya material, nemət və xidmətlərin istehlakı proseslərini muşayiət edir. Onun əhəmiyyətli hissəsi ictimai istehsalla bağlıdır ki, bu da istehsal informasiyası kimi qəbul edilir. İnformasiya iqtisadiyyatı dedikdə, elə iqtisadiyyat başa düşülür ki, orada ÜDM-ın böyük hissəsi informasiyanın və biliyin istehsalı, emalı, saxlanılması üzrə fəaliyyət hesabına formalaşır. Bu isə informasiya texnologiyalarının birbaşa inkişafı nəticəsində mümkün olur. Lakin bu iqtisadiyyatı bütövlükdə, kompüterlərin, telekommunikasiya avadanlılarının istehsalını və istifadə edilməsini, telekommunikasiya şəbəkələrinin inkişafını, proqram məhsulları və xidmətlərini və s. əhatə edən rəqəmsal iqtisadiyyat kimi də tam qəbul etmək olmaz. Beləliklə, bilik iqtisadiyyatında informasiya və bilik təkcə ənənəvi istehsal faktorlarının səmərəliliyinin yüksəlməsinə təsir göstərmirlər, həm də müstəqil istehsal faktoruna çevrilirlər.

Bilik iqtisadiyyatı şəraitində informasiya və bilik iqtisadiyyata və cəmiyyətə daha güclü təsir göstərir. Onlar yüksək səviyyədə iqtisadi resursa və rəqabətə davamlılıq faktoruna çevrilirlər. Ümumiyyətlə, dünyada əsas inkişaf perspektivləri bilavasitə bilik iqtisadiyyatı ilə əlaqədardır. Belə ki, iqtisadi artım daha çox yeni ideyaların tətbiqi və yeni texnologiyaların inkişaf etdirilməsi hesabına həyata keçirilə bilər. Hazırda bilik kifayət dərəcədə arta və yayıla bilmək qabiliyyətinə malik olduğundan, onun mübadiləsi hər bir fərdin biliyinin artmasına gətirir.

İnformasiya iqtisadiyyatının əsas məsələlərindən biri də informasiyanın yayılması və istifadəsi məqsədilə onun müxtəlif mənbələrdən əldə olunmasıdır. İnformasiya iqtisadiyyatında müxtəlif mənbələrdən alınmış informasiyadan iqtisadiyyatın digər sahələrinə təsir göstərmək üçün də istifadə edilir. Bu səbəbdən də iqtisadi elmlər sistemində informasiya iqtisadiyyatı xüsusi rola malik olur. İnformasiya iqtisadiyyatı bilik sahəsi kimi sahə iqtisadiyyatları ilə münasibətdə metaekonomika kimi xarakterizə edilə bilər. Həmin istiqamətdə informasiya texnologiyalarının maddi-texniki bazasının, onların effektiv tətbiqinin, elmi biliklərin yaradılmasının və onların ötürülməsi vasitələrinin iqtisadi aspektləri öyrənilir. Biliyin bu sahəsi informasiyanın resurs kimi tədqiqində və onun ayrı-ayrı aspektlərinini vahid obyektdə birləşdirilməsində özünü daha qabarıq göstərir. Onun əsas rolu elektron sferanın inkişafının qanunauyğun tendensiyalarının, informasiya-texnoloji ukladın inkişaf funksiyalarının və bu funksiyaların səmərəli yerinə yetirilməsi şərtlərinin və amillərinin tədqiqindən ibarətdir [3].

Formalaşmaqda olan informasiya iqtisadiyyatının tərkib elementlərinə 1)mövcud biliklərdən istifadə, eləcə də, yeni biliklərin yaradılmasını, sahibkarlığın inkişafını stimullaşdıran tənzimləyici əlverişli mühitin, 2)müasir informasiya infrastrukturunun, informasiya xidməti və bilik istehsalı sənayesinin formalaşması, 3)kifayət qədər bilik və təcrübəsi olan yüksək intellektli insanların əksəriyyət təşkil etməsi, 4)şəbəkə strukturlu səmərəli milli innovasiya sisteminin yaradılması və s. aiddir. Bilik iqtisadiyyatında bilik xammal və materiallardan daha vacib hesab edilir. O, dəyərin yaradılmasının əsas mənbəyinə çevrilir. İnsanlar və onların qarşılıqlı səmərəli münasibətləri bu iqtisadiyyatda əsas amilə çevrilir.

İnformasiya iqtisadiyyatının xarakterik xüsusiyyətləri sırasında 1)intellektual insan kapitalının maddi elementlərlə müqayisədə daha çox investisiya töhvəsi verməsini, 2)qeyri-maddi aktivlərin yüksək spesifikasiyalılığı və unikallığını, 3)şirkətlərin fəaliyyətində innovasiya tərkibinin nisbətən çox olması, 4)istehsalın yeniləşmə tempinin yüksək olmasını, 5)istehsal texnologiyalarının və ya məhsullarının və xidmətlərinin yeniləşməsinin fasiləsiz xarakter daşımasını qeyd etmək olar [4]. Proqnozlara görə, hazırda dünya, o cümlədən Avropa ölkələri 2020-ci ilədək texniki və iqtisadi cəhətdən əhəmiyyətli dərəcədə dəyişəcək [5]. Smart şəbəkələr kimi İKT-nin ağıllı tətbiqləri, intellektual logistika, intellektual müəssisə, intellektual şəhərlər qitədə və ondan kənarda geniş yayılacaqdır. 2020-ci ilə kimi, genişzolaqlı İnternetə qlobal giriş, mobil texnologiya, sosial media, ümumiləşdirilmiş maşın-maşın (M2M) kommunikasiyasından geniş istifadə, rəqəmsal məlumatların eksponensial artımı (40.000 eksabayt miqdarında) baş verəcəkdir.

Bazarların tədqiqi məlumatlarına görə, Avropada böyük həcmli məlumatlar bazarının 2016-cı ildə orta hesabla 32% artacağı təxmin edilir. İqtisadi məlumatların dəyərinin getdikcə artması nəticəsində araşdırmalar və innovasiyanı dəstəkləmək üçün Avropa Birliyində yaxın 5 ildə böyük miqdarda İT-vakansiyaların yaradılması nəzərdə tutulur. Almaniya bazarında 2016-ci ildə böyük həcmli məlumatların dəyərinin 1,7 milyard avroya çatacağı təxmin edilmişdir. Bu isə orta hesabla Avropa İttifaqında 20 milyard dollarlıq böyük verilənlər və biliklər bazarının formalaşmasını nəzərdə tutur. Göründüyü kimi, yeni iqtisadiyyatın bilik və informasiya ilə əlaqədar xüsusiyyətlərinin hərtərəfli öyrənilməsi yeni yanaşma, yeni metod və vasitələr tələb edir.

III. MİLLİ İNNOVASİYA SİSTEMİNİN İKT YÖNÜMLÜLÜYÜ ZƏRURƏTİ

Dinamik inkişaf edən ölkələrin təcrübəsi göstərir ki, iqtisadi tərəqqi səmərəli Milli İnnovasiya Sisteminin (MİS) yaradılmasından çox asılıdır. MİS-in yaradılmasının əsas şərtləri qlobal informasiya resurslarından istifadə, yeni biliklərin ölkədaxili ehtiyaclara uyğun adaptasiyası, milli inkişaf naminə onların tətbiqi və s. məsələləri əhatə edir.

Məlum olduğu kimi, innovasiya fəaliyyəti iqtisadi inkişafı və rəqabətliliyi təmin etmək üçün yeni ideyaların, elmi biliklərin, texnologiya və məhsulların müxtəlif istehsal və idarəetmə sahələrinə tətbiq edilməsidir. İnnovasiyanın əldə edilməsi, onun cəmiyyətin maddi mühitində təkrar istehsalı və reallaşdırılması ilə bağlı olan innovasiyadan istifadə prosesi innovasiya prosesini təşkil edir [6]. Aparilan təhlillər göstərir ki, ixtiyari sahədə innovasiya layihələrinin işlənməsi və tətbiqi İKT-dən istifadə etmədən mümkün deyildir və məhz bu sahənin inkişaf səviyyəsi qarşıya qoyulan məqsədlərə çatmağa təsir edən əsas amillərdəndir. İKT-dən istifadə etməklə innovasiyaların uğurlu tətbiqi üçün əlverişli şərait yaradılmalıdır. İnformasiya texnologiyaları sahəsinin inkişafı üçün bu sahədə müvafiq proqram və qanunlar qəbul olunmalıdır. Formalaşmaqda olan MİS-in infrastrukturunda İKT daha vacib və önəmli yer tutmalıdır. Burada iki amilə xüsusi fikir verilməlidir: 1)İKT-nin özü innovativ vasitə kimi bütün digər layihələrin tərkibində iştirak edir və onun innovativ səviyyəsini daha da artırır. 2)İKT sferasının özünün innovativ inkişafına daxili sahə innovasiyaları vasitəsilə nail olunur ki, bunun da vasitəsilə yenə də digər sahələr inkişaf etdirilir.

İnformasiya cəmiyyətinə keçid istiqamətində İKT yönümlü innovasiya sisteminin yaradılması məqsədi ilə elm və təhsil sahəsində informasiya və biliklərin mübadiləsini təmin edən kompüter şəbəkələrinin inkişafı və əhalinin bütün təbəqələrinin informasiya resurslarına müraciət imkanlarının genişləndirilməsi vacib məsələlərdəndir.

Ümumiyyətlə, innovativ inkişaf yolunda ənənəvi olmayan qərarların qəbul olunması digər prosesdə risklərin meydana çıxması ilə bağlı olur. İnnovasiyaların müəyyən sahəyə tətbiqi həmin sahədə yeniliyin dərk olmasını tələb edir. İKT yönümlü MİS-in formalaşmasının başqa bir aspekti sosial-iqtisadi informasiyasının toplanması, saxlanılması, təhlili və qeydiyyatının vahid sisteminin yaradılmasıdır. İKT vasitələrinin üstünlük təşkil etdiyi MİS-in strukturunda Big data texnologiyaları mühüm funksiya və vəzifələrin yerinə yetirilməsinə imkan yaradır.

IV. İQTİSADİYYATIN İNFRASTRUKTUR ELEMENTLƏRİNİN FORMALAŞMASINDA İKT AMİLİ

Yeni iqtisadi şərait əmtəə və xidmətlərin inkişaf edən bazarlarının informasiya təminatını, iqtisadiyyatın inkişafı üçün uyğun informasiya infrastrukturunun, informasiyanın mübadiləsi və yayılmasının effektiv sisteminin yaradılmasını tələb edir. İnformasiya bazarının infrastrukturu kimi informasiya məhsulları və xidmətləri təklif edən insanlar və ya təşkilatlar qrupunu birləşdirən müxtəlif xarakterli və fəaliyyətli sektorların məcmusu çıxış edir. İnformasiya infrastrukturunu biliklərin və informasiyanın əlyetərliyini, onların ötürülməsini və nəticə etibarilə, ictimai istehsalda informasiyanın tətbiqi üçün şəraitin yaradılması mexanizmi kimi biliklərin tətbiqinin mümkünlüyü üçün yaradılmış imkan və şərait kimi xarakterizə etmək olar [7]. İqtisadiyyatın dayanıqlı və səmərəli inkişafı onun strukturundan birbaşa asılıdır. Struktur dəyişikliyi iqtisadiyyatın sektorları arasındakı münasibətin dəyişdirilməsidir. İqtisadiyyatın strukturu dinamik xarakter daşıyır və bir sıra kateqoriyalarla ifadə olunur. İqtisadiyyatın strukturunun dəyişilməsinin əsas faktoru elmi-texniki inkişafdır. O, iqtisadiyyatda yeni istehsal sahələri və sektorları yaradaraq onu dəyişir və müasirləşdirir. Struktur dəyişikliyi həm də yeni texnoloji ukladın üstünlük təşkil etdiyi iqtisadi infrastrukturun yaranmasıyla xarakterizə olunur.

İqtisadiyyatın struktur dəyişikliyinə bazar və onun dövlətlə münasibəti də təsir edir. İqtisadiyyatın və bazarın strukturları arasında bir çox istiqamətlərdə qarşılıqlı asılılıq mövcuddur. İqtisadiyyatın yeni yaranmış strukturu da eyni zamanda bazarlara fəal təsir edir. Bazarların rəqabətədavamlı mallarla və xidmətlərlə təmin olunmasına imkan yaradır, məşğulluğun və əmək bazarının strukturuna təsir edir, kapital dövriyyəsini sürətləndirir. İqtisadiyyatın yeni sektorlarının yaranmasıyla yeni bazarlar da meydana çıxır. Daha səmərəli iqtisadi struktur siyasətinin həyata keçirilməsi üçün dövlətin tənzimləmə mexanizmlərinin tətbiqinə ehtiyac yaranır. İqtisadi strukturun dəyişilməsinə təsir edən faktorlardan biri də iqtisadiyyatın qloballaşmasıdır. İKT-nin o cümlədən genişzolaqlı İnternetin yayılması, Böyük verilənlərin tətbiqi iqtisadiyyatın bir çox sektorlarında, xüsusilə də xidmət sahəsində əmək məhsuldarlığının və rəqabətədavamlılığın artmasına təsir edərək iqtisadiyyatın struktur dəyişikliyində əhəmiyyətli rol oynayır. Qloballaşma şəraitində Big data texnologiyaları əsasında göstərilən xidmətlər bütün dünyada resurslardan daha optimal istifadə etməyə imkan verdi. O, bazarların genişlənilməsi hesabına əmək məhsuldarlığının, biznesin səmərəliliyinin artmasına şərait yaradaraq iqtisadi artımın sürətlənməsinə əhəmiyyətli töhfə verdi. Xidmətlərin çatdırılmasının, yayılmasının və koordinasiyasının çox hissəsinin Big data texbologiyaları vasitəsilə həyata keçirilməsi, bütovlükdə, gəlirlərin artmasına imkan yaratmaqla yanaşı, eyni zamanda, peşələrin sabitləşmiş funksiyalarının dəyişməsinə təsir etdi. Big data texbologiyalarının təsirinin müxtəlif aspektlərini ümumiləşdirərək belə nəticəyə gəlmək olar ki, iqtisadiyyatda baş verən struktur dəyişiklikləri və orada yeni sektorların yaranması onların inkişaf sürəti ilə düz mütənasibdir [8, 9]. Tədqiqatlar göstərir ki, Big data texbologiyalarının tətbiqi məhsuldarlığın artmasına və iqtisadiyyatın struktur dəyişikliyinə müsbət təsir edir.

V. İNFORMASİYA İQTİSADİYYATININ İNKİŞAF SƏVİYYƏSİNİN QİYMƏTLƏNDİRİLMƏSİNDƏ BİG DATA TEXNOLOGİYALARININ TƏTBİQİ

Cəmiyyətin və iqtisadiyyatın müasir inkişaf dövrünü qiymətləndirmək üçün bir sıra indikatorlar sistemi (İS) işlənilmiş və geniş yayılmışdır. Buna səbəb onların sosial-iqtisadi proseslərin təhlili nəticəsində əldə olunmuş məlumatların asan şərhini verə bilməsi. Eyni zamanda, ölçmə sistemində yeni kompozit indikatorlardan (composite indicators) da geniş istifadə olunmasıdır.

Bəzi kompozit indikatorlar (Kİ) sisteminə aşağıdakıları daxil etmək ola [10]: 1)İnformasiya cəmiyyəti indeksi, 2)Şəbəkə hazırlığı indeksi, 3)İKT-nin inkişafı indeksi, 4)Elektron hazırlıq indeksi, 5)Bilik iqtisadiyyatı indeksi, 6)Qlobal innovasiya indeksi, 7)Qlobal rəqabətədavamlılıq indeksi və s.

Bilik iqtisadiyyatı indeksi (Knowledge Economy Index, KEI) müxtəlif ölkələrdə və regionlarda elmtutumlu iqtisadiyyatın inkişaf səviyyəsini xarakterizə edən indikator kompleksinin əsasında hesablanır. KEI Dünya Bankı tərəfindən “İnkişaf üçün bilik” (“Knowledge for Development”) proqramı çərçivəsində hazırlanmış biliyin qiymətləndirilməsi metodologiyasının bir hissəsidir. İnnovasiya indeksi biliklərə əsaslanan inkişaf modelinə keçidin hazırlığını qiymətləndirir və 76 göstəricini özündə birləşdirən dörd tərkib hissəsindən ibarətdir: “İnstitutsional rejim”, “Təhsil və əhalinin ixtisaslaşması”, “Səmərəli innovasiya sistemi”, “Müasir İKT infrastrukturu”. Bu indeks 0-10 ballıq şkala üzrə ölkələrin reytinqini müəyyən edir. Bu indeksə görə, dünya reytinqində ilk yerləri uyğun olaraq İsveç (9.43 bal), Finlandiya (9.33 bal), Danimarka (9.16 bal), Niderland (9.11 bal), Norveç (9.11 bal) tutur. Azərbaycan bu indeksə görə 4.56 balla 79-cu yerdə qərarlaşmışdır.

Göstərilən beynəlxalq kompozit və milli inkişaf indekslərinin hazırlanması çox böyük həcmdə ilkin verilənlərin paylanmış şəkildə emalını tələb etdiyindən həmin proseslər Big data texnologiyalarının tətbiqi olmadan mümkün deyildir. Belə ki, bir çox müəssələlərdə böyük həcmli verilənlərdən istifadə aşağı səviyyədədir. Məsələn, Orta Asiya, Latın Amerikası, Şimali Amerikanın 47% müəssisələrində böyük həcmli verilənlərdən istifadə səviyyəsi 40% təşkil edir.

Böyük həcmli verilənlərdən istifadə edən müəsisələrin içərisində investisiya qoyuluşlarının orta gəlirliliyi 2012-ci ildə 40%-dən çox olmuşdur. İnvestisiya qoyan müəəsəsəsələrin rəhbərləri müəssisələrdə böyük həcmli verilənlərdən istifadə etməklə daha çox gəlir götürməyi əsas məqsədlərdən biri hesab edirlər [11]. Ona görə də müasir biznes-proseslər yeni İKT imkanlarına açıq olmalı və onların nailiyyətlərindən yararlanmalıdır. Hazırda yeni proqram həllərinin 80%-dən çoxu bulud texnologiyalarından istifadə etməklə həyata keçirilir. Mobil trafikin 95%-ni verilənlər təşkil edir. Hər gün minlərlə, milyonlarla qetabayt verilənlər generasiya olunurlar. Sosial şəbəkələrin milyonlarla istifadəçisi informasiya mənbəyidir. Həmin şəbəkələrdə milyonlarla proqram həlləri və saytlar inteqrasiya olunur. Onlar arasında mürəkkəb informasiya mübadiləsi baş verir. Müasir biznes böyük verilənlərlə yanaşı, sürətli verilənləri də tələb edir. Ona görə də böyük verilənlərin iqtisadiyyatının yenidən qurulmasına, formalaşmasına və səmərəli təşkili və idarə olunmasına kəskin ehtiyac hiss edilməkdədir. Böyük verilənlərə, bir qayda olaraq, aşağıdakı məlumat tipləri aiddir: 1)müəssisənin ənənəvi məlumatları (buraya - CRM sistemləri üzrə müştəri məlumatları, ERP məlumatlarının tranzaksiyası, veb mağazaları üzrə əməliyyat məlumatları və s. daxildir), 2)maşın – generasiyası/sensor məlumatları, 3)sosial məlumatlar -müştəri ilə əlaqə axınları, saytlar, mikro-bloqlar, sosial media-platformaları, Facebook, Twitter və s. kimi sosal şəbəkələr. Böyük həcmli verilənlərin tətbiqi ilə bağlı formalaşan xüsusi perspektivi olan sektorlara aşağıdakıları aid etmək olar [12]: 1)səhiyyə, 2)mobil rabitə, nəqliyyat və logistika, 3)energetika, 4)ictimai sektor, 5)media, kontent, 6)istehsal, 7)pərakəndə satış, 8)ətraf mühit və s.

Böyük həcmli verilənlər Avropanın bilik iqtisadiyyatında əsas hərəkət verici qüvvə hesab olunur [5]. İqtisadiyyatın bütün digər real sektorlarında böyük həcmli məlumatlardan istifadə innovasiya amili kimi qiymətləndirilir. Big datanın səhiyyə xidmətlərindən təhsilə qədər, sudan enerji alınmasının idarə olunması, stasionar müəssisədən mobilliyə qədər hər yerdə tətbiq olunması, iqtisadi və idarəetmə proseslərində yeni biznes modellərinin inkişaf etdirilməsinə, məhsuldarlığın artımının dəstəklənməsinə, sosial sektorlardakı fəaliyyətin inkişaf etdirilməsi sahəsində problemlərin həllinə əlavə imkanlar yaradır.

Böyük həcmli məlumatların təsirinin qiymətləndirməsi prosesində bir çox amillərə diqqət yetirilməlidir [13, 14] 1)məlumat mənbələrindən istifadə imkanlarının genişləndirilməsi, 2)aktual məlumatların iqtisadi qiymətləndirilməsi, 3)biznes imkanlarının dəstəkləməsi üçün lazım olan böyük həcmli verilənlər texnologiyaları və alətlərinin inkişaf etdirilməsi, 4)ölçülə bilən və dəyər təmin edən məlumatların idarə olunması üzrə biznes modellərinin inkişaf etdirilməsi, 5)sınaq-benchmarking texnologiyaları, tətbiqi proqramlar və biznes modelləri, 6)bacarıq və praktik təcrübə problemi, 7)təhlükəsizlik və məxfiliklə bağlı məsələlər, 8)məlumatlarda etibarlılıq dərəcəsinin yüksəldilməsi və s.

İnformasiya və bilklərin idarə olunmasında aşkaretmə, müşahidə, əldəetmə, yayma, tətbiq, ləğvetmə kimi mərhələ və proseslərdən istifadə olunur. Korporativ biliklər elmi tədqiqatlar, təcrübi eksperimentlər, patent əldə etmək, texnologiya gətirmək və s. kimi yollarla əldə edilir. Geniş yayılmış üsullardan biri də biznes-proseslərin optimallaşdırılmasını reallaşdıran ERP, CRM, MRL, S&OP, PLM sistemləri kimi proqram təminatlarının və ya yeni məhsulların layihələndirməsinə imkan verən CAD/CAM/CAE kimi həll modellərinin əldə edilməsidir [15] .

Böyük həcmli verilənlər texnologiyası ikili funksiya rolunu daşıyır. Onlardan biri onun iqtisadiyyatın inkişafında hərəkətverici rola malik olması ilə, digəri isə iqtisadiyyatın real sektorunun yaranması ilə əlaqədardır [16, 17]. Ona görə də Big data iqtisadiyyatının öyrənilməsinə onun öz prinsip və metodlarının göstəricilərinin ayrıca öyrənilməsinə də ehtiyac vardır.

NƏTİCƏ

İnformasiya iqtisadiyyatının mahiyyəti istehsal və istehlak arasında, maddi və qeyri-maddi istehsal mərhələləri arasında informasiyanın əsas əlaqələndirici vasitə kimi çıxış etdiyi yeni tip iqtisadi münasibətlərin reallaşdırılmasından ibarətdir. Müasir informasiya texnologiyalarının geniş tətbiqi və istifadəsi ilə yeni tipli dinamik sistem kimi formalaşan informasiya cəmiyyəti iqtisadiyyatının inkişaf perspektivləri məhz sərvətin ən əhəmiyyətli potensial mənbəyi rolunda çıxış etdiyi informasiya resurslarının olmasındadır.

Araşdırmalar göstərir ki, Böyük verilənlərdən istifadə bir çox imkan və yeniliklərin yaranmasına təkan verəcək. Birinci növbədə innovativ iqtisadiyyatda və idarəetmə proseslərində şəffaflığa nail olunur. Bunun bariz nümunəsi E-hökumət portalının yaradılmasında və əhaliyə elektron xidmətlərin göstərilməsində özünü büruzə verir. İqtisadiyyatın diversifikasiyalaşmasında, fəaliyyət nəticələrinin yaxşılaşdırılmasında zəruri olan şəraitin yaradılmasında da bu texnologiyalar mühüm rol oynaya bilər. Big data texnologiyaların tətbiqi məhsul və xidmətlərin fərdlərə uyğunlaşdırılmasında, qərar qəbuletmə sistemlərinin dəstəklənməsində, yeni innovativ biznes-modellərin, məhsul və xidmətlərin işlənilməsində, rəqabətə davamlılığın artırılmasında əvəzolunmaz rola malik olacaqdır.

ƏDƏBİYYAT

[1] R.M.Alguliyev, A.G.Aliyev. Elaboration Development Strategy and the Analysis of Trends Shaping the Information Economy in Azerbaijan on the Basis of the Experience of Foreign Countries. WASET. International Journal of Social, Education, Economics and Management Engineering Vol:9, No:3, 2015. p. 935- 938.

[2] Д.М.Губайдуллина. Необходимость совершенствования инфраструктуры информационной экономики в современной России. Социально-экономические явления и процессы, №4 (020), 2010, 25-27.

[3] А.М.Михайлов, М.В.Михайлов. Информация и знания в системе факторов постиндустриального производства. Ж. Экономические науки, 2010, №7(68), ст. 49-53

[4] Н.С.Микова. Анализ систем мониторинга развития экономики знаний и информационного общества. Академический журнал «Интеллект. Инновации. Инвестиции» № 2/2011, ст. 53-64

[5] European Big Data Value contractual Public Private Partnership (cPPP) -Strategic Research and Innovation Agenda. 2014.

[6] Р.М.Алгулиев, А.Г.Алиев. Некоторые аспекты формирования ИКТ-ориентированных национальных инновационных систем. Проблемы современной экономики. Евразийский международный научно-аналитический журнал. № 4, 2011, стр. 1-8.

[7] Ə.Q. Əliyev. Müasir iqtisadiyyatın strukturlaşmasında IKT-nin təsiri məsələləri. AMEA İnformasiya Texnologiyaları İnstitutu. “İnformasiya Cəmiyyəti Problemləri” jurnalı. №1, 2015, s.24-33

[8] McConnell, Elizabeth Anne, Birkett, Michelle, Shattell, Mona. The Future of Big data: Innovative Methodological Approaches. 2015, vol. 36. Issue 6, p.478-480. www.web.a.ebscohost.com

[9] Ю.А.Зеленков. Суперкомпьютерные вычисления в контексте экономики знаний. Суперкомпьютерные дни в России 2015 // Russian Supercomputing Days 2015, стр. 465-475

[10] Ə.Q.Əliyev. İnformasiya iqtisadiyyatının inkişaf səviyyəsinin qiymətləndirilməsi üzrə indikatorlar sisteminin işlənilməsi məsələləri. AMEA İnformasiya Texnologiyaları İnstitutu. “İnformasiya cəmiyyəti problemləri” jurnalı. 2014, №1, səh. 65-74.

[11] Sofia Berto Villas-Boas. Big Data in Firms and Economic Research. Applied Economics and Finance, Vol.1, №1, 2014. pp.65-70.

[12] Tushar M. Chavan, S. P. Akarte. Opportunities and challenges of big data in economics research and enterprises. International Journal of Computer Science and Mobile Computing. Vol. 3, Issue 4, 2014, pg.1155 – 1161

[13] Ike Picone, Caroline Pauwels. Belgium: big changes in a small news economy. Series Title Media Business and Innovation. 2013. pp 149-162. Springer Berlin Heidelberg

[14] Max Nathan, Anna Rosso, Francois Bouet. Mapping ‘Information Economy’ Businesses with Big Data: Findings for the UK. 2014. Discussion Paper № 8662

[15] Wuheng Luo. Enterprise data economy: A hadoop-driven model and strategy. Big Data, IEEE International Conference. 2013. pp.65- 70.

[16] Liran Einav. Jonathan Levin. Economics in the age of big data. Science 346. 2014. www.sciencemag.org

[17] Guichuan Zhou, Xiaowen Jie. Research on Big Data Under Enterprise Technological Innovation Management. International Conference on Management Science and Engineering Management. 2014. Vol. 280. pp 443-453. http://link.springer.com

Virtual Məkanda Onomastik Vahidlərin Transliterasiyası

Yedgar Cəfərov
AMEA Nəsimi adına Dilçilik İnstitutu, Bakı, Azərbaycan
yedgar@yandex.ru

***Xülasə* — Məruzə virtual məkanda onomastik vahidlərin (şəxs adları, soyadlar, coğrafi adlar və s.) transliterasiyası probleminə həsr edilmişdir. Xarici adların Azərbaycan, Azərbaycan adlarının xarici dillərdə transliterasiyası ilə bağlı mövcud vəziyyətin təhlili verilmiş, bu sahədə özünü göstərən problemlərin həlli üçün müəyyən təkliflər irəli sürülmüşdür.**

Açar sözlər* — *onomastika, transliterasiya, transkripsiya, xüsusi ad, əlifba, orfoqrafiya, orfoepiya

I. GİRİŞ

Onomastik vahidlərin, xüsusilə də şəxs adları və toponimlərin yerli dillərdən xarici, xarici dillərdən isə yerli dillərə düzgün transliterasiyası qloballaşma şəraitində müasir Azərbaycan dilçiliyinin ən aktual problemlərindən biridir. Çünki qloballaşma dövründə, virtual dünyanın formalaşması prosesində milli identikliyin, onun ayrı-ayrı subyektlərinin (adlar, soyadlar, coğrafi adlar, tarixi hadisələr və s.) qorunması istiqamətində transliterasiya çox mühüm əhəmiyyət kəsb edən məsələ kimi çıxış edir [1, s. 43].

Transliterasiya - bir yazı sistemindəki hərflərin başqa yazı sisteminin hərfləri ilə verilməsidir. Qrafik transliterasiyanın 4 əsas tipi vardır: 1) bir qrafik elementin digər qrafik elementlə əvəz edilməsi; 2) bir qrafik elementin bir neçə digər qrafik elementlə əvəz edilməsi; 3) bir neçə qrafik elementin digər bir qrafik element ilə əvəz edilməsi; 4) bir neçə qrafik elementin digər bir neçə qrafik element ilə əvəz edilməsi.

II. PROBLEMİN AKTUALLIĞI

Təxminən iki əsrə yaxın müddət ərzində Rusiya imperiyasının tərkib hissəsi kimi rus dilinin dominantlığı şəraitində yaşamış Azərbaycan xarici aləmlə, demək olar ki, bütün əlaqələri məhz bu dil vasitəsilə qurmuşdur. Bu gün dilimizdə işlənən Avropa mənşəli sözlərin, o cümlədən onomastik vahidlərin mühüm qismi ana dilimizə məhz rus dili vasitəsilə, bu dildə transliterasiya mərhələsini keçdikdən sonra daxil olmuşdur. 1939-cu ildən etibarən Azərbaycanda kiril qrafikasının tətbiqi də Avropa mənşəli sözlərin ana dilimizdə transliterasiyasının rus dilinin normalarına uyğunlaşdırılmasına böyük təsir göstərmişdir. Azərbaycan dilində xarici adlar səslənmə forması əsas götürülməklə kiril qrafikasının hərflərindən istifadə edilərək yazılmışdır. Orijinal dili bilməyən tərcüməçilərin mətni rus dilindən Azərbaycan dilinə çevirərkən transliterasiyanı da rus dilindən olduğu kimi qəbul etmələri bu sahədə bir sıra səhvlərin yaranmasına səbəb olmuşdur. Məsələn, fransız yazıçı Viktor Hüqonun adı dilimizdə *Üqo* şəklində transliterasiya edilməli ikən, fransız dilində /h/ fonemi mövcud olmadığı halda, rus dilindəki *Гюго* nümunəsinə əsaslanaraq bu ad Azərbaycan dilində də *Hüqo* şəklində sabitləşmişdir.

Başqa bir misalda, Perulu yazıçı, ədəbiyyat üzrə Nobel mükafatı laureatı (2010) *Mario Barqas Yosa*-nın *(Mario Vargas Llosa)* adı, ispan dilinin oxunuş qaydalarına görə *Barqas Yosa* şəklində keçməli ikən sırf rus dilinin dominantlığı səbəbindən *Lyosa (rusca: Льоса)* şəklində qeyd olunur.

Fransız yazıçısı, filosof Albert Kamünün adı da rusca *Камю* yazıldığına görə uzun illər Azərbaycan dilində *Kamyü* şəklində qeyd olunmuş, bu yanlışlıq yalnız son illər aradan qaldırılmışdır [2]. Əcnəbi xüsusi adların dilimizə transliterasiyası zamanı rus dilinin təsiri nəticəsində yol verilən qüsurlarla yanaşı, öz məzmun və formasında yüksək milli potensial daşıyan bir sıra Azərbaycan onomastik vahidlərinin xarici dillərə rus dili vasitəsilə yanlış şəkildə transliterasiya edilməsi də o dövrün milli mədəniyyətimizə “qazandırdığı” problemlərdən olmuşdur. Qərb mətbuatında hələ də Azərbaycan adının bəzən *Azerbaidzhan, Gəncə* adının *Gyandzha, Gandja* və s. *Sumqayıt* adının *Sumgait, Gədəbəy* adının *Gedabek* kimi yazılması sovet dövründə bu sahədə yol verilmiş yanlışlıqların nəticəsidir [2].

Azərbaycan müstəqillik əldə etdikdən sonra bu sahədə vəziyyət xeyli dəyişdi. Müstəqil dövlət kimi beynəlxalq aləmə fəal şəkildə inteqrasiya olunan Azərbaycanda rus dilinin dominantlığı azaldı. Eyni zamanda, müxtəlif sahələr üzrə beynəlxalq əlaqələrin getdikcə güclənməsi dilimizdə əcnəbi xüsusi adların funksionallığını artırdı. Həm də xüsusi adların dilimizə artıq rus dili vasitəsilə deyil, orijinal dillərdən keçidi prosesi genişləndi. Cəmiyyətdə işgüzar dil kimi ingilis dilinin mövqeyinin güclənməsi, milli informasiya məkanının qlobal informasiya mühitinə açıqlığı, ədəbi-bədii mətnlərin, işgüzar sənədlərin və s. dilimizə birbaşa mənbədən çevrilməsi, həmçinin beynəlxalq sosial-iqtisadi, siyasi-mədəni əlaqələrin subyekti rolunda çıxış edən soydaşlarımızın, coğrafi ərazi vahidlərimizin, milli əmtəə nişanlarımızın timsalında Azərbaycan onomastik vahidlərinin qlobal çevrədə çox geniş funksionallıq qazanması adların qarşılıqlı transliterasiyası məsələsini daha da aktuallaşdırmış oldu.

Nəticə etibarilə, hazırda yalnız rus adlarının bu dildəki tələffüz formasında latın qrafikasının hərfləri ilə yazılmasına ehtiyac qalmışdır. Digər qrafikalı dillərə aid onomastik vahidlərin, eləcə də yer adlarının, ad və soyadların yazılışında isə rus dilinin norma kimi nəzərə alınmasına heç bir zərurət yoxdur. Bu onomastik vahidləri latın qrafikalı Azərbaycan

əlifbasında olmayan müəyyən hərfləri müvafiq olanları ilə əvəzləməklə, olduğu kimi, orijinal, rəsmi yazılış formasında yaza bilərik. Eyni zamanda, Azərbaycan onomastik vahidləri də xarici dillərdə birbaşa ana dilimizdən transliterasiya edilməlidir. Amma bu problemin həlli istiqamətində atılan addımlar, əfsuslar olsun ki, hələlik onun aktuallığına adekvat deyildir. Latın qrafikasının qəbulundan 24 ildən çox keçməsinə baxmayaraq indiyə qədər kiril əlifbasının mental dairəsindən və yazı qaydalarından tam ayrıla bilməmişik. Bu illər ərzində nə elmi-nəzəri mənbələrdə, nə dillə bağlı yaradılmış komissiyalarda xüsusi adların düzgün təqdim edilməsi ilə bağlı ciddi bir tədbir görülməmiş, müvafiq nəzarət mexanizmləri işlənib hazırlanmamışdır. Ona görə də tərcümə mətnlərində linqvistik baxımdan həm transliterasiya, həm də transkripsiya qaydaları kobud şəkildə pozulur, ədəbi dil normalarına xələl gəlir. Bu problem nəinki bədii tərcümə mətnlərini, habelə tarixi, elmi, siyasi, iqtisadi mətnləri də əhatə edir. Nəticədə xüsusi adların daşıdığı informativ məna yükü oxucuya yanlış çatdırılır. Bununla da həm ədəbi dil təhrif olunur, həm də orijinalda ifadə edilmiş milli kolorit pozulur, ya da zəif nəzərə çarpır.

Aparılan müşahidələr son onilliklərdə həm xarici mətbuatda, həm də yerli kütləvi informasiya vasitələrində dərc olunan materiallarda, rəsmi sənədlərdə, çapdan çıxan kitablarda, efir dilində onomastik vahidlərdən istifadə ilə bağlı qüsurların nəinki azaldığını, əksinə, artdığını göstərir.

Hazırda Azərbaycan mediasında əcnəbi dillərdən alınma bir sıra terminlər kimi müxtəlif onomastik vahidlər də gah ingilis dilindəki kimi, gah da Azərbaycan dilinə transliterasiya variantında yazılır. Məsələn, *Human Rights Watch* təşkilatının adına indi yerli mediada həm ingiliscə necə varsa, həm də Azərbaycan dilində tələffüz olunduğu formada (*Hyuman Rayts Votç*) rast gəlmək mümkündür. Eyni sözü *Freedom House* təşkilatı, *The Washington Post*, *The New York Times* qəzetləri, BBC telekanalı haqqında da demək olar. Bu adlar Azərbaycan kütləvi informasiya vasitələrinin bir qismində ingiliscə, digər bir qismində isə *Fridom Haus, Vaşinqton Post, Nyu York Tayms, Bi Bi Si* kimi yazılır. Bu cür variantlılıq hadisəsi, adı KİV-də tez-tez hallanan yerli *Space* telekanalına da aiddir. Bu televiziya kanalının adı bəzən *Speys* kimi yazılır.

Türkiyə Respublikasının Prezidenti Rəcəb Tayyib Ərdoğanın ad və soyadının Azərbaycan kütləvi informasiya vasitələrində ən azı on beş formada yazılması faktları da bu qeyri-normal vəziyyətə bir nümunədir: *Rəcəb Tayyib Ərdoğan (az.apa.az, anspress.com, deyerler.org, ikisahil.com, azadliq.org); Rəcəb Tayyib Erdoğan (azadlıqradiosu.az); Rəcəb Tayyib Ərdoğan (azadliq.info – “Azadlıq” qəzeti); Rəcəb Tayib Ərdoğan (deyerler.org, az.apa.az, bizimyol.az, azadliq.org); Rəcəb Tayip Ərdoğan (deyerler.org); Rəcəb Təyyub Ərdoğan (azadliq.org, qaynar.info); Rəcəb Təyyib Ərdoğan (azadliq.org); Rəcəb Teyyub Ərdoğan (mediaforum.az, azadliqradiosu.az, musavat.com, deyerler.org); Rəcəb Tayyip Ərdoğan (az.apa.az, azadlıqradiosu.az); Rəcəb Tayyip Erdoğan (azhumanrights.az, icmal.az); Rəcəp Tayyip Ərdoğan (azadlıq.info–“Azadlıq” qəzeti, qafqazinfo.az, merkez-az.com, xeber.azeri.net); Receb Tayyib Ərdoğan (xazar.tv); Recep Tayyib Erdoğan (bbc.co.uk/azeri); Recep Tayyip Erdogan (azadliqradiosu.az); Recep Tayyip Ərdoğan (qol.az)* və s. Qeyd edək ki, Avropa və Amerika mətbuatında bu ad yalnız əlifbalarında olmayan “ğ” hərfi əlifbalarında olan “g” hərfi ilə əvəz edilməklə *Recep Tayyip Erdogan* kimi yazılır [5].

Bu gün Azərbaycan onomastik vahidlərinin xarici dillərə transliterasiyasında da ciddi problemlər özünü göstərməkdədir. Məsələn, uzun illərdən bəri beynəlxalq miqyasda ən çox istifadə edilən Azərbaycan toponimlərindən biri - Qarabağ coğrafi adı ingilis dilində *Karabakh* kimi transliterasiya edilir. Bəzi mənbələrdə bu toponim *Naqorniy Karabax*, digərlərində isə *Upper Karabakh* kimi təqdim olunur. Bü cür təqdimat qüsurlu və yolverilməzdir. Gəncə toponiminin xarici dillərə gah *Genje*, gah *Ganja*, gah da *Genche*, Naxçıvanın gah *Nakhichevan*, gah da *Nakhchivan* kimi transliterasiya edilməsi də xarici informasiya mühitində xaotiklik yaradır və çaşqınlığa səbəb olur.

Müharibə şəraitində yaşayan, informasiya müharibəsinin subyekti kimi çıxış edən, öz milli həqiqətlərini dünyaya çatdırmağa can atan ölkənin dövlət dilinə aid onomastik vahidlərin, o cümlədən yer və şəxs adlarının xarici dillərdə yanlış anlaşılması tarixilik baxımından çox arzuolunmazdır.

Transliterasiya problemi hazırda dünyanın əsas məlumat mənbələrindən olan Vikipediya açıq ensiklopediyasının Azərbaycan bölməsində ana dilimizdə məqalələrin yerləşdirilməsi zamanı da müəyyən narahatlıqlar doğurur. Bir çox Vikipediya redaktorları latın qrafikasında yaradılmış digər Vikipediya bölmələrində olduğu kimi Azərbaycan bölməsində də əcnəbi adların olduğu kimi yazılmasına üstünlük verirlər. Lakin dilimizdə adların transliterasiyası ilə bağlı müəyyən qaydaların olduğunu və indiyə qədər hazırlanmış bütün məlumat kitablarında, ensiklopediyalarda bunun nəzərə alındığını əsas gətirən bir çoxları, eyni qaydalara Vikipediya mühitində də riayət edilməsini vacib sayırlar. Bu ikili yanaşma isə qarışıqlıqlara gətirir. Eyni zamanda, bu problem internet axtarış sistemlərində məlumatların əlçatanlığına da müəyyən mənada mənfi təsir göstərir.

Onomastik vahidlər özündə yerin, yurdun, məmləkətin, xalqın tarixini, milli-mənəvi dünyasını əks etdirir. Bu adlarda yol verilən yanlışlıqlar təbii olaraq, virtual məkanda milli sərvətlərin identifikliyinə böyük zərbə vurur, gələcək problemlər üçün zəmin hazırlayır. Ona görə də onomastik vahidlər düzgün transliterasiya edilməlidir ki, gələcəkdə hansısa qüvvələr bu faktlardan xalqımıza qarşı sui-istifadə hallarına yol verə bilməsinlər. Bu cür adlara müəyyən əlavələrin edilməsinə, yaxud onların tərcüməsinə əsla yol vermək olmaz. Belə adlar sadəcə transliterasiya ilə təqdim olunmalı, hətta yazılışında transkripsiyası da düzgün verilməlidir.

III. TƏKLİFLƏR

Bu problemlərin əsas yaranma səbəblərindən biri budur ki, onomastik vahidlərin Azərbaycan dilindən başqa əcnəbi dillərə, eləcə də əcnəbi dillərdən ana dilimizə

transliterasiyası ilə bağlı normalar hələ tam sabitləşməmiş, bu sahədə ziddiyyətlər ortadan götürülməmişdir. Ona görə də dilin yazı qaydaları ilə bağlı üç prinsipi - fonetik, morfoloji və tarixi-ənənəvi prinsipləri nəzərə alaraq transliterasiya normalarının tam sabitləşməsi zəruridir. Dilçilərimiz tərəfindən əlifbalar və qrafikalararası iki istiqamətli keçidləri özündə əks etdirən konvertasiya cədvəllərinin işlənməsi təmin olunmalıdır. Bu cür cədvəllərin müvafiq standartlar kimi milli və beynəlxalq standartlaşdırma qurumları tərəfindən təsdiqlənməsi zərurəti nəzərə alınmalıdır. Konvertasiya cədvəlləri hazırlanaraq təsdiqləndikdən sonra informasiya texnologiyaları mütəxəssisləri həmin cədvəllər əsasında xüsusi proqram təminatı vasitələri işləməlidirlər. Təkcə ötən əsrdə ölkəmizdə 3 qrafikadan - ərəb, kiril və latın qrafikalarından istifadə olunması indiyə qədər dərc olunmuş informasiya resurslarının - kitab, jurnal və digər yazı materiallarının təşkil etdiyi milli informasiya məkanımızı da 3 hissəyə bölmüşdür. Bunun hansı problemlər yaratdığı isə hamıya məlumdur. Həm bütün dünya azərbaycanlıları arasında qrafika problemini aradan qaldırmaq, həm də əlifbanın bir neçə dəfə dəyişməsi ilə Azərbaycanda nəsillər arasında qırılmış yazılı rabitəni yenidən bərpa etmək, ümumi bir informasiya mühiti yaratmaq üçün müxtəlif dillərin əlifbasında olan hərflərin qarşılıqlı çevrilməsini - konvertasiyasını həyata keçirən xüsusi proqram təminatı vasitələrinin hazırlanması bu baxımdan da son dərəcə vacibdir. Belə proqram təminatı vasitələri hazırlandıqdan sonra bütün Azərbaycandilli elektron resurslar - veb-saytlar, elektron dərsliklər, elektron qəzet və jurnallar və s. həmin konvertasiya cədvəlləri ilə təmin olunmalıdır. Bunun da nəticəsində dünyanın istənilən ölkəsində yaşayan hər bir azərbaycanlı kompüterin ekranında yerləşdirilmiş cədvəldə bir düyməni basmaqla istədiyi qrafikanı seçib lazımi məlumatı oxuyacaq, eyni zamanda, digər qrafikadan istifadə edən azərbaycanlılarla heç bir maneə olmadan elektron yazışmalar aparacaqdır. Transliterasiya problemi həll olunacağı halda, İnternetdə Azərbaycana məxsus tarixi, coğrafi, musiqi, folklor və s. incilərin adları da müəyyən bir standart əsasında başqa dillərin qrafikalarında yazılar, virtual məkanda axtarış vaxtı bu adlarla bağlı çoxçeşidli adlar meydana gəlməz [13, s. 250].

Azərbaycan Milli Elmlər Akademiyasının İnformasiya Texnologiyaları İnstitutunun 2010-cu ildə hazırladığı “Adlar və soyadlar” dövlət informasiya sisteminin yaradılması konsepsiyası”nda bu cür proqram təminatlarının tətbiq mexanizmləri aşağıdakı nümunələr əsasında ümumi şəkildə təsvir edilmişdir [1]:

CƏDVƏL I

HACIBƏYOV	HAJIBEYOV	GÖYÇAY	ГЕОКЧАЙ
	HADJIBEYOV		
	HADJIBEKOV		ГЕКЧАЙ
	QADJIBEKOV		
	XADJIBEKOV		ГОКЧАЙ

CƏDVƏL II

GƏNCƏ	GANJA	NAXÇIVAN	NAXCHIVAN
	GANDJA		NAKHCHIVAN
	GENJA		NAXİCHEVAN
	GENDJA		NAHCHIVAN

Adlar və soyadlar” Dövlət İnformasiya Sisteminin reallaşdırılması istiqamətində görülən işlərin sürətləndirilməsi, bu sistemdə onomastik vahidlərin transliterasiyası ilə bağlı modulların təkmilləşdirilməsi də problemin həllinə həlledici kömək göstərə bilər.

Onomastik vahidlərin yazılış qaydaları ilə bağlı müvafiq kitabların nəşri, bu sahədə maarifləndirmə işlərinin aparılması, kütləvi informasiya vasitələrinin, ədəbiyyatın bu sahədə normayaratma funksiyası yerinə yetirdiyini nəzərə alaraq jurnalistlər, tərcüməçilər üçün müvafiq treninqlərin təşkili, mediada mütəmadi monitorinqlərin həyata keçirilməsi, xüsusi adların yazılış qaydaları ilə bağlı elmi-tədqiqat işlərinin gücləndirilməsi də bu prosesə böyük fayda verərdi.

Həmçinin latın qrafikalı yazı prinsipləri əsasında yeni orfoqrafiya, habelə onomastik lüğətlərin yaradılmasına və bir çox məsələlərin yeni prinsiplər nəzərə alınmaqla dəqiqləşdirilməsinə də böyük ehtiyac vardır. Bu lüğətdə Azərbaycan onomastik vahidlərinin dünyanın aparıcı xarici dillərində, habelə xarici dillərə aid xüsusi adların da Azərbaycan dilində düzgün yazılış formaları öz əksini tapa bilər. Ümumilikdə, bu cür aktual mövzuların Azərbaycanda dilçilik qurumları, universitetlər, filoloqlar, tərcüməçilər tərəfindən forumlarda, seminarlarda müzakirəyə çıxarılması həm dilimizin qloballaşma şəraitində inkişafı baxımından son dərəcə faydalı olar, həm də Azərbaycan dilinin digər dillərlə təmasında vasitəçi dillərin təyinediciliyini ortadan qaldırmağa təkan verərdi.

ƏDƏBİYYAT

[1] “Adlar və soyadlar” dövlət informasiya sisteminin yaradılması konsepsiyası”. Ekspress-İnformasiya İnformasiya Cəmiyyəti Seriyası, “İnformasiya Texnologiyaları” nəşriyyatı, Bakı, 2010

[2] Xəlilli Ə. Azərbaycan dili və xarici dillər arasında transliterasiya problemi // http://hispaniya.com/arsivler/357

[3] “Milliyyət” qəzeti (Türkiyə), 24 noyabr 1982, s.21.Allahverdi F. Transliterasiya problemi // http://fallahverdi.blogspot.com/2011/07/transliterasiya-problemi_26.html

[4] Azerbaijanian Prose; an Anthology Ibragimov Mirza Editor Published by Progress Books, Moscow (1977)

[5] Əliyeva N. Hansı əsasla, hansı məntiqlə? // http://www.azadliq.org/content/article/24306743.html

[6] Məmmədov N.Ş. Axundov A.A. Dilçiliyə giriş. Bakı, “Maarif”, 1966

[7] http://www.mediaforum.az/az/print/2006/05/02/ALINMA-S%C3%96ZL%C6%8FR%C4%B0-NEC%C6%8F-YAZAQ-YAZILDI%C4%9EI-YOXSA-OXUND… 2/2

[8] Allahverdi F. Transliterasiya problemi // http://fallahverdi.blogspot.com/2011/07/transliterasiya-problemi_26.html

[9] Mirzoyev E. Siz latınca yaza bilirsiniz? // http://elmirmirzoev.wordpress.com/2010/05/20/siz-latinca-yaza-bilirsiniz

[10] Л.В.Щерба. Транслитерация латинскими буквами русских фамилий и географических названий // Известия Академии Наук Союза ССР, Отделения литература и языка, 1940, № 3, с.118 - 119

[11] TDK (2012) Yazım Kılavuzu, 27. Baskı, Ankara: TDK Yayınları.

[12] Yakup Yılmaz. Azerbaycan ve Türkiye türkçesinde yabançı özel adların yazımı meselesi // C:/Users/Rauf/Downloads/5000105428-5000151195-1-PB.pdf

[13] Cəfərov Y.M. Azərbaycan dilinin onomastikası müstəqillik dövründə. Bakı, Elm-təhsil nəşriyyatı, 2015

Ovuc İzləri Əsasında İnsanın İdentifikasiya Olunmasında Big Data-nın Rolu

Günel Aslanova
İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
gunel_aslanova90@mail.ru

***Xülasə* — Bu işdə ovuc izləri əsasında insanın identifikasiya olunmasında Big Data-nın rolundan bəhs olunur. Dünyada terror təhlükəsinin artması ilə əlaqədar olaraq təhlükəsizliyi təmin edən biometrik identifikasiya sistemlərindən istifadə olunur. Bu identifikasiya sistemləri ilə işləyərkən daha çox Big Data texnologiyalarından istifadə etmək zərurəti meydana çıxır.**

***Açar sözlər* — daktiloskopiya, dermatoqlifika, irq, big data, ovuc izi, biometriya**

I. GİRİŞ

Biometrik texnologiyaların tətbiqi dövlət sərhədlərindən keçən şəxslərin etibarlı yoxlanılmasını, pasport-viza rejiminə və şəxsiyyəti təsdiq edən digər sənədlərə nəzarəti gücləndirir. Burada tanınma sistemləri xüsusi əhəmiyyət kəsb edir. Bu sistemlərdən kriminalistika ilə yanaşı, digər sahələrdə də geniş istifadə olunur. ABŞ-da baş vermiş 11 sentyabr 2001-ci il terror hadisəsindən sonra biometrik texnologiyaların inkişafına diqqət daha da artmağa başlamışdır. Dünyanın hər yerində biometrik pasportların yaranmasına ehtiyac duyulmuşdur. Biometrik identifikasiya vasitələrindən biri də ovuc izləri əsasında insanın irqi-etnik mənsubiyyətinin müəyyən olunmasıdır.

Son illərdə biometrik texnologiyalar sahəsində tədqiqat işlərinin sayı bir neçə dəfə artıb, bu isə problemin aktuallığından xəbər verir. Aktuallığın əsas səbəbləri bunlardır:

- ✓ Fərdi məlumatların mühafizə edilməsi, terrorçuluğa qarşı mübarizə;
- ✓ Elektron hökümət, elektron təhsil, elektron tibb və s. sahədə informasiya və komunikasiya vasitələrinin inkişafının yeni sistemlərin yaradılmasını zəruri etməsi;
- ✓ “Azərbaycan Respublikasının biometrik eyniləşdirmə sisteminin yaradılması üzrə” Dövlət proqramının qəbul edilməsi və s.

Məlumdur ki, insanın ovuc izləri əsasında tanınma sistemləri mövcudur. Lakin bunun üçün çoxlu sayda verilənlər analiz olunmalıdır. Əgər biz şəxsin hansı irqə mənsub olduğunu müəyyən etsək bu zaman həmin irqə məxsus verilənlər arasından identifikasiya aparmaq olar və nisbətən daha az sayda mövcud verilənlər əsasında şəxsin müəyyən edilməsi mümkündür. Bu isə vaxt baxımından qənaətə səbəb olacaqdır.

II. İNSANIN OVUC İZLƏRİNƏ GÖRƏ İDENTİFİKASİYA EDİLMƏSİ

Big Data texnologiyalarının geniş istifadə sahələrdən biri də insanın ovuc naxışları əsasında identifikasiya olunmasıdır. Bu sahədə bəzi terminləri qeyd edək.

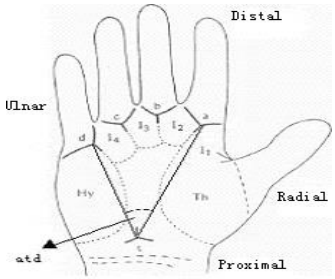
Daktiloskopiya–dərinin təsvirinin təkrarolunmazlığına əsaslanan barmaq izləri əsasında insanın tanınması üsuludur [1]. Dermatoqlifika–insanın ovuc içi və pəncələrinin dərisinin naxışlarının əlamətini öyrənən elmdir [2].

İnsanın ovuc içi izləri bir neçə məqsəd üçün istifadə oluna bilər:

- Şəxsi idetifikasiya;
- Diaqnozun qoyulmasında;
- Qanla əlaqəli identifikasiyalarda;
- İrqi etnik mənsubiyyətin müəyyən olunması.

Ovuc izlərinin tanınmasında istifadə olunan bəzi üsullar verilmişdir [3]:

1. Kəmiyyət analizi (Quantitative analysis):
   1.1. Naxış sıxlığı (Pattern intensity)
   1.2. İzlərin sayı (Ridge counts)
   1.3. “atd” bucağı (atd angle)
   1.4. “Enin” nisbəti (Breadth ratio)
2. Topologiya .



Şəkil1. Əlin naxış sahəsinin təsviri

Ovuc naxışlarından insanın irqi və etnik mənsubiyyətinin müəyyən olunmasında da istifadə etmək mümkündür. İnsanın irqi və etnik mənsubiyyətinin müəyyən olunması sahəsindəki problemlərlə ictimai elmlərdən Antropologiya, Etnoqrafiya və İrqşünaslıq sahəsində çalışan mütəxəssislər, texnika elmləri sahəsində isə Biometrik texnologiyalar, Tibb, Surətlərin tanınması sahəsindəki mütəxəssislər tədqiqat aparır. İrqlərin

ətraf mühitə uyğunlaşması nəticəsində elə xarici fiziki əlamətlər kompleksi yaranmışdır ki, bu əlamətlərə görə dünya xalqalarına əvvəllər “ağ”, “qara” və “sarı” rənglərə və üç əsas irqə ayırırdılar: avropoid, neqroid və monqoloid. Son vaxtlarda monqoloid irqindən amerikanoid, ekvatorial irqindən avstraloid irqinin ayrılması tövsiyyə olunur [4]. Ümumiyyətlə insanın irqi və etnik mənsubiyyətini aşağıdakı əlamətlərə görə müəyyən etmək olar:

- Kəllə sümüklərinə görə
- Gözə görə
- Barmaq və ovuc naxışlarına görə
- Dişlərə görə
- Sifətin təsvirləri əsasında və s.

Naxışlar insan həyatı boyu daimi və dəyişməzdir. Əslində, naxışlar dölün ilkin formalaşması mərhələsində inkişaf edir və dərinin son parçalanmasına qədər dəyişməz olaraq qalır. Lakin, bəzi xarici, eləcə də təsadüfi səbəblərdən dəri köhnələ və ya zədələnə bilər. Dəri naxışlarının öyrənilməsinin müxtəlif tətbiq sahələri var. Daun sindromu, autizm kimi genetik xəstəliklərin identifikasiyası və mübahisəli atalıq məsələlərinin həllində, o cümlədən irqi-etnik mənsubiyyətin müəyyən edilməsində dəri naxışların böyük əhəmiyyəti var. Tətbiq sahəsi geniş olduğuna görə dəri naxışlarının oxunmasında yeni və asan üsulların yaradılmasına ehtiyac vardır [5].

III. BİG DATA və OVUC İZİ

Böyük miqyaslı biometrik sistemlər ümummilli şəxsiyyət vəsiqəsi və güvənli mobil ödəniş üsullarının əsasıdır. Digər Big Data sistemləri ilə kimi, biometrik sistemlər də "dörd V" problemləri ilə qarşılaşır. Şəxsi məlumatların toplanması zamanı həcm (volume), tranzaksiyaya (əməliyyata) cavab vaxtı (sürət–velocity), əməliyyatın aparılması zamanı fərqlərin yaranması (velocity–çeşidlilik), müxtəlif biometrik identifikatorların olması (variety).

Bir sıra müxtəlif və real biometrik tətbiqlərdə olan bu problemlərin həlli üçün aşağıdakı üsullar vardır.

1) Biometrik verilənlər bazasının miqyasından (həcmindən) asılı olmadan yaxın sabit zaman ərzində girişə imkan verən (sürət) indeksləmə metodları
2) Risk və dəqiqlik məsələlərinin həlli üçün çoxsaylı (müxtəliflik) biometrikaların birləşdirilməsi
3) Çoxsaylı təkrar qeydiyyatın qarşısını almaqla və biometrik identifikatorların oğurluğundan qorunmaqla biometrik verilənlər bazalarının tamlığının təmin edilməsi

Bir çox ümumilikləri universal Big Data sistem-dizayn məsələləri ilə paylaşdıqda, biometrik sistemlər həm də bu məsələlərin necə yaranmasını və unikal, predmetyönümlü yolla həll edilməsini özünə daxil edən zəngin tematik tədqiqatları təmin edir [6]. Mürəkkəb informasiya sistemləri və verilənlərin analitik təhlilinə artan tələbat nəticəsində *big data* termini son illərdə geniş istifadə olmaqdadır. “Big data” termini isə ilk dəfə 1998-ci ildə Silicon Graphics kompaniyasının kompüter elmləri üzrə mütəxəssisi Con Meşi tərəfindən istifadə edilmişdir [7]. “Mövcud texnologiyaların, metodların və alqoritmlərin köməyilə toplanılması, saxlanılması, analizi və vizuallaşdırılması çətin və ya mümkün olmayan verilənlər çoxluğuna “Big data” deyilir”. Biometrik texnologiyalardan istifadə edərkən çoxlu sayda verilənlərin emalı qaçınılmazdır. Biometrik məlumatlar biometrik sistemdə mərhələlərlə emal olunur. Beləki, sistemə biometrik giriş olunur və verilənlər bazasındakı şablonlar təqdim olunan nümunə ilə bir-bir yoxlanılır. Prosesin sonunda biometrik sistem daxil edilən biometrik nümunə ilə yüksək oxşarlığı olan identifikatorların siyahısını verir. İdentifikasiya mənfi və ya müsbət ola bilər. Verilənlərin sayı çox olduqda daha çox müqayisə əməliyyatı aparılır və bu prosesə daha çox zaman sərf olunur. Bu zaman ənənəvi sistemlər artıq kifayət etmir və böyük verilənlərə müraciət etmək zərurəti yaranır. Bu baxımdan böyük verilənlərin xidməti böyükdür. Verilənlər çox olduqca onlarla işləmək və təhlil etmək daha da çətin olur. Az sayda verilənlər üçün mövcud olan riyazi modellər artıq böyük verilənlər üçün işləmir. Big data-nın bir sıra üstünlükləri var ki, bu üstünlükləri identifikasiya apararkən müşahidə etmək olar.

Big datanın üstünlükləri bunlardır:

- ✓ Məlumat nə qədər çoxdursa tədqiqatçının səhv qərar qəbul etmək ehtimalı da o qədər azdır;
- ✓ Böyük həcmdə verilənlər olmadıqda internet istifadəçilərinin fəaliyyətini dəqiq tədqiq etmək praktiki cəhətdən mümkün deyil;
- ✓ Verilənlərin saxlanılması daha ucuz və əlçatandır, belə ki, böyük həcmdə verilənlərin yadda saxlanması və analiz olunması yanlış identifikasiyanın aparılmasından daha səmərəlidir [8].

Bu üstünlükləri nəzərə alaraq ovuc izləri əsasında insanın identifikasiya olunması məsələsində də Big Data-dan istifadə olunması daha səmərəlidir.

NƏTİCƏ

Təqdim olunan işdə biometrik texnologiyaların tətbiq sahələri, ovuc izləri əsasında insanın identifikasiya olunması, ovuc izləri əsasında insanın identifikasiya olunması məsələsinə, həmçinin burada istifadə olunan biometrik sistemlərdə Big Data-nın rolu və üstünlüklərindən bəhs olunmuşdur.

ƏDƏBİYYAT

[1] https://ru.wikipedia.org/wiki/Дактилоскопия
[2] https://ru.wikipedia.org/wiki/Дерматоглифика
[3] Y.Zhou, Y.Zeng, Lizhen and Wenlei Hu, Application and development of palm print research Biomedical Engineering Center, Beijing Polytechnic University, P.R.China, p. 383–390, 2001.
[4] https://az.wikipedia.org/wiki/İrq
[5] Г.Л. Хить, И.Г. Широбоков, И.А. Славолюбова, дерматоглифика в антропологии российская академия наук институт этнологии и антропологии им. н.н. миклухо-маклая, Санкт-Петербург, 2013.
[6] N.K. Ratha, J.H. Connell; S.Pankanti, "Big Data approach to biometric-based identity analytics," in IBM Journal of Research and Development , vol.59, no.2/3, pp.4:1-4:11, 2015.
[7] F. Diebold, On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.
[8] http://habrahabr.ru/company/io/blog/264125/

Azərbaycan Milli Elmlər Akademiyası
İNFORMASİYA TEXNOLOGİYALARI İNSTİTUTU

“İnformasiya Texnologiyaları” Nəşriyyatı

Az1141,Bakı şəh., B.Vahabzadə küç., 9A
Tel.: (+99412) 539 01 67 Faks: (+99412) 539 61 21
secretary@iit.ab.az,

www.ikt.az