# 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.

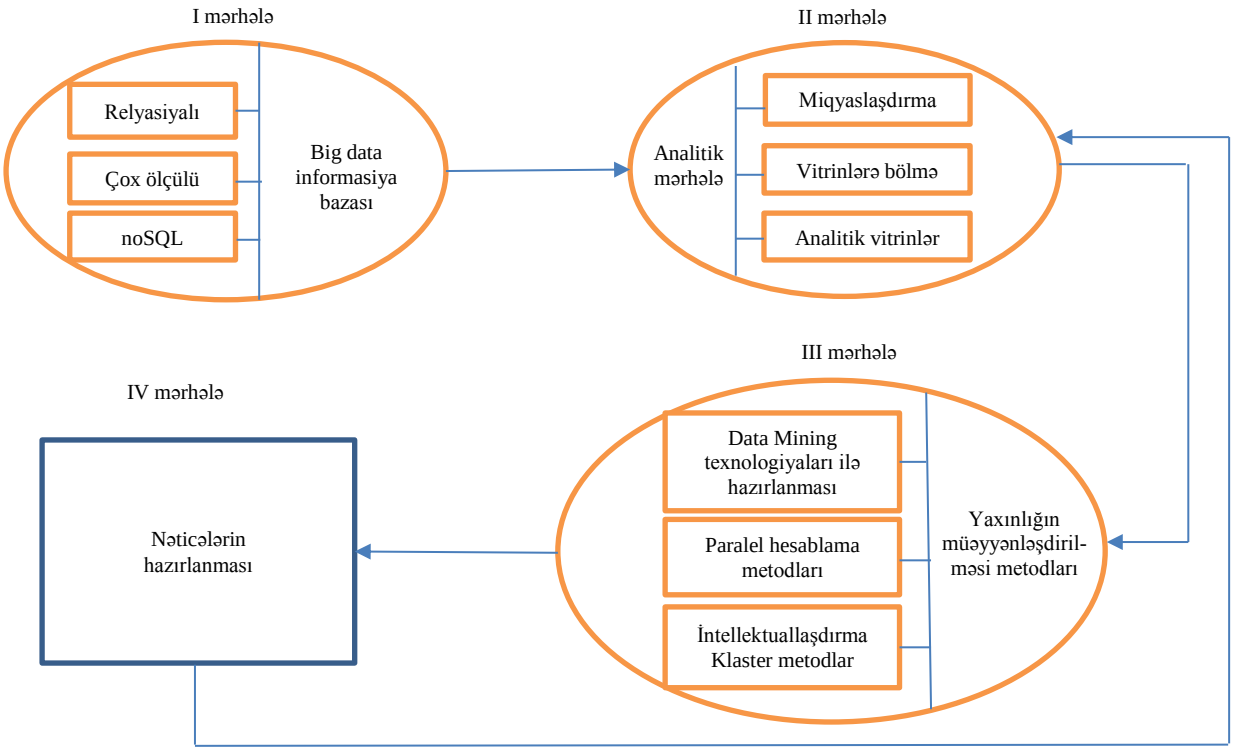
**IV mərhələ**: normal başa çatmış yoxlamadan sonra nəticələr haqqında hesabatın hazırlanması.

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