Böyük həcmli fərdi məlumatların analizi üçün iterativ çəkili k-means alqoritmi

Ramiz Alıquliyev[1], Şəlalə Tahirzadə[2]
[1,2]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1] r.aliguliyev@gmail.com, [2] fmv.shalala@gmail.com

***Xülasə*—İnformasiya Kommunikasiya Texnologiyalarının (İKT) inkişafı rəqəmsal informasiyanın sıçrayışlı artımına və nəticədə, “big data” konsepsiyasının populyarlaşmasına səbəb olmuşdur. Bu səbəbdən big data və onun mahiyyətini, müxtəlif mənbələrini, analiz texnologiyalarının imkanlarını, problemlərini, təhlükəsizlik məsələlərini hərtərəfli tədqiq etməyə ehtiyac yaranmışdır. Məqalədə mövcud analiz texnologiyalarının çatışmazlıqlarına, fərdi məlumatların analizindəki təhlükəsizlik məsələlərinə baxılır və həlləri araşdırılır. Böyük məlumatların analizindəki çatışmazlıqları aradan qaldırmaq üçün iterativ və çəkili iterativ k-means alqoritmləri təklif edilir.**

Açar sözlər—big data, fərdi məlumatlar, fərdi məlumatların təhlükəsizliyi, klasterləşdirmə, iterativ və çəkili klasterləşdirmə.

I. GİRİŞ

İKT sahəsindəki yüksək inkişaf meylləri yeni texnoloji həllərin geniş şəkildə istifadəsinə, cəmiyyətin kompüterləşməsinə və s. gətirib çıxarmışdır. Verilənlərin toplanması və saxlanması vasitələrinin də sürətli inkişafı ilə hazırda gündəlik olaraq böyük həcmli verilənlər toplanılır və kompüter şəbəkələrinə, internetə, müxtəlif verilənlər saxlancına axın edir. Bu isə böyük həcmli verilənlər anlayışının yaranmasına səbəb olmuşdur. Böyük həcmli verilənlərin analizi vasitəsilə faydalı məlumatlar – bilik əldə etmək mümkündür.

Böyük həcmli verilənlər müxtəlif məzmunlu mənbələrdən toplanıla bilər. Tibbi, hərbi, mühəndislik, məişət, müxtəlif elm sahələri və s. buna misal ola bilər. Belə müxtəlif məzmunlu verilənlərin təhlükəsizliyinə qoyulan tələblər də müxtəlifdir. Məlumdur ki, analiz texnologiyalarının tətbiqi nəticəsində böyük verilənlərdən əlavə bilik və qiymətli məlumatlar əldə etmək mümkündür. Bu da öz növbəsində verilənlərin məzmunundan asılı olaraq təhlükəsizlik baxımından müxtəlif problemlərə gətirib çıxara bilər. Misal olaraq qeyd etmək olar ki, hərbi və ya fərdi məlumatların təhlükəsizliyi ilə məişət verilənlərinin təhlükəsizlik tələbləri eyni deyildir. Belə ki, hərbi və ya fərdi məlumatların analizi vasitəsilə dövlət əhəmiyyətli biliklər əldə etmək mümkündür.

Aydındır ki, verilənlərin həcminin sürətli artımı mövcud analiz texnologiyalarının imkanlarını məhdudlaşdırır. Bu məqsədlə məqalədə Big data texnologiyalarında fərdi məlumatların təhlükəsizlik məsələləri araşdırılır. Mövcud analiz texnologiyalarının imkanları, çətinlikləri nəzərdən keçirilir. Təklif edilən iterativ və iterativ çəkili k-means alqoritmlərinin tətbiqi ilə eksperiment həyata keçirilir. Eksperimentin nəticələrinə əsasən k-means alqoritmi ilə təklif edilən alqoritmlər müqayisə edilir, müxtəlif qiymətləndirmə indeksləri vasitəsilə nəticələr qiymətləndirilir.

II. “BIG DATA” KONSEPSİYASI

Böyük verilənlər həcminin hazırki eksponensial artım sürəti də məhz informasiyaya əlyetərliyin asanlaşması və kommunikasiya vasitələrinin çoxalması ilə sıx əlaqədardır [8]. Böyük verilənlər dedikdə, çox böyük həcmə malik, sürətli və emal üçün çox mürəkkəb verilənlər başa düşülür.

Böyük verilənlərin “3V” adlandırılan xarakteristikası olan həcm (*ing. volume*) – verilənlərin həcmi; müxtəliflik (*ing.variety*) – verilənlərin müxtəlifliyi və bir çox hallarda kifayət qədər strukturlaşdırılmamış; sürət (*ing.* velocity) – verilənlərin böyük sürətlə emal edilməsi olaraq qeyd edilir:

“Google Trends”-in hesablamalarına google axtarış mühərriki vasitəsilə saniyədə 40000, günlük 3.46 milyon, illik olaraq isə 1.2 trilyon axtarış həyata keçirilir [16]. Dünyada ikinci ən çox ziyarət edilən və nəhəng şirkətlərdən olan “Youtube”-a dəqiqədə 500 saat həcmində videolar yüklənir [17]. Bütün bunlar verilənlərin həcminin fasiləsiz olaraq sürətli artımına gətirib çıxarır. Belə böyük həcmli verilənlərin isə analizinə ehtiyac yaranır. Müxtəlif emal texnologiyaları vasitəsilə isə onları analiz etmək, qiymətli biliklərə çevirmək mümkündür. Böyük verilənlər konsepsiyanın geniş yayılması özü ilə birgə təhlükəsizlik problemlərini də gətirir [1]. Bu isə informasiya mühafizəsi məsələlərinin aktuallığının artmasına gətirib çıxarır. İnformasiyanın mühafizəsi informasiya təhlükəsizliyinin təmin olunmasına yönəlmiş tədbirlər kompleksidir.

III. İNFORMASİYA TƏHLÜKƏSİZLİYİ

Məlumdur ki, İKT sahəsinin nəzərə çarpacaq səviyyədə inkişafı elmi-texniki inqilaba səbəb olmuşdur. Bu da cəmiyyətə təsirsiz qalmamışdır. Nəticədə, elmi ədəbiyyatlarda informasiya cəmiyyəti anlayışından istifadə edilməyə başlanmışdır. İnformasiyanın toplanması, saxlanması, emalı, istifadəsi və ötürülməsi ilə bağlı fəaliyyət sahələri bu cəmiyyətdə aparıcı mövqeyə yüksəlməklə iqtisadiyyatın əsasını təşkil edir. İnformasiya və biliklər isə informasiya cəmiyyətinin ən mühüm resursu və başlıca elementidir [2].

İnformasiya təhlükəsizliyinin təmin olunması probleminin aktuallığını və vacibliyini şərtləndirən səbəblərdən aşağıdakıları qeyd etmək olar:

- şəbəkə texnologiyalarının geniş yayılması və lokal şəbəkələrin qlobal şəbəkələr halında birləşməsi;

DOI: 10.25045/NCInfoSec.2019.02

- informasiya təhlükəsizliyinin pozulmasına praktik olaraq mane olmayan qlobal İnternet şəbəkəsinin inkişafı;
- minimal təhlükəsizlik tələblərinə belə cavab verməyən proqram vasitələrinin geniş yayılması.

A. Fərdi məlumatların təhlükəsizliyi

İnformasiya cəmiyyətinin hər bir üzvü olan fərd bu məlumatların həm istehsalçısı, həm də istehlakçısı hesab edilir. Məzmun baxımından informasiya müxtəliflik təşkil edir. Fərdi məlumatlar dedikdə şəxsin kimliyini birbaşa və ya dolayısı ilə müəyyənləşdirməyə imkan verən istənilən məlumat başa düşülür. Fərdi məlumatlar toplandığı andan mühafizə olunur və hər bir fərd ilkin mənada öz məlumatlarının qoruyucusuna çevrilir. Bu tip məlumatlar daxilolma növünə görə konfidensial və açıq kateqoriyalara bölünür [3]. Açıq məlumatların analizi belə konfidensial informasiyaların əldə olunmasına səbəb ola bilər. Bu səbəbdən böyük məlumatların məxfiliyi əvvəllər fərdi mənada müzakirə edilsə də, tədricən “milli təhlükəsizlik” istiqamətində inkişaf etmiş, nəticədə, siyasi və milli xarakter qazanmışdır. Bu baxımdan informasiya və İKT hər bir ölkənin milli təhlükəsizliyinin təmin olunması üçün mühüm vasitədir. İnformasiya təhlükəsizliyinin təmin edilməsi isə prioritet məsələlərdən biridir. Bunun nəticəsidir ki, hazırda fərdi məlumatların toplanması, emalı, ötürülməsi, mühafizəsi və s. qanunvericilikdə müəyyən edilmiş hüquqi normalar əsasında həyata keçirilir.

IV. BÖYÜK VERİLƏNLƏRİN ANALİZİ TEXNOLOGİYALARI

Böyük verilənlərin analizi nəticəsində daha sürətli və daha optimal qərarlar qəbul etmək, çəkilən xərcləri minimallaşdırmaq, sərf edilən zamana qənaət etmək və yeni növ xidmətlər təklif etmək mümkündür. Verilənlərin analizi üçün müxtəlif vasitə və üsullardan istifadə etmək mümkündür. Bu məqsədlə ən geniş tətbiq edilən üsul klasterləşdirmə alqoritmləridir. Klasterləşdirmə alqoritmlərinə misal olaraq K-means, K-medoids, DBSCAN, E-M və s. qeyd etmək olar. Verilənlərin analizi üçün ən çox tətbiq edilən klasterləşdirmə alqoritmi kimi K-means və onun müxtəlif modifikasiya olunmuş versiyalarını qeyd etmək olar. K-means alqoritminin ən böyük çatışmazlığı onun başlanğıc klaster mərkəzlərinin seçilməsindən asılı olmasıdır. Belə ki, ilkin mərkəzin seçilməsindən asılı olaraq nəticələrin keyfiyyəti də müxtəlif ola bilər [5]. Klasterlərin keyfiyyətinin ilkin mərkəzin seçilməsindən asılılığı, böyük həcmli verilənlərin analizində çətinliklər, emal prosesində texniki təminata qoyulan yüksək tələblər kimi bir sıra məsələləri aradan qaldırmaq məqsədi ilə k-means alqoritminin müxtəlif modifikasiya edilmiş versiyalarına nəzər yetirilir. K-means alqoritminin çatışmazlıqlarını aradan qaldırmaq üçün onun modifikasiyası olan iterativ və çəkili iterativ k-means alqoritmi təklif edilir. Emal prosesində texniki təminata qoyulan tələbləri nəzərə alaraq verilənlərin paketlərə bölünməklə emal edilməsi təklif edilir [4,6,7]. Qeyd edilən alqoritmlərin nəticələri eksperiment aparılmaqla k-means alqoritmi ilə müqayisə edilmişdir. Təklif edilən alqoritmlərin əsas addımları aşağıdakı kimi verilmişdir.

A. İterativ k-means alqoritmi

İterativ k-means alqoritminin addımları aşağıdakı kimidir:

- $\mathbf{P} = \{p_1,..., p_n\}$ verilənlər çoxluğu $\mathbf{B} = \{b_1,..., b_m\}$ paketlərinə bölünür;
- b_1 paketinə *k*-means alqoritmi tətbiq edilir;
  - Mərkəzlər növbəti paketə əlavə edilmək üçün seçilir;
  - Yeni paketə *k*-means alqoritmi tətbiq edilir. Əvvəlki mərhələdə tapılmış klaster mərkəzləri bu mərhələ üçün başlanğıc klaster mərkəzləri kimi qəbul edilir;
  - b_m^* paketinə *k*-means alqoritmi tətbiq edildikdən sonra iterasiya dayandırılır;
  - Yekunda alınan mərkəzlər $\mathbf{P}$ verilənlər çoxluğunun klaster mərkəzləri kimi qəbul edilir;
- $\mathbf{P}$ verilənlər çoxluğunun hər bir obyekti ilə sonuncu klaster mərkəzləri arasındakı Evklid məsafəsi hesablanır;
- $\mathbf{P}$ çoxluğunun obyektləri məsafəcə yaxın olduğu mərkəzə uyğun klasterlərə aid edilir.

B. Çəkili iterativ k-means alqoritmi

Çəkili İterativ k-means alqoritminin addımları aşağıdakı kimidir:

- $\mathbf{P} = \{p_1,..., p_n\}$ verilənlər çoxluğu $\mathbf{B} = \{b_1,..., b_m\}$ paketlərinə bölünür;
- b_1 paketinə *k*-means alqoritmi tətbiq edilir;
  - Tapılmış klasterlərin mərkəzləri üçün (1) düsturunun köməyilə çəkilər hesablanır;
  - Çəkili mərkəzlər növbəti paketə əlavə edilir və alınmış nöqtələr çoxluğu üçün çəkili k-means alqoritmi tətbiq edilir.
  - Yeni paketə çəkili *k*-means alqoritmi tətbiq edilir. Bu mərhələdə klasterlərin başlanğıc mərkəzləri kimi əvvəlki mərhələdə hesablanmış çəkili klaster mərkəzləri qəbul edilir;
  - Bu proses bütün paketlər üçün iterativ olaraq təkrarlanır. Və sonuncu mərhələdə b_m^* paketinə çəkili *k*-means tətbiq edildikdən sonra iterasiya dayandırılır;
- Sonuncu mərhələdə tapılmış mərkəzlər $\mathbf{P}$ verilənlər çoxluğunun klaster mərkəzləri kimi qəbul edilir;
- $\mathbf{P}$ verilənlər çoxluğunun hər bir obyekti ilə sonuncu mərhələdə tapılmış klaster mərkəzləri arasındakı Evklid məsafəsi hesablanır;
- $\mathbf{P}$ çoxluğunun obyektləri məsafəcə yaxın olduğu mərkəzə uyğun klasterlərə aid edilir.

Burada klasterlərin mərkəzlərinin çəkisi aşağıdakı düsturun köməyilə hesablanır:

$$\omega_{ij} = 1 + \frac{|C_{ij}|}{|b_i|},\ i = 1,...,m;\ ,\ j = 1,...,k \qquad (1)$$

$|b_i|$ – b_i paketinin ölçüsü (paketdəki obyektlərin sayı); $|C_{ij}|$ – b_i paketinin klasterləşməsindən sonra alınan C_{ij} klasterinin ölçüsü (C_{ij} klasterindəki obyektlərin sayı); ω_{ij} – C_{ij} klaster mərkəzinin çəkisidir.

V. TƏKLİF EDİLƏN ALQORİTMLƏRİN EKSPERİMENTİ VƏ MÜQAYİSƏLİ ANALİZİ

Təklif edilən alqoritmləri qiymətləndirmək üçün Intel core (i7), 2.20 GHz, 8 Gb RAM xarakteristikalarına malik kompüterdə R proqramlaşdırma dilindən (R-3.5.3) istifadə edilmişdir. Eksperiment üçün bir neçə müxtəlif ölçülü verilənlər üzərində ənənəvi k-means metodu və təklif edilən alqoritmlər tətbiq edilmişdir. Eksperimentin nəticələrinin müqayisəli analizini reallaşdırmaq məqsədilə bir neçə qiymətləndirmə indeksləri seçilmişdir. İstifadə edilmiş verilənlərin təsviri və seçilmiş qiymətləndirmə indeksləri haqqında məlumat aşağıda qeyd edilmişdir.

A. Verilənlər

Eksperimentlər “Poker-hand-testing and Training”, “Localization Data for Person Activity” , “UScensus data 1990” verilənləri üzərində tətbiq edilmişdir və onların xarakteristikası cədvəl 1-də verilmişdir.

CƏDVƏL 1. VERİLƏNLƏR ÇOXLUĞUNUN XARAKTERİSTİKASI

Verilənlər çoxluğu	**Obyektlərin sayı**	**Atributların sayı**	**Siniflərin sayı**
Poker-hand-testing and Training	1025010	11	10
Localization Data for Person Activity	164860	4	11
UScensus data 1990	2458285	68	–

“Poker hand testing and Training” verilənlərindəki hər bir nümunə 52-lik standart dəstədən çəkilmiş 5 oyun kartını ifadə edir [9, 10, 11]. Hər bir kart “suit” və “rank” adlandırılan 2 atributla xarakterizə edilir. 11-ci atribut isə sinifləri əks etdirir.

İstifadə edilmiş “Localization Data for Person Activity” verilənləri müxtəlif hərəkətlər icra edən 5 fərdin məlumatlarından əldə edilmişdir [11,12]. Onların hər birinə eyni ssenarini 5 dəfə yerinə yetirən, sağ və sol ayaq biləyi, bel və sinə nahiyələrinə bərkidilmiş 4 sensor birləşdirilmişdir. Bu sensorlar vasitəsilə 4 atributa malik nümunə toplanmışdır ki, onlardan 4-cüsü sinifləri əks etdirir.

“US census 1990 data” 1990-cı ildə siyahıya alma zamanı əldə edilmiş məlumatlardan toplanmışdır [11,13].

Eksperimentin nəticələrinin qiymətləndirilməsi üçün (2) düsturu ilə hesablanan funksiyanın qiyməti, Purity indeksi (3), Davies-Bouldin indeksi (4) istifadə edilmişdir.

$$F = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - O_j\right\|^2 \to \min \qquad (2)$$

burada O_j – C_j klasterinin mərkəzi, $x_i^{(j)}$ – C_j klasterinə aid edilən i -ci obyektdir.

B. Qiymətləndirmə indeksləri

Purity indeksinin hesablanması üçün aşağıda göstərilmiş (3) düsturundan istifadə edilir:

$$Purity = \frac{1}{n}\sum_{i=1}^{k}\max_j\left|C_i \cap A_j\right|, \qquad (3)$$

burada n – obyektlərin sayı, k – klasterlərin sayı, C_i – i-ci klaster, A_j – j -ci sinifdir. Purity indeksinin qiyməti nə qədər böyük olarsa, alqoritm bir o qədər effektiv hesab edilir [14].

Klasterləşdirmə alqoritmlərinin qiymətləndirilməsi üçün istifadə edilən digər metrika isə Davies-Bouldin (DB) indeksidir [15]. DB indeksinin hesablanması üçün aşağıda göstərilmiş (4) və (5) düsturlarından istifadə edilir:

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_i \qquad (4)$$

$$R_i = \max_{j=1,...k, i\neq j} R_{ij} \text{ və } R_{ij} = \frac{\mathrm{var}(C_i) + \mathrm{var}(C_j)}{\left\|O_i - O_j\right\|}, \qquad (5)$$

burada C_i - i-ci klaster, O_i – C_i klasterinin mərkəzidir. DB indeksi 0 və 1 aralığında qiymətlər alır. Bu indeksin aşağı qiymətləri klasterləşdirmə üçün daha yaxşı hesab edilir.

Cədvəl 2-də təklif edilən iterativ və iterativ çəkili k-means alqoritmlərinin yekun nəticələri k-means alqoritmi ilə müqayisə edilir. Burada əsas məqsəd (2) düsturu ilə hesablanan funksiyanın qiymətini minimallaşdırmaq və nəticələrin dəqiqliyini artırmaqdır.

CƏDVƏL 2. K-MEANS, IK-MEANS, IWK-MEANS ALQORİTMLƏRİNİN EKSPERİMENTAL NƏTİCƏLƏRİ

Localization Data for Person Activity				
Metod	*DB*	*Purity*	*Funksiyanın qiyməti*	*Vaxt (san.)*
k-means	1.0451	0.4560	31,588.52	1.36
ik-means	1.1673	0.4609	35,292.28	3.73
iwk-means	1.1399	0.4410	32,721.03	4.07
Poker-hand-testing and Training				
Metod	*DB*	*Purity*	*Funksiyanın qiyməti*	*Vaxt (san.)*
k-means	1.4441	0.4758	40,653,041	23.49
ik-means	1.4238	0.5012	38,676,726	32.21
iwk-means	1.4497	0.5001	38,694,990	32.66
UScensus data 1990				
Metod	*DB*	*Funksiyanın qiyməti*		*Vaxt (san.)*
k-means	0.5774	615,910,906		69.29
ik-means	0.5774	532,548,026		22.56
iwk-means	0.5774	529,369,871		24.84

Burada əsas məqsəd təklif edilmiş alqoritmlər vasitəsilə daha keyfiyyətli klasterləşdirməni həyata keçirməkdir. Qeyd etmək lazımdır ki, qiymətləndirilmə üçün istifadə edilmiş DB indeksinin, (2) funksiyasının və zamanın minimal qiymətləri,

Purity indeksinin isə maksimal qiyməti daha keyfiyyətli nəticəni ifadə edir. Təklif edilən alqoritmlər üçün ən yaxşı nəticə “US census 1990 data” verilənlərinin nəticələridir. Bu verilənlərinin nəticələrinə görə DB indeksi hər üç alqoritm üçün eyni qiymət alsa da, funksiyanın qiyməti və zaman göstəricilərinə əsasən IK-means və IWK-means alqoritmləri K-means alqoritmini üstələyir.

İterativ və çəkili k-means alqoritmləri vasitəsilə klasterlərin keyfiyyətinin ilkin mərkəzin seçilməsindən asılılığını aradan qaldırmaq mümkündür. Belə ki, k-means alqoritmindən fərqli olaraq təklif edilən alqoritmlərin əsas üstünlüyü odur ki, iterativ prosesdə bir öncəki klasterləşmədən əldə edilmiş mərkəzlər növbəti addım üçün başlanğıc mərkəz kimi seçilir. Təklif edilən alqoritmlərin digər üstünlüyü verilənlərin paketlərə bölünməklə analiz edilməsidir ki, bununla emal prosesində texniki təminata qoyulan yüksək tələbləri minimallaşdırmaq olar. Beləliklə, eksperimentin nəticələrini əks etdirən cədvəl 2-dən də göründüyü kimi verilənlərin ölçüsünün böyük qiymətlərində təklif edilən alqoritmlərin effektivliyi artır.

NƏTİCƏ

Məqalədə böyük verilənlər konsepsiyanın geniş yayılması nəticəsində özü ilə birgə gətirdiyi problemlər, emal texnologiyalarındakı çətinliklər, təhlükəsizlik məsələləri araşdırılmış və risklər təhlil edilmişdir. Eyni zamanda analiz texnologiyalarının tətbiqi ilə böyük verilənlərin təhlükəsizlik məsələlərinə baxılmışdır. Qeyd edilən məsələlərin həllinə nail olmaq üçün iterativ və çəkili iterativ *k*-means alqoritmləri təklif edilmişdir. Təklif edilən alqoritmlərin əsas məqsədi *k*-means klasterləşdirmə alqoritminin ilkin mərkəzdən asılılığını aradan qaldırmaqla nəticələrin etibarlığını və dəqiqliyini artırmaqdır.

Təklif olunan alqoritmlərin tətbiqi ilə müxtəlif verilənlər üzərində eksperimentlər həyata keçirilmişdir. Eksperimentin nəticələri bəzi qiymətləndirmə indekslərinə əsasən müqayisə edilmişdir. Araşdırmalar onu göstərdi ki, bu alqoritmlərin əsas üstünlüyü ənənəvi metodlarla müqayisədə daha sürətli olmasıdır. Eksperimentin nəticələrinə əsasən qeyd etmək olar ki, təklif edilən alqoritmlər çox böyük həcmli verilənlərin analizi üçün daha effektivdir. Beləliklə, verilənlərin analizi zamanı təhlükəsizlik baxımından həssaslığı nəzərə alınmaqla mürəkkəb məsələlərin həlli üçün təklif edilən alqoritmlər tətbiq edilə bilər.

MİNNƏTDARLIQ

Bu iş Azərbaycan Respublikası Dövlət Neft Şirkətinin (SOCAR) Elm Fondunun mailyyə yardımı ilə yerinə yetirilmişdir – **Müqavilə № 03LR – AMEA**.

İSTİNADLAR

[1] R.M. Alguliyev, Y.N. Imamverdiyev, “Big data: big promises for information security”, Proc. 2014 IEEE 8th International Conference on Application of Information and Communication Technologies, pp.1-4, 2014.

[2] R.M. Əliquliyev, M.Ş. Hacırəhimova, A.Ş. Əliyeva, “Big Data-nın aktual elmi-nəzəri problemləri”, İnformasiya cəmiyyəti problemləri, №2, s.37-49, 2016.

[3] R.M. Alguliyev, R.M. Aliguliyev, F.J. Abdullayeva, “Privacy-preserving deep learning algorithm for big personal data analysis”, Journal of Industrial Information Integration, vol.15, pp.1-14, 2019.

[4] R. Alguliyev, R. Aliguliyev, A. Bagirov, R. Karimov, “Batch clustering algorithm for big data sets”, Proc. 2016 IEEE 10th International Conference on Application of Information and Communication Technologies, pp.79-82, 2016.

[5] R.M. Alguliyev, R.M. Aliguliyev, L.V. Sukhostat, “Weighted consensus approach for big data clustering”, Proceedings of the 13th IEEE International Conference Application of Information and Communication Technologies, pp.143-146, 2019.

[6] N. Karmitsa, A.M. Bagirov, S. Taheri, “Clustering in large data sets with the limited memory bundle method”, Pattern Recognition, vol.83, pp.245-259, 2018.

[7] N. Karmitsa, A.M. Bagirov, S. Taheri, “New diagonal bundle method for clustering problems in large data sets”, European Journal of Operational Research, vol.263, no.2, pp.367-379, 2017.

[8] D. Zhang, “Big data security and privacy protection”, “Advances in Computer Science Research”, vol.77, pp.275-278, 2018.

[9] R.A.Rossi, N.K.Ahmed, “The Network Data Repository with Interactive Graph Analytics and Visualization”, AAAI, 2015.

[10] R. Cattral, F. Oppacher, D. Deugo, “Evolutionary Data Mining with Automatic Rule Generalization”, Recent Advances in Computers, Computing and Communications, WSEAS Press, pp.296-300, 2002.

[11] D. Dheeru, G. Casey, “{UCI} Machine Learning Repository”, University of California, Irvine, School of Information and Computer Sciences.

[12] B. Kaluza, V. Mirchevska, E. Dovgan, M. Lustrek, M. Gams, An Agent-based Approach to Care in Independent Living, International Joint Conference on Ambient Intelligence (AmI-10), Malaga, Spain

[13] C. Meek, B. Thiesson, D. Heckerman, “The Learning-Curve Sampling Method Applied to Model-Based Clustering”, Journal of Machine Learning Research, vol.2, no.2, pp.397-418, 2002

[14] A.M. Rubinov, N.V. Soukhorukova, and J. Ugon, Classes and clusters in data analysis, Euro. J. Operational Research, vol. 173, pp. 849–865, 2006

[15] D.L.Davies, and D. W. Bouldin. “A Cluster Separation Measure”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[16] Google Trends, https://trends.google.com/trends/?geo=US

[17] Statista, Statistics&Fact, https://www.statista.com/topics/2019/youtube/

AN ITERATIVE WEIGHTED K-MEANS ALGORITHM FOR BIG PERSONAL DATA ANALYSIS

Ramiz M. Aliguliyev[1], Shalala F.Tahirzada[2]
[1,2] Institute of Information Technology of Azerbaijan National Academy of Sciences
[1] r.aliguliyev@gmail.com, [2] fmv.shalala@gmail.com

Abstract -- The development of information and communication technologies (ICT) has led to a significant increase in digital information and consequently led to the spread of the concept of “big data". Therefore, there is a need for a comprehensive study of big data and its essence, various sources, capabilities of analytics technologies, problems, and security issues. The article addresses the weaknesses of existing analytical technologies, security issues in the analysis of personal data, and examines their solutions. Iterative and iterative weighted k-means algorithms are proposed to eliminate insufficiency in big data analysis.

Keywords *-- big data, personal data, personal data security, clustering, iterative and weighted clustering.*