UOT 004.051

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

VİKİ-MÜHİTDƏ BÖYÜK VERİLƏNLƏRLƏ BAĞLI PROBLEMLƏR VƏ ONLARIN HƏLLİ YOLLARI

Məqalədə viki-mühitdə toplanan və durmadan artan böyük həcmli verilənlərin yaratdığı problemlər analiz edilmiş, müxtəlif məqsədlər üçün qərarların qəbulunda vikimetrik tədqiqatlardan istifadə perspektivləri müəyyənləşdirilmişdir. Viki-mühitdə böyük verilənlərin saxlanması və emalı üçün zəruri şərtlər göstərilmiş, Big Data ilə əlaqədar problemlərin həlli üçün bəzi mövcud yanaşmalardan birgə istifadə təklif edilmişdir.

***Açar sözlər**: viki-mühit, viki-texnologiya, Vikipediya, Big Data, vikianalitika, xromoqram, intellektual analiz, Map-Reduce.*

Giriş

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

Viki-texnologiyaları ilə idarə olunan Vikipediya (ing.Wikipedia) virtual ensiklopediyası və onun törəmə layihələri (Vikikitab, Vikimənbə, Commons, Vikinövlər və s.) ümumilikdə viki-mühit təşkil edirlər və onun nəhəng məlumat bazası hər saniyə yeni məqalə, xəbər, kitab, foto, audio və video-fayllarla zənginləşir [2, 3]. Tədqiqatın əsas məqsədi 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.

Viki-mühitdə böyük verilənlərlə bağlı problemlər

Viki-mühit özündə milyonlarla veb-səhifələri birləşdirir. 2016-cı ilin yanvar ayına olan məlumata görə, yalnız Vikipediya layihəsi 60 milyondan artıq qeydiyyatdan keçmiş istifadəçiyə və 40 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. On milyonlarla istifadəçinin şəxsi və müzakirə səhifələrini, sorğuları, yaradılan şablon və bot-proqramları da əlavə etsək, informasiyanın həddən artıq çox olmasını təsəvvür etmək çətin deyil. İstifadəçilərin sayının durmadan artması və yüksək aktivliyi qoyulan məsələlərin həllində BD texnologiyalarından istifadənin aktuallığını önə çəkir. BD-də olduğu kimi, viki-mühitdə də problemlər əsasən verilənlərin həddən artıq çox olması və qeyri-strukturlaşdırılmış formatda olması ilə bağlıdır.

Əsas problemlər aşağıdakılardır:

– *Verilənlərin ölçüsü.* Vikipediya layihələrinə aid VB-lərdə verilənlərin həcmi baytlarla göstərilir. Vikipediyanın milyonlarla səhifələrində və fayllarında toplanan verilənlərin həcmi onların BD kimi analiz olunub-olunmamasını müəyyən edir. Böyük həcmdə verilənlərin emalı və saxlanması üçün isə xüsusi şərtlər tələb olunur.

- *Verilənlərin müxtəlifliyi.* Viki-kontentlər müxtəlif tipdə ola bilərlər. Viki-səhifələrə mətn, səs, şəkil və video-fayllar yerləşdirmək mümkün olduğuna və viki-texnologiyalar faylların əksər formatlarını dəstəklədiklərinə görə verilənlər müxtəlif tipdə və strukturda toplanırlar. Onları bir yerə toplamaq və eyni zamanda emal etmək mümkün deyil. İlk növbədə verilənləri analiz üçün uyğun formata 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-də qalır və istənilən zaman istifadəçi tərəfindən bərpa edilməsi mümkündür. Saniyədə milyonlarla informasiya saxlayan Vikipediyanın VB-də heç bir informasiyanın silinməməsi zaman keçdikcə verilənlərin həddən artıq çoxalması deməkdir. Verilənlərin emalında sürət nə qədər çox olarsa, interaktivlik şərtləri daha yaxşı ödənilər.
- *Verilənlərin dəyişkənliyi.* Viki-səhifələr açıqdırlar və hər an birbaşa brauzerdən dəyişdirilirlər. Baxılan səhifə yenilənmədən sonra artıq tam başqa strukturda və dizaynda təqdim oluna bilər. Səhifənin bütün köhnə versiyaları (yaranma tarixi, şərhlər, silinmiş və dəyişdirilmiş kontentlər, müəlliflər göstərilməklə) bazada saxlanılır. Bu faktorlar analiz zamanı verilənlərin idarə olunmasında müəyyən problemlər yaradırlar.
- *Mürəkkəblik.* Verilənlər müxtəlif mənbələrdən daxil olurlar. Viki-səhifələr arasında mürəkkəb semantika mövcuddur. Səhifələr arasında linklər müxtəlif tipdə olurlar: layihədaxili, dil faktoruna görə (intervikilər), kateqoriyalara görə, digər təyinatlı viki-layihələrə və nəhayət, xarici layihələrə (İnternetin müxtəlif veb-saytlarına) linklər. Bu linklər hər an dəyişdirilə və ləğv edilə bilər. Belə bir mürəkkəblik vikimetrik tədqiqatları da mürəkkəbləşdirir.

Verilənlərin ənənəvi emalı zamanı onlar yoxlanılır, müəyyən formata gətirilir və sistemə yüklənilirlər. Bu cür ardıcıllıq viki-mühitdə toplanan verilənlərin saxlanması və emalı məsələlərində özünü doğrultmur. Müxtəlif elmi ədəbiyyatlarda göstərildiyi kimi, 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 [2, 7–9]. Bu məsələlərin həlli vikianalitik tədqiqatlarda da aktuallığını saxlayır.

Vikianalitika

Vikianalitika viki-mühitdə toplanan verilənlərin ölçülməsi, analizi və qiymətləndirilməsi haqqında elmdir. Viki-səhifələrdəki kontentlər və Vikipediya layihələrinə müraciət edən istifadəçilər haqqında informasiyanın ölçülməsi, analizi və təsviri vikianalitika vasitəsi ilə həyata keçirilir. Vikianalitik tədqiqatların əsas məqsədi veb-auditoriyanı təyin edən verilənlər əsasında viki-səhifələrdə yerləşdirilən kontentin keyfiyyətinin qiymətləndirilməsi, bu səhifələrdə trafikin monitorinqinin aparılması və müxtəlif məqsədlər üçün qərarların qəbulunda Vikipediya istifadəçilərinin davranışlarının öyrənilməsidir. Vikimetrik tədqiqatlardan istifadə etməklə viki-səhifələrin informativliyini, təqdim olunan məlumatın düzgünlüyünü, aktuallığını, əhatəliliyini, viki-səhifələr arasında əlaqələrin rahatlığını, məntiqi strukturunu tədqiq etmək, viki-səhifələrin yaradılmasında iştirak edən istifadəçilərin fəaliyyət məqsədlərini və viki-mühitdə informasiya müharibəsinin reallaşdırılmasında bilavasitə iştirak edən, müəyyən ideoloji məqsəd daşıyan gizli sosial şəbəkələri təyin etmək mümkündür.

Viki-səhifələrdə yerləşdirilən kontentin keyfiyyətinin təyin edilməsi və qiymətləndirilməsi vasitəsilə aşağıdakı məlumatlar əldə oluna bilər:

- viki-səhifələrdə vandalizm halları;
- səhifənin adının mövzuya uyğunluğu;
- viki-səhifələr arasındakı semantika;
- viki-mühitdə konfliktli situasiyalar;
- informasiya qarşıdurmalarında iştirak edən sosial qrupların aşkarlanması.

Viki-səhifələrin qiymətləndirilməsində mövzunun əhatəliliyi, struktur elementlərinin sayı, digər İnternet resursları ilə əlaqələrin sayı, istifadəçi tərəfindən viki-səhifəyə olunan müraciətlərin sayı və vaxtı, viki-səhifələrdə informasiyanın həcmi, yüklənən multimedia fayllarının keyfiyyəti və mövzuya uyğunluğu, qrammatik və orfoqrafik səhvlərin sayı, informasiyanın etibarlılığı, əlyetərliliyi, hiperistinadlardan təşkil olunmuş siyahıların sayı və dinamikliyi, informasiyanın təqdim olunma stili, materialın başqa dillərdə təqdim olunmasında rahatlıq və etibarlılıq və s. kimi parametrlər də nəzərə alınmalıdır.

Viki-səhifələrin trafikinin statistikasını aparmaqla, aşağıda göstərilən müxtəlif məlumatları əldə etmək olar:

– viki-səhifələrə müraciətlərin sayı;
– səhifənin axtarışında istifadə olunan açar sözlər və ifadələr;
– istifadəçilərin sosial-demoqrafik portreti;
– istifadəçilərin Vikipediya layihələrində və ayrı-ayrı viki-səhifələrdə keçirdikləri vaxt;
– viki-səhifələrdəki məlumatların aktuallığı;
– kateqoriyalar üzrə istifadəçilərin sayı: daimi və təsadüfi.

Analitik nöqteyi-nəzərdən yuxarıda göstərilən imkanlar müəyyən situasiyaların qiymətləndirilməsində dəqiqliyi yüksəldir və hadisələrin inkişafında mümkün ssenarilərin hazırlanmasına və tətbiqinə imkan verir. Viki-mühitdə yaranan BD bu gün müxtəlif sahələrdə tədqiqatlar aparmaq üçün Vikipediyanı bir poliqona çevirmiş və elmi tədqiqatlar üçün böyük imkanlar yaradır. Böyük verilənləri emal etməklə oradan gizli informasiyanı çıxarmaq bütün elmi sahələrin əsas mövzusudur. Viki-mühitdə BD ilə əlaqədar problemlər verilənlərin emalında yeni yanaşmaların axtarılmasına ehtiyac yaradır. Nəzərə almaq lazımdır ki, hər bir viki-səhifə ayrı-ayrılıqda elektron sənəddir və odur ki, sənədlərin keyfiyyətini və kəmiyyətini ölçmək üçün istifadə olunan bir çox parametrlərdən viki-səhifələrin ölçülməsində də istifadə olunur [2, 3, 7].

Viki-mühitdə verilənlərin kütləvi şəkildə toplanması və emalı dövlət orqanlarında, biznes və təhsildə müəyyən qərarların qəbuluna da təsir edə bilər. Viki-layihələrdə daima yenilənən kollektiv biliyin yaranma prosesini öyrənmək, Vikipediya ensiklopediyasında toplanan məqalələrin bütün sahələri nə dərəcədə əhatə etməsini, viki-kontentlərin informasiya təsirində və qarşıdurmasında rolunu təyin etmək üçün müqayisəli analiz metodları, qeyri-səlis altçoxluqlar nəzəriyyəsi, klasterləşmə metodları, qraflar nəzəriyyəsi üstünlük təşkil edir. Viki-mühitin analizində – məqalələrin müəyyən əlamətlərə görə təsnifatlandırılmasında və sosial şəbəkənin təyinində isə semantik analiz metodlarından geniş istifadə edilir. Semantik analiz metodlarında HITS (*Hyperlink-Induced Topic Search*), PageRank, Random Forest alqoritmlərindən istifadəyə geniş yer verilir [8, 9, 10].

Vikimetrik tədqiqatlarda ilkin olaraq açıq olan və hər kəsin istifadə edə bildiyi bazalara müraciət olunmalıdır. Bu bazalardakı göstəricilər aşağıdakılardır:

– viki-kontentlərin saxlandığı VB-nin həcmi (bütün viki-səhifələrin həcmi, müzakirə, kateqoriya və istiqamətləndirmə səhifələri də daxil olmaqla);
– viki-layihələrdə ensiklopedik məqalələrin sayı;
– hər bir ensiklopedik məqalədə sözlərin sayı;
– viki-səhifələrə müraciətlərin sayı;
– viki-səhifələrdə daxili linklərin sayı;
– viki-səhifələrdə xarici linklərin sayı;
– viki-səhifələrin baytlarla həcmi;
– viki-layihələrdə qeydiyyatdan keçmiş istifadəçilərin və anonimlərin sayı;
– aktiv istifadəçilərin sayı (bir ay ərzində 5 redaktədən çox redaktəsi olan istifadəçilər);
– daha aktiv istifadəçilərin sayı (bir ay ərzində 100 redaktədən çox redaktəsi olan istifadəçilər).

Viki-mühitin intellektual analizi tədqiqatların daha dərindən aparılmasını və gizli məlumatların əldə olunmasını mümkün edir. Gizli məlumatlardan problemlərin həllində və

qərarların qəbulunda istifadə olunur. İntellektual analiz üçün viki-mühit geniş imkanlara malikdir və bu imkanlar aşağıda göstərilmişdir:

a) Viki-mühit sənədləri idarə edir:
- Milyonlarla viki-səhifələrin mövcudluğu.
- Səhifələrin müxtəlif tipli yüzlərlə linklərə malik olması (Dense link structure).

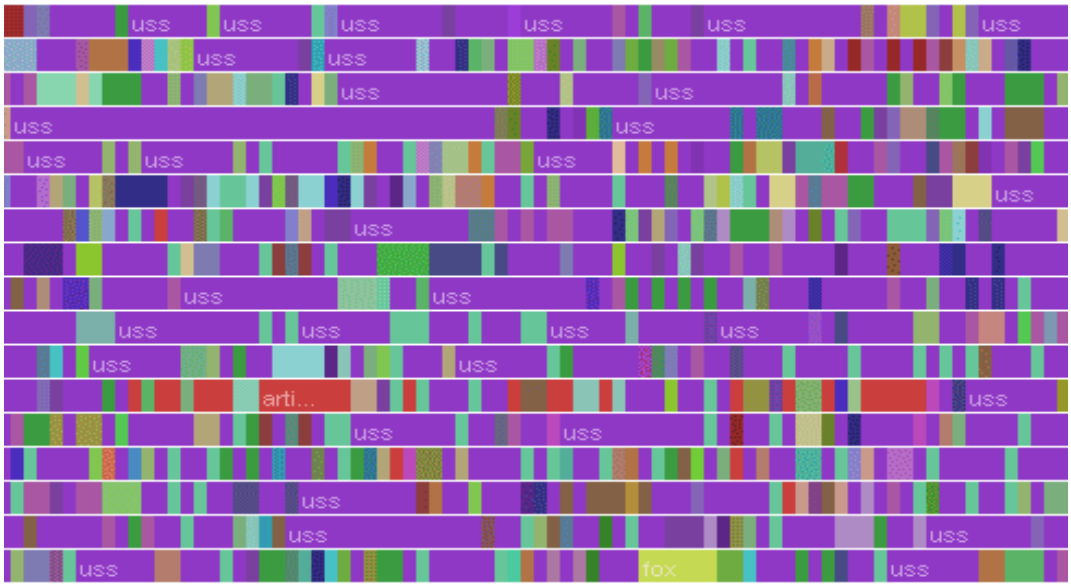
b) Viki-mühit bilikləri idarə edir:
- Fasiləsiz olaraq yeni kontentin yaradılması.
- Kontentin dinamik dəyişdirilməsi.
- Kontentin sistemləşdirilməsi.

c) Viki-mühit nəhəng sosial şəbəkədir.

Vikipediyanın yuxarıda göstərilən imkanları ilə əlaqədar Wiki Mining, Text Mining və Link Mining metodları əsasında viki-mühitdə səhifələr arasında semantik əlaqələrin analizini aparmaq, ensiklopedik məqalələrin əhatə dairəsini və keyfiyyətini təyin etmək, gizli sosial şəbəkələri aşkar etmək mümkündür [11, 12].

Problemin həllinə mövcud yanaşmalar

Vikimetrik tədqiqatlarda müxtəlif metodlardan istifadə (sayğac, jurnal fayllarının analizi, müraciətlər, müşahidələr və s.) edilir. İnternetdə BD ilə bağlı problemlər Vikipediya resurslarının ölçülməsi və istifadəçilərin davranışlarının öyrənilməsində vikianalitikadan güclü vasitə kimi istifadə olunmasını tələb edir. IBM şirkəti Vikipediyada istifadəçilərinin 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 ranglər mətndəki sözlərə verilən ranglərdən asılıdır. Mətnin ranglərlə təsviri metodu xromoqramın əsas fərqləndirici cəhətidir [13]. Məsələn, şəkil 1-də göstərilən xromoqram hərbi-dəniz qüvvələrinin tarixi ilə bağlı məqalələrdə edilən düzəlişlər əsasında təşkil olunmuşdur.



Şəkil 1. Viki-mühitdə istifadəçinin maraq dairəsini göstərən xromoqram

Xromoqramda üstünlük təşkil edən bənövşəyi rəng hərbi-dəniz gəmilərinin adları olan məqalələri göstərir. Yəni, xromoqram göstərir ki, istifadəçi müxtəlif mövzulara müraciət etsə də, hərbi gəmilərə olan maraq üstünlük təşkil edir. Xromoqramdan istifadə etməklə istifadəçinin hansı mövzuya maraq göstərməsini, vandalizm hallarını, viki-mühitdə fəaliyyət göstərən sosial şəbəkələri, konfliktli situasiyaları və digər maraqlı məsələləri analiz etmək mümkündür [12, 13].

Xromoqramdan istifadə etməklə istifadəçinin fəaliyyətinin xarakterini də müəyyən etmək mümkündür. Sistemdə hər bir fəaliyyət növü xüsusi rənglə işarələnir. Belə ki, viki-mühitdə hər bir istifadəçinin özünəməxsus iş stili vardır və onlar aşağıda göstərilmişdir:

- yeni səhifənin yaradılması;
- mövcud səhifəyə kontentin daxil edilməsi;
- informasiyanın səhifədən silinməsi və ya dəyişdirilməsi;
- dizayn işləri və xüsusi şablonların əlavə edilməsi;
- səhifələr arasında semantikaya nəzarət;
- müəyyən mövzu ətrafında müzakirələrdə iştirak.

Vikipediyada böyük verilənlərin analizini nəzərdə tutan digər sistem SGI UV 2 (Silicon Graphics International, ultraviolet) sistemidir [13]. SGI şirkəti və İllinoys Universitetinin əməkdaşı K.Liptau (Kalev H. Leetaru) tərəfindən yaradılan bu sistem Vikipediyada tammətnli 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. SGI UV 2 sistemi vasitəsi ilə dünyanın 1900-cü il üçün vizual tarixi xəritəsinin müəyyən edilməsi

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" [14]. Sistemdə verilənlərin emalı əsas yaddaşda toplanan verilənlərin intellektual analizi (*ing. in-memory data-mining*) vasitəsi ilə həyata keçirilir. İntellektual analiz alqoritmləri, ilk növbədə, mətnin xüsusiyyətini, məzmununu, səhifələr arasındakı əlaqələri üzə çıxarır. Nəticələr metaverilənlərdə saxlanılır və sonrakı analizlərdə istifadə olunur.

Bəzi yanaşmalarda BD problemləri ənənəvi 3V (volume (həcm), variety (müxtəliflik), velocity (sürət)) [15] ilə deyil, 3C parametrləri ilə göstərilir: cardinality (həddən artıq böyüklük), continuity (fasiləsizlik) və complexity (mürəkkəblik) [16]. 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 kontentlərin tipinə və formatına görə müxtəlifliyi, qeyri-strukturlaşdırılmış formatda olması, hər saniyə dəyişdirilmə ehtimalı verilənlərin sürətli emalını çətinləşdirir.

Problemin Big Data texnologiyaları əsasında həlli

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 proqnozlaşdırsın. 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 köçürülməlidir. 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 müxtəlif serverlərdəki verilənlərin strukturlaşdırılması və bazada toplanması müxtəlif olur. Verilənlərin analizi baxımından onları 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 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 sürətli emal imkanına malik intellektual analiz alqoritmlərinə ehtiyac yaranır. 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.

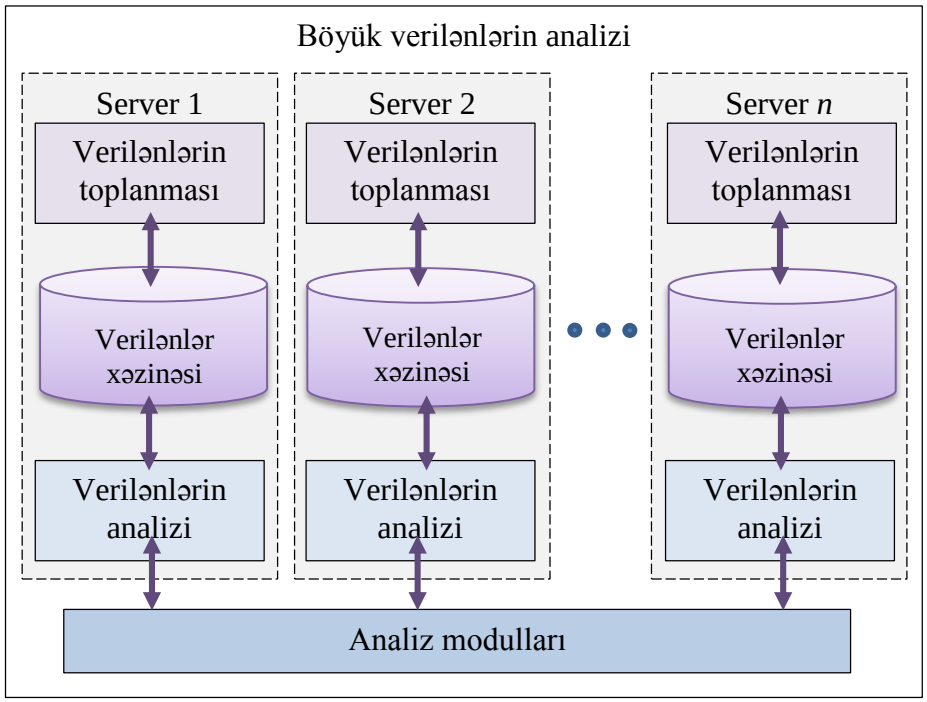
Nəzərə almaq lazımdır ki, böyük verilənlərin emalı prosesində təşkil olunan massivlər heç də həmişə təsadüfi seçmədən əldə olunmur. Bəzən statistik verilənlər daha böyük əhəmiyyət kəsb edirlər. Burada əsas problem ondan ibarətdir ki, böyük, mürəkkəb, keterogen verilənlərin statistik analizi səhv və uyğun 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 [17].

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. Buna misal olaraq, agent modelləşdirməni göstərmək olar. Agent modelləşdirmə paylanmış verilənlərin davranışını tədqiq edir və bu metodla bütün sistemin davranışı təyin edilir [18]. Eyni zamanda faktor və klaster analizi kimi çoxölçülü analiz metodlarından istifadə etməklə verilənlərin gizli stukturunu tədqiq etmək mümkündür [19]. Böyük verilənlərin klasterləşdirilməsində Hadoop Distributed File Systems (HDFS) [20] modelindən,

Hive database və ZooKeeper metodlarından [21], Cloud texnologiyalarından [22] 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 və paylmış əlavələrin qurulmasında istifadə olunur. Hive – paylanmış verilənlər xəzinəsidir. Hive xüsusi sorğu dili olan HiveQL dilində işləyir və HDFS-də toplanmış verilənləri idarə edir.

BD problemlərinin həlli ilə bağlı yanaşmaların qısa təhlili göstərir ki, verilənlərin paylanmış şəkildə emalı onların 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 serverdə yerləşən verilənlər xəzinəsində saxlanılır və indeksasiya edilir (şəkil 3). 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. VB üçün verilənlərin paylanması sorğuların yerinə yetirilməsi zamanı qovşaqlar arasında informasiya mübadilələrinin sayının minimum olmasına gətirib çıxaracaqdır.



Şəkil 3. Böyük verilənlərin analizində istifadə olunan paylanmış sistem

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 yaradırlar. BD ilə işləmək üçün ən populyar texnologiyalara misal olaraq, 2004-cü ildə Google şirkəti tərəfindən təklif olunan Map-Reduce modelini və Apache Software Foundation şirkəti tərəfindən təklif olunan açıq kodlu Hadoop proqramını göstərmək olar. Bu sistemlərdə verilənləri cədvəllərə yığmaq və ciddi iyerarxiyaya tabe etmək lazım gəlmir. Analizdən öncə verilənləri növlərə bölmək kifayətdir. Map-Reduce modeli verilənlərin düzgün paylanması işində səmərəliliyi təmin edir. Map-Reduce – paylanmış verilənlərin hesablanması modelidir və böyük

verilənlər üzərində paralel hesablamalar üçün istifadə olunur. Giriş verilənlərin ilkin emalı üçün nəzərdə tutulmuş xəritə (Map) işçi qovşaqlara (*ing. 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 qovşağında aralıq cəmləri hesablayacaq və sonrakı mərhələdə bu 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 (*ing. non-deterministic polynomial*) sinfindəki 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 validasiyası vacib məsələlərdəndir. Verifikasiya və validasiya proqram təminatının keyfiyyətinə nəzarət etmək və sistemdəki nasazlıqları müəyyən etmək üçün istifadə olunur. Verifikasiya hər hansı məsələnin həllində təklif edilən yanaşmaların əvvəlki yanaşmalarla müqayisəsini təmin edir və proqram təminatına qoyulan tələblər, standart normalar, istifadəçi sənədləri arasındakı uyğunluqları yoxlayır. Validasiya isə təklif edilən yanaşmanın istifadəçi və ya sifarişçilərin tələbləri ilə uyğunluğunu yoxlayır. Bu tələblər çox zaman rəsmi sənədləşdirilməsələr də, tələblərin ümumi təsvirində yer alırlar. Nəzərə almaq lazımdır ki, verilənlər həddən artıq böyük olduqda xətaların ölçülməsi də vacib məsələlərdən biridir. Böyük verilənlərin emalında xətalar və ölçmə zamanı yaranan xətaların sayı qarşılıqlı əlaqəlidir.

Nəticə

Tədqiqat nəticəsində məlum olmuşdur ki, viki-mühitdə böyük verilənlərin toplanması onların emalında və saxlanmasında problemlər yaradır və məsələlərin həllində yeni yanaşmaların tətbiqini tələb edir. Viki-mühitdəki böyük verilənlərin vizuallaşdırılması və analizində istifadə olunan mövcud sistemlərin iş prinsipləri belə deməyə əsas verir ki, Vikipediyada toplanmış verilənlər son illər özlərində BD-nın xüsusiyyətlərini daşımaqdadırlar: ölçüləri böyükdür – onların səmərəli strukturlaşdırılması və saxlanması üçün xüsusi şərtlər tələb olunur və 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.

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. Digər tərəfdən, viki-mühitdəki verilənlərin müxtəlifliyi və dəyişkənliyi onların analizi və təsvirində kompüter qrafikasının yeni texnologiyalarından istifadəni qaçılmaz edir.

Ədəbiyyat

1. Əliquliyev R.M., Hacırəhimova M.Ş. "Big Data" fenomeni: problemlər və imkanlar // İnformasiya texnologiyaları problemləri, 2014, №2, s.3–16.
2. Alakbarova I.Y. 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), Baku, Sept. 12–14, 2012, vol.I, pp.119–120.

3. Алгулиев Р.М., Алыгулиев Р.М., Алекперова И.Я. Викиметрические исследования: современное состояние и перспективы // Телекоммуникации, 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. Ələkbərova İ.Y. Viki mühitdə reallaşdırılan bəzi informasiya müharibəsi texnologiyalarının analizi // İnformasiya cəmiyyəti problemləri, 2011, №2, səh.18–28.
8. Arazy O., Stroulia E.A. Utility for estimating the relative contributions of wiki authors / Proceedings of the Third International ICWSM Conference, San Jose, California, May 17–20, 2009, pp.171–174.
9. Halfaker A., Keyes O., Taraborelli D. Making Peripheral Participation Legitimate: Reader engagement experiments in Wikipedia / Proceedings of the 2013 conference on Computer supported cooperative work, ACM, NY. USA, 2013, pp.849–860.
10. Əliquliyev R.M., Ələkbərova İ.Y. Vikimetrik tədqiqatlar: müasir vəziyyəti, problemləri və perspektivləri, Ekspress-informasiya, Bakı, "İnformasiya texnologiyaları" nəşriyyatı, 2015, 87 səh.
11. Iba T., Nemoto K., Peters B., Gloor P.A. Analyzing the creative editing behavior of Wikipedia editors: through dynamic social network analysis // Procedia – Social and Behavioral Sciences, 2010, vol.2, no.4, pp.6441–6456.
12. Müller C., Meuthrath B., Baumgra A. Analyzing wiki-based networks to improve knowledge processes in organizations // Journal of Universal Computer Science, 2008,vol.14, no.4, pp. 526–545.
13. Wattenberg M., Viégas F.B., 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.
14. www.sgi.com/go/wikipedia/
15. Alguliyev R., Imamverdiyev Y. 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.
16. Hilbert M. Big Data for Development: From Information – to Knowledge Societies, 2013. http://ssrn.com/abstract=2205145
17. Couper M. Is the sky falling? New technology, changing media, and the future of surveys // Survey Research Methods, 2013, vol.7, no.3, pp.145–156.
18. Suthaharan S. 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.
19. Alguliev R.M., Aliguliyev R.M., Alekperova I.Ya. 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.
20. Thusoo A., Sarma J.S., Jain N., Shao Z., Chakka P., Zhang N, Antony A., Liu H., Murthy R. 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.
21. Hunt P., Konar M., Junqueira F.P., Reed B. 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.
22. Carlin S., Curran K. Cloud Computing Technologies // International Journal of Cloud Computing and Services Science (IJ-CLOSER), 2012, vol.1, pp.59–65.

УДК 004.051

Алекперова Ирада Я.
Институт Информационных Технологий НАНА, Баку, Азербайджан
airada.09@gmail.com

Проблемы, связанные с большими данными в вики-среде, и способы их решения

В статье анализированы проблемы, которые возникли из-за собранных и непрерывно растущих данных большого объема в вики-среде, определены перспективы использования викиметрических исследований при принятии решений для различных целей. Показаны необходимые условия для хранения и обработки данных большого объема в вики-среде, предложено совместное использование некоторых существующих подходов для решения проблем, связанных с Big Data.

Ключевые слова: *вики-среда, вики-технология, Википедия, Big Data, вики аналитика, хромограмма, интеллектуальный анализ, Map-Reduce.*

Irada Y. Alakbarova
Institute of Information Technology of ANAS, Baku, Azerbaijan
airada.09@gmail.com

The problems associated with big data in wiki-environment and their solution ways

The article analyzes the problems arisen due to the collected and continuously growing large-scaled data in wiki-environment. The prospects of using wiki-metric research in decision-making for various purposes are defined. The necessary conditions for the storage and processing of large amounts of data in a wiki-environment are presented. It is proposed to jointly use some existing approaches to solve the problems associated with Big Data.

Keywords: *wiki environment, Wiki technology, Wikipedia, Big Data, wiki-analytics, hromogram, data mining, Map-Reduce.*