UOT 004.02 DOI: 10.25045/jpit.v11.i1.12

Qasımova R.T. [1]***, Abbaslı R.N.*** [2]
[1]AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan,
[2]Thinking capital, Montreal, Kvebek, Kanada
[1]renakasumova@gmail.com, [2]rahim.abbasli@gmail.com

BÖYÜK HƏCMLİ RƏQƏMSAL VERİLƏNLƏRDƏ AXTARIŞ ALQORİTMLƏRİNİN ANALİZİ

Daxil olmuşdur: 10.04.2019 Düzəliş olunmuşdur: 08.05.2019 Qəbul olunmuşdur: 14.05.2019

Rəqəmsal materiallar fasiləsiz artan, müxtəlif formatlarda təqdim edilən mətn sənədlərindən, verilənlər bazasından, təsvirlərdən, səsli və qrafik materiallardan, proqram təminatı və veb-səhifələrdən ibarətdir. Rəqəmsal informasiyanın sürətli artımı, verilənlərin müxtəlifliyi unikal verilənlər strukturunun vaxtında analiz edilməsi, daha uyğun, dolğun axtarışı və emalı zərurətini yaratmışdır. Bu məqsədlə məqalədə qeyri-ənənəvi verilənlər strukturuna malik böyük verilənlərin emalı zamanı süni intellektə əsaslanan alqoritmlərin xüsusiyyətləri analiz olunmuşdur. Müəyyən edilmişdir ki, axtarışın keyfiyyətinin yaxşılaşdırılması, verilənlərin həcminin böyüklüyü, istifadəçi sorğularının axınının intensivliyi ilə bağlı məsələlərin həlli zamanı süni intellektə əsaslanan alqoritmlərdən istifadə olunması zəruridir. Eyni zamanda məqalədə axtarış alqoritmləri təhlil olunmuş, onların problemləri təsnifatlandırılmış və axtarış sistemlərinin imkanlarının maksimum istifadəsi üçün potensial həllər təklif edilmişdir.

Açar sözlər: *rəqəmsal irs, rəqəmsal verilənlər, Big Data, Big Data analitikası, axtarış sistemləri, informasiya axtarışı, süni intellekt, maşın təlimi.*

Giriş

Bəşəriyyətin irsi həmişə ən davamlı vasitələrdən istifadə edərək əsrlərdən keçmişdir. Amma keçmişdən yığılmış verilənlərin tam olmaması səbəbindən indiki nəsil öz tarixini başa düşməkdə çətinlik çəkir. Hazırkı dünyada insanlar da başqa sürətlə işləyirlər. Obyektlərin zəruri olaraq rəqəmsallaşdırılması onu qiymətləndirə bilən nəsil yaratmışdır. Yeni nəslin tələbatı sürətlə artdığından, şirkət və hökumətləri zamanla ayaqlaşmağa məcbur edir. Bu gün rəqəmsal irs (Rİ) ideyası qlobal rəqəmsal dünyanın bir hissəsidir.

Rİ mədəniyyət, təhsil, elm, idarəetmə resurslarından, həmçinin texniki, hüquqi, tibbi və digər xarakterli informasiyadan ibarətdir. Yəni, Rİ rəqəmsal formada yaradılmış və ya mövcud analoq resurslarından rəqəmsal formaya çevrilmiş mədəniyyət, təhsil, elm və idarəetmə sahələrinə aid resursları, həmçinin texniki, hüquqi, tibbi və digər informasiya növlərini əhatə edir. İstənilən halda onlar insanlar tərəfindən zaman və ərazi sərhədlərindən asılı olmayaraq istifadə edilir [1].

Rəqəmsal resurslar geniş diapazonu – tibbi informasiyaları, peyklərdən müşahidə verilənlərini, multimedia resurslarını, insan genomunu qeyd edən elmi verilənlər bazalarını, şəbəkə informasiya arxivlərindən muzey kataloqlarını və s. əhatə edir. *Rəqəmsallaşdırma* – orijinal (analoq) materialın rəqəmsal formaya transformasiyası prosesidir [2].

Rİ-nin saxlanmasında inkişaf etmiş ölkələrin təcrübəsi göstərir ki, irs materiallarının toplanması, qorunub saxlanılması funksiyaları ilk olaraq milli kitabxanaların, arxivlərin, muzeylərin və irs idarələrinin üzərinə düşür [3, 4]. Hazırda dünya kitabxanalarında, muzeylərində, arxivlərində toplanan müxtəlif növ sənədlərə, kolleksiyalara, materiallara əlçatanlığı təmin etmək üçün onların rəqəmsallaşdırılması prosesi daima həyata keçirilir. Bu yalnız kitabxana və arxivlərə deyil, həm də bütün rəqəmsal informasiya resurslarını istehsal edən və onlara uzunmüddətli əlçatanlığa maraqlı olanlara da aiddir. Bunun nəticəsidir ki, son zamanlar böyük həcmdə informasiya massivlərinin toplanması, eyni zamanda onların çox böyük sürətlə artması həm akademik mühitdə, həm də informasiya texnologiyaları sahəsində daha çox diqqət cəlb etməyə başlamışdır.

Məqalədə tədqiqat obyekti kimi kitabxanaların, arxiv sənədlərinin və irs materiallarının rəqəmsallaşdırılması prosesində əsasən strukturlaşdırılmamış verilənlərdən ibarət olan Rİ

seçilmişdir. İrs materiallarının rəqəmsallaşdırılması verilənlərin emalında sadə axtarış sistemlərinə (AS) və ya analitik alətlərə nisbətən daha yaxşı həllər tələb edir. Baxılan problemin araşdırılması, elmi-nəzəri tədqiqatların təhlili əsasında alınan nəticə göstərir ki, bu gün istifadə olunan alqoritmlər dəyişən tələblərdən geri qalmamaq, onlardan asılı olmamaq üçün çevik olmalıdırlar. Bu mühüm məsələləri nəzərə alaraq böyük həcmli verilənlərdə sürətli və səmərəli axtarışın aparılmasında AS-in elmi-nəzəri problemlərinin araşdırılması və elmi-tədqiqat obyekti kimi öyrənilməsi vacibdir, aktualdır.

Big Data erasında axtarış sistemlərinin problemləri

İnformasiya axtarışını AS reallaşdırır. AS istifadəçiləri maraqlandıran informasiyanın saxlanılması, axtarılması və təqdim edilməsi üçün nəzərdə tutulmuş proqram-texniki sistemdir. İnformasiya axtarışı iterativ prosesdir, yəni istifadəçinin tələbləri ödənilmədikdə, ona sorğusunda dəyişikliklər etmək və yeni sorğu üzrə təkrar axtarış aparmaq imkanı verilir. AS-in əsas məqsədi istifadəçilərin informasiya tələbatlarının ödənilməsini həyata keçirməkdən ibarətdir [5].

International Data Corporation (IDC) beynəlxalq analitik şirkətinin apardığı statistik hesabata əsasən, 2025-ci ildə dünyada rəqəmsal informasiyanın həcmi 175 zetabayt olacaqdır. Bu da 2018-ci illə müqayisədə 10 dəfə çoxdur. 2020-ci ildə İnternetə qoşulan qurğuların sayının 50 milyard olacağı gözlənilir. Hesabatda, həmçinin informasiya dünyasında baş verən digər tendensiyalar, cəmiyyətin həyatı üçün informasiyanın artan əhəmiyyəti, onun buludlara qlobal miqrasiyası, insan müdaxiləsi olmadan məlumatlar generasiya edən yeni qurğuların dalğası və s. öz əksini tapır. Eyni zamanda tədqiqatın əsas proqnozlarında aşağıdakılar da qeyd edilir [6, 7]:

– İnformasiyanın həcmi növbəti səkkiz il ərzində hər iki ildən bir ikiqat artacaqdır. Bu artımın əsas amillərindən biri avtomatik generasiya olunan verilənlərin payının artmasıdır.

– Faydalı verilənlərin böyük həcmi itirilir. Bu gün böyük verilənlər (Big Data) texnologiyasının tətbiqi ilə isifadə oluna biləcək potensial faydalı verilənlərin 23%-indən 3%-i istifadə olunur.

– İnformasiyanın çox hissəsi yaxşı qorunmur.

– 2010-cu ildə informasiyanın üçdə birindən azının qorunması tələb olunurdusa, 2020-ci ilə belə informasiyanın payı 40%-i ötə bilər.

– Qorunma səviyyəsi regiondan asılı olaraq dəyişilir, inkişaf etməkdə olan bazarlar üçün bu, çox aşağıdır.

Rəqəmsal informasiyanın belə sürətli artımı, verilənlərin müxtəlifliyi, onların ötürülmə sürətinin yüksək artımı Big Data sahəsində çoxsaylı problemlərin yaranmasına səbəb olur. Yüzlərlə terabayt və ekzabayt həcmində Big Data-nin 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, real zamanda onların analizini aparmaq, daha dərin intellektual analiz apararaq nəticələri vizuallaşdırılmaq Big Data analitikanın əsas məsələlərindəndir. Bu da Big Data-nın ən əsas problemlərindən sayılan böyük verilənlərin analitikası (*ing. Big Data Analytics*) istiqamətinin yaranmasına gətirib çıxarmışdır [8–10].

Qeyd etmək lazımdır ki, bu gün qarşılaşdığımız əsas problemlərdən biri Big Data-nı analiz edə bilən ixtisaslı elmi kadrların çatışmazlığıdır. Həmçinin, bütün bunlara baxmayaraq ən böyük problem böyük həcmli verilənlərdən istənilən informasiyanı tez və dolğun əldə edə bilən intellektual alətin yaradılmasıdır. Big Data-nın müxtəlif mənbələrdən toplanması istənilən tip informasiya axtarışı probleminin mürəkkəbliyinə səbəb olur. Belə ki, istifadəçilər AS-dən istifadə edərək axtarışın yüksək dolğunluğunun və dəqiqliyinin təmin olunmasını istəyirlər. Yəni, onlar öz informasiya ehtiyaclarına**,** istəklərinə uyğun olmayan bütün sənədlərin atılmasını və axtarışın nəticəsi qismində onlara lazım olanın təqdim olunmasını tələb edirlər.

Ona görə də böyük həcmli verilənlərdə səmərəli axtarışın aparılması problemlər arasında ən başlıca pol oynayır. Yəni, yaradılan sürətli AS hələ də mövcud problemlərin həlli üçün kifayət

deyil. Hazırda istifadə edilən alətlər strukturlaşdırılmamış verilənlərlə işləyərkən aşağı keyfiyyət göstərirlər. Strukturlaşdırılmamış verilənlər daha sürətlə artan verilənlərdir. Bu tip verilənlər səsdən, təsvirdən (GIF, JPG, PNG, BMP və s. formatlarda), videofayldan, animasiyadan və s. ibarətdir. Strukturlaşdırılmamış verilənlərin gündəlik həyata daxil olması ünsiyyət üsulunu və nəticədə axtarışı da dəyişmişdir. Keçən il *Google*-da axtarış sorğularının 50%-dən çoxu səsli sorğulardan ibarət olmuşdur. İnsanlar gözləyirlər ki, alqoritmlər təkcə onların dediklərini deyil, həm də nə demək istədiklərini, niyyətlərini seçəcəklər. Bu fərq kifayət qədər çevik olmayan mövcud alqoritmlər üçün ciddi problem yaradır [11, 12].

Bütövlükdə cari analitik alqoritmləri və AS-in bəzi əsas problemlərini aşağıdakı şəkildə təsnifatlandırmaq olar:

– Müxtəlif tipli verilənlərin strukturuna uyğun olaraq tez adaptasiya olmur (məsələn, müxtəlif formatlı şəkillərin axtarışı, səsli axtarış və s.);

– Sorğunun mənasını başa düşmür, niyyətlə bağlı axtarışın nəticələrini qura bilmir (məsələn, axtarışın istifadəçinin niyyətinə və ehtiyaclarına uyğunluğu təmin olunmur və s.);

– Müxtəlif tipli, fərqli axtarış üslublarına uyğunlaşa bilmir (məsələn, müxtəlif dillərdə axtarışın aparılması və s.);

– Axtarışın nəticələrini tez və effektiv təsnifatlandırıb sıralaya bilmir (məsələn, təkrar başlıqları (*ing. Repetitive title*), eyni informasiyanı dəfələrlə göstərir, artıq "quyruqlar" əmələ gəlir, sənədlər məzmuna uyğun sıralanmır və s.) və s.

Hazırda dünya yenidən bir çox mürəkəb məsələlərin öhdəsindən gələ bilən süni intellekt (Sİ, *ing. **Artificial İntelligence***) **adlanan** sahəni tədqiq etməyə başlamışdır. Mürəkkəb məsələlərin həlli və insan əqlinin modelləşdirilməsi üçün insanın süni oxşarının yaradılması ideyası qədim zamanlardan mövcuddur. Elmi istiqamət kimi Sİ-nin yaranması yalnız XX əsrin 40-cı illərində kompüterin yaranmasından sonra baş verdi. Sİ, bir elmi istiqamət olaraq, müxtəlif elm sahələrinin inteqrasiyasına əyani nümunədir. O, riyaziyyat, kibernetika, proqramlaşdırma, linqvistika, biologiya, psixologiya kimi elm sahələrinin sintezidir. Sİ insan kimi düşünən və insan kimi qərar qəbul edə bilən texniki qurğu yaratmaq məqsədi daşıyır [13].

Sİ-nin alət və alqoritmləri yeni deyil. Onların əksəriyyəti 1960-cı və 1970-ci illərdə işlənmişdir. Maşın təlimi (MT, *ing. Machine Learning*) öz-özlüyündə Sİ-nin bir hissəsidir. Sİ və MT-nin vaxtilə İnternet dünyasının başlıqlarında olmamasının əsas səbəbi həmin zamanlarda hesablama gücünü nəzərə almaqla alqoritmlərin reallaşmasının məhdud imkanları ilə izah olunur. Verilənlərin saxlanması metodları ilə birlikdə kompüterlərin artan sürəti ona gətirib çıxardı ki, Sİ alqoritmləri bir çox məsələlərin həlli üçün seçilməyə başlandı. İnsanların düşüncəsində maşınla qarşılıqlı əlaqəyə münasibətdə dəyişikliklər baş verdi.

Reqressiya daxil olmaqla, hazırda istifadə edilən proqnostik modellərin əksəriyyəti artıq MT-nin hissələridir. Sİ MT ilə məhdudlaşmır. Reqressiya kimi sadə alqoritmlərin gündəlik həyatımıza tətbiqi öz effektliliyini sübut etmişdir. Reqressiya və modelləşdirmənin proqnostik gücü daha yaxşı analitik həllərə tələbat yaratmışdır.

Müasir AS kifayət qədər inkişaf etmişdir. Burada istifadə edilən mürəkkəb alqoritmlər AS-ə sorğunu tez qəbul etməyə və böyük həcmli informasiyalar arasından dəqiq nəticələri qaytarmağa imkan verir. Ümumiyyətlə, AS erkən prototiplərinin mövcud olduğu vaxtdan bu günə qədər uzun bir yol keçmişdir. Buraya İnternetdə veb-skanerlərdə, təsnifatlandırmada və indeksləşdirmədə yaxşılaşmaları, veb-ustaların veb-səhifələrin skan edilməsində istifadə etdikləri robots.txt kimi yeni protokolların tətbiqini aid etmək olar. Səsli axtarışın tətbiqi isə müxtəlif AS-də layihələndirilmiş bir neçə axtarış texnologiyalarının kulminasiyası oldu.

İnternetdə ümumdünya hörümçək toru yaranana qədər ilk axtarış alətlərindən biri 1990-cı ildə işlənib hazırlanmış *Archie* olmuşdur. O, *FTP* (*ing. File Transfer Protocol*) ehtiyatlarının siyahısını emal edir və faylların adlarına görə axtarış aparmağa imkan verir. İnternetdə informasiya axtarışını həyata keçirən *Web-Crawler*, *Excite*, *İnfoseek*, *Yahoo*, *Tradeware Galaxy* və s. AS-lər 1994-cü ildən fəaliyyətə başlayıblar. Bu gün *Yahoo* dünyanın ən məşhur AS-indən biridir. *Yahoo* özündə

Yahoo!Directory tematik kataloqunu, eləcə də İnternetin ilk və ən məşhur elektron poçtu olan *Yahoo!Mail* xidmətlərini birləşdirir. 1995-ci ildə istifadəyə verilən *Alta Vista* təbii dili sorğuların emalı üçün birinci AS idi. 1995-ci ildə meydana gələn *Lycos* AS də İnternetin ilk axtarış maşınlarından biridir. O, sorğuya uyğun siqnalları təsnifatlandıran, açar sözləri prefikslərlə və sözün yaxınlığıyla müqayisə edən sistemlə işə başladı. 1997-ci ildə *Ask Jeeves* redaktorlardan istifadə edərək istifadəçinin faktiki sorğularına uyğunluğunu təmin etmək üçün öz AS-ini təqdim etdi [5].

Bu gün 1999-cu ildən *Google* şirkəti tərəfindən öz xidmətlərini təqdim etməyə başlayan *Google* dünyanın ən tanınmış və geniş istifadə olunan AS-dir. Eyni zamanda hazırda daha geniş istifadə edilən AS bunlardır: *Microsoft Bing*, *Yahoo!*, *Ask*, *Baidu*, *Yandex*, *AOL* və s. Lakin *Google*-nin istifadəsi rəqibləri kölgədə qoyur. 2015-ci il fevral ayının *Google* qiymətləndirməsinə əsasən, o, hər ay 1,1 milyard unikal istifadəçi qazanır. *Bing* 350 milyonla 2-ci, *Yahoo!* 300 milyonla 3-cü, *Ask* isə 245 milyonla 4-cü yerdədir. Baxmayaraq ki, *Google* köhnə AS deyil, lakin o, bu gün ən məşhurlardan biri olmuşdur. *Google*-nin hər gün emal etdiyi verilənlərin həcmi təxminən 20 petabayt qiymətləndirilir. Bütün bu trafik *Google* üçün faydalıdır, onun gəlirlərinin əsas hissəsi reklamdan daxil olur [14, 15].

Böyük verilənlər yüksək keyfiyyətli onlayn AS-in inkişafına imkan vermişdir. Axtarış sorğuları əsasında işləyən AS çoxlu sayda sorğular emal etmək imkanına malik mürəkkəb alqoritmlər tələb edir. Nəzərə almaq lazımdır ki, AS verilənlər bazasının xususiyyətlərinə görə yaradılır. AS-lər müxtəlif olduğu üçün hər bir AS-in özünün problemləri var. Araşdırmalar göstərir ki, AS-də aşağıdakı əsas məsələlərin həlli təmin edilməlidir:

– Sorğuya ***relevantlıq***, axtarış nəticəsində tapılaraq istifadəçiyə təqdim edilmiş sənədlərin məzmununun informasiya sorğusunun məzmununa uyğun təqdim edilməsi;

– Sorğuya ***pertinentlik,*** axtarış nəticəsində tapılmış sənədlərin məzmununun istifadəçinin axtarış sorğusu şəklində formalaşdırılmış informasiya tələbatına uyğun axtarılıb tapılaraq təqdim edilməsi;

– Əldə olunmuş informasiyanın ***ümumiləşdirilməsi***, ***dəqiqləşdirilməsi*** və ***dolğunluğu***;

– Sorğuya ***repetitive***-lik **(**təkrarçılıq) təkrar başlıqların və artıq "quyruqların" aradan qaldırılması və s.

Mütəxəssislər axtarış obyektinin növündən, formasından, məzmunundan və formatından asılı olaraq, AS-də informasiya axtarışının aşağıdakı növlərini qeyd edirlər [5, 16]:

– *Tammətnli axtarış,* sənədlərin məzmunu (bütün mətni) üzrə axtarış prosesidir.

– *Meta-verilənlər üzrə axtarış*, sənədlərin AS tərəfindən qəbul edilən atributlarına və ya rekvizitlərinə (sənədin adı, yaradılma tarixi, ölçüsü, müəllifi və s.) görə axtarılması prosesidir.

– *Təsvirə görə axtarış*, məzmununa görə təsvirlərin axtarılması prosesidir.

– *Ünvanlı axtarış*, axtarış sorğusunda göstərilən formal əlamətlərə görə sənədlərin axtarılması prosesidir.

– *Semantik axtarış*, sənədlərin məzmununa görə axtarılması prosesidir.

– *Sənədli axtarış*, istifadəçinin sorğusuna uyğun olaraq, informasiya-axtarış sisteminin bazasında ilkin sənədlərin və ya verilənlər bazalarında ikinci dərəcəli sənədlərin axtarılması prosesidir.

– *Faktoqrafik axtarış*, axtarış sorğusuna uyğun olan faktların axtarılması prosesidir.

Virtual fəzada böyük verilənlərin həcminin və müxtəlifliyinin kəskin artması nəticəsində istifadəçinin tələbatına uyğun informasiyanın axtarılması və tapılması çox ciddi problemə çevrilmişdir. İnternetdə informasiyanın axtarılması və əldə edilməsinin aşağıdakı üsulları qeyd edilir [17]:

– URL-ünvanın daxil edilməsi;

– Hiperəlaqələr üzrə hərəkət (naviqasiya);

– AS-in istifadəsi və s.

Böyük həcmli verilənlər üçün genişmiqyaslı AS tələb olunur. Həmçinin, AS müxtəlif tip sorğuları yerinə yetirərək strukturlaşdırılmamış və strukturlaşdırılmış verilənlər üzrə axtarışı

həyata keçirmək imkanına malik olmalıdır. Hazırda böyük həcmli verilənləri axtara bilən bir neçə açıq mənbə kodlu alətlər var. Verilənlərin axtarışı real vaxt rejimində mümkündür. Açıq kodlu alətlərdən bəziləri haqqında aşağıda məlumat verilmişdir [18].

Lucene – Apache lisenziyalı axtarış alqoritmidir və yüksək məhsuldarlıqlı və miqyaslanan indeksləşdirmə təklif edir. O, həmçinin minimal yaddaş tələb edir. Alqoritm ranqlaşdırılmış axtarışı, sahələrin axtarışı, verilənlərin diapazon üzrə axtarışı, eləcə də bir neçə indeks üzrə axtarışı təklif edir. Bir neçə sorğu variantı mövcuddur və tamamilə *Java*-ya inteqrasiya olunur.

Apache Solr – müəssisənin *Apache* lisenziyalı *Java*-da yaradılmış AS-in avtonom serveridir. O, tammətnli axtarış serveri kimi işləyə bilər və faset axtarışı, dinamik klasterləşdirmə, real vaxtda indeksləşdirməyə yaxın və coğrafi məkan üzrə axtarış kimi funksiyalar təklif edir. Eyni zamanda miqyaslanan və imtinaya davamlıdır. *Solrdan* İnternetin mühüm AS-indən çoxu istifadə edir.

Elasticsearch – açıq mənbə kodlu *Apache Solrun* üstündə yaradılmış alətdir. Verilənlər HTTP vasitəsilə *JSON*-dan (*ing. JavaScript Object Notation*, insanlar tərəfindən asan oxuna bilən verilənlərin çevrilməsi üçün formatdır) istifadə etməklə indeksləşdirilir və miqyaslana bilir. *Elasticsearch* bir neçə indeks üzrə axtarışı yerinə yetirə bilər. İndekslər qalıqlara bölünür və onlar qovşaqlar üzrə paylanır. Bu, sürətli işi və asanlıqla yenidən balanslaşmanı və yönəldilməni təmin edir.

Süni intellekt və maşın təlimi – əsas trendlər

Sİ – müxtəlif mürəkkəb tətbiqi məsələlərin həlli üçün insan düşüncəsinə və ya canlı və cansız təbiətdə gedən proseslərə analoji prinsip və yanaşmalardan istifadə edilməklə işlənib hazırlanmış metodlar və alətlər məcmusudur. Bu, maşın formatında insan intellektidir, burada kompüter proqramları adətən insanın yerinə yetirdiyi tapşırıqları yerinə yetirirlər. Sİ informatika, psixologiya, fəlsəfə, linqvistika, iqtisadiyyat, optimallaşdırma, məntiq nəzəriyyəsi və bir sıra başqa sahələrə əsaslanan tədqiqat sahəsi olmaqla, informatikanın xüsusi bölməsidir.

Bu günədək Sİ sahəsindəki tədqiqatlar müxtəlif, məsələn, biliklərin təqdimatı, mühakimələrin modelləşdirilməsi, biliklərin əldə edilməsi, MT və verilənlərin intellektual təhlili və obrazların tanınması, qərar qəbulunun dəstəklənməsi, proseslərin və sistemlərin idarə edilməsi, dinamik intellektual sistemlər, planlaşdırma və s. istiqamətlərdə aparılmışdır [13].

MT Sİ-nin bir hissəsi olaraq verilənlər yığımındakı müxtəlif problemlərə tətbiq edilə bilər. *Dərin təlim* **(***ing. Deep Learning***)** isə MT-dən götürülmüş daha əhatəli alqoritmik yanaşmadır. O, nitqin və təsvirlərin tanınması və s. kimi məsələlərin yerinə yetirilməsində məntiqə əsaslanır və verilənləri çoxlaylı neyron şəbəkədən istifadəyə yönəldən metodlardan istifadə edir [19].

MT üçün verilənlərin çox olması yaxşı nəticə deməkdir, çünki yeni verilənlər kompüter proqramını təlimə və özünü təkmilləşdirməyə imkan verir. Bu, məşhur tədqiqat mövzusudur. Belə ki, MT-nin potensialı çoxdur. Bu, kompüterlərə prinsipial müxtəlif və daha güclü olmağa imkan verir, nəticədə yeni tətbiqlər əmələ gəlir. Verilənlərin intellektual analizi ilə MT arasında əlaqə mövcuddur. Belə ki, hər ikisi böyük verilənlərin analizi metodudur və şablonların axtarışına əsaslanır. Əsas fərq ondadır ki, verilənlərin intellektual emalı insanlara başa düşmək və istifadə etmək imkanı verir. MT isə bu şablonları öz proqramını və qavramanı yaxşılaşdırmaq üçün istifadə edir [20].

Hazırda MT İnternetdə səmərəli axtarışın aparılmasında, obrazların və nitqin tanınmasında və s. kimi bir çox tətbiqlərdə istifadə edilir. *Google*-nin *RankBrain* alqoritmi MT-yə nümunədir. O, açar sözlər və digər amillərin uyğunluğunun proqramlaşdırılmış qaydaları əsasında sadəcə nəticələr təqdim etmir, hər bir axtarış sorğusunun dolğunluğunu və kontekstini qiymətləndirir. Digər bir nümunə kimi *Facebook* sosial şəbəkəsinin xəbərlər axınını göstərmək olar. İstifadəçi nəyi bəyənirsə, *Facebook* ona həmin yönümlü reklamları təklif edir. Bütövlükdə MT alqoritmləri təlim üsuluna görə üç hissəyə bölünə bilər: *nəzarət edilən, nəzarət edilməyən və yarım-nəzarət edilən maşın təlimi alqoritmləri.*

Nəzarət edilən alqoritmlər giriş verilənləri və ya təlim verilənləri adlanan verilənlər yığımı tələb edir. Verilənlər yığımı nişanlara malik olmalıdır. Ən əsası odur ki, layihələndirilmiş təlim

verilənləri yığımının məqsəd və nəticələri bilinməlidir. Məlum olduğu kimi, bu halda alqoritmin istənilən nəticəsi yalan təsnifat və proqnoz edə bilər. Nəzarət edilən alqoritmlərə misal olaraq, reqressiya və ya neyron şəbəkələri göstərmək olar.

Nəzarət edilməyən təlim alqoritmləri nişanlanmamış verilənlər yığımı ilə işləyirlər. Verilənlər yığımının məqsəd və nəticəsi məlum deyil. Alqoritm naməlum nəticənin axtarışına əsaslanır. Alqoritmlərin əksəriyyəti verilənlər yığımında izafiliyi azaltmaqla problemi həll etməyə cəhd edirlər. Belə alqoritmlərə misal olaraq ölçünün azaldılması, assosiativ qaydalar, *Apriori, K-means* və s. göstərmək olar.

Yarım nəzarət edilən alqoritmlər onlar haqqında bəzi informasiyaya malik olduğumuz verilənlər yığımına əsaslanır. Nişanlar, nəticələr və məqsəd tam məlum deyil, lakin nəticənin necə göründüyünə dair nümunələr var. Modellər verilmiş nümunələri öyrənməyə davam etməli və bütün verilənlər yığımı üçün nəticələri tapmalıdırlar [21].

Xətti reqressiya (*ing. Linear regression*), Loqistik reqressiya (*ing. Logistic regression*), Qərarlar ağacı (*ing. Decision Tree*), Naive Bayes (*ing. Naive Bayes*), K-yaxın qonşu (*ing. K-Nearest Neighbor*), Dayaq vektor maşınları (*ing. Support Vector Machines*), Təsadüfi meşə (*ing. Random Forest and Bagging*), Qradiyent bustinqi (*ing. Gradient Boosting*), Apriori (*ing. Apriori*), Neyron şəbəkələri (*ing. Neural networks*) hazırda tez-tez istifadə edilən MT metodlarındandır.

Axtarış sistemlərində maşın təlimi metodlarının tətbiqləri

Tədqiqatlar göstərir ki, MT alqoritmləri növbəti 10 il ərzində bütün dünya üzrə 25% işçi yerini əvəz edəcəkdir. MT proqramları özləri modifikasiya olunur, insanın minimal müdaxiləsi ilə zamanla yaxşılaşırlar. Hazırda *Google, Microsoft Bing, Yahoo!, Ask, Baidu, Yandex* kimi AS-lər Sİ-yə əsaslanan alqoritmlərdən istifadəyə yönəliblər [22].

İntellekt tələb olunan hər şey çox vaxt MT alqoritmlərinin köməyi ilə həll olunur. Məsələn, *Google RankBrain* (hərfi tərcümədə – ranqlaşdırıcı intellekt, *Hummingbird* – ranqlaşdırma alqoritmi) – Sİ-yə əsaslanan özüöyrənən sistemdir. Onun istifadəsi 26 oktyabr 2015-ci ildə *Google* AS tərəfindən təsdiq edilmişdir. O, *Google* axtarış nəticələrinin emalını dəstəkləyir və istifadəçilər üçün daha müvafiq nəticələr təmin edir. Belə ki, əgər *RankBrain* sözü və ya ifadəni görürsə, lakin onunla tanış deyilsə, alqoritm hansı söz və ya ifadələrin analoji məna daşımasına dair fərziyyə edə və nəticəni filtrləşdirə bilər, bu da onu istifadəçilərin hələ heç vaxt vermədiyi axtarış sorğularının emalında daha effektiv edir. *RankBrain* keçmiş axtarış sorğuları haqqında verilənləri yığır və onları analiz etməklə axtarışın nəticələrini necə sazlamağı bilir və real vaxt rejimində yenidən işləyir. Bundan başqa, digər əksər AS *Google*-nin axtarış şablonundan istifadə edir. Əgər səhifə *Google*-də optimallaşdırılıbsa, o, digər əksər AS üçün optimallaşdırılmış sayılır.

Bu gün bəzi transmilli şirkətlər AS-də olan problemlərin aradan qaldırılması üçün MT metodlarından istifadə edirlər. Məsələn, *Amazon Echo Aleks* ingilis dilində səsli informasiyanı başa düşən qurğudur, virtual köməkçidir. O, real vaxtda xəbərlər, musiqi, hava haqqında məlumatlar, həyəcan siqnalının qurulması, audio kitabların yaradılması və s. kimi səsli cavab vermə imkanlarına malikdir.

SIRI (*ing. Speech Interpretation and Recognition Interface*) Apple-nin iOS əməliyyat sistemi üçün nəzərdə tutulmuş fərdi ağıllı assistent və naviqasiya proqramıdır. *SIRI* – Süni İntellekt Beynəlxalq Mərkəzinin layihəsidir, hazırda bəlkə də ən böyük Sİ layihəsi olaraq ABŞ-ın Müdafiə Nazirliyinin Perspektiv Tədqiqat layihələri Agentliyi (*ing. Defence Advanced Research Projects Agency*) tərəfindən maliyyələşdirilir. 11 iyun 2014-cü ildən Apple Rus dilində də layihələndirib, istifadəyə verildi. *SIRI,* Azərbaycan dili istisna olmaqla, 21 dili başa düşür.

Beləliklə, MT AS-də aşağıdakı məsələlər üçün istifadə olunur:

1. *Axtarışın ranqlaşdırılması*. Çox vaxt AS başlanğıc axtarış, ilkin reytinq, kontekstli ranqlaşma, fərdiləşdirilmiş reytinq və s. kimi ardıcıl baş verən bir neçə ranqlaşdırma fazasına malik olur. MT bütün bu mərhələlərdə ranqlaşdırma üçün istifadə olunur.

2. *İstifadəçi sorğularının dəqiqləşdirilməsi.* MT-dən istifadəçinin daxil etdiyi axtarış sorğusunun mənasının başa düşülməsi üçün istifadə olunur. Bunun bəzi nümunələri:

– Sorğuların təsnifatı: AS axtarış sorğusunda müxtəlif təsnifatlandırıcılardan istifadə edir. Məsələn, naviqasiya, informasiya və tranzaksiya sorğularının aşkarlanması və s.

– Orfoqrafiyanın təklifi və ya düzəlişi;

– Sinonimlər/sorğunun genişləndirilməsi: AS-də sorğunun açar sözlərinin və nəticələr yığımının genişləndirilməsi üçün sinonimlərdən istifadə edilir.

– Axtarışın istifadəçinin niyyətinə və ehtiyaclarına uyğunluğu. Məsələn, "Qartal" sözünü axtarışa verdikdə, bu, qrup, quş, ölkə adı və s. ola bilər.

3. *URL-ünvanlı axtarış***.** Bu, URL-ünvanın, yəni axtarışın nəticəsinin qavranması üçün edilən hər şeyi özünə daxil edir və çox vaxt AS URL-ni skanlaşdıranda/indeksləşdirəndə yerinə yetirilir. Bunun bəzi nümunələri:

– Səhifələrin təsnifatı;

– Spamın aşkarlanması;

– URL-ünvanların arzuolunmayan/keyfiyyətsiz aşkarlanması;

– Əhval-ruhiyyələrin analizi;

– Münasibətlərin aşkarlanması.

4. *AS-də göstərilmiş digər kontentin yaradılması.* Məsələn, oxşar sorğular, sayta istinadlar və s.

5. *Skanerlər, MT-dən optimal sürətin təyin edilməsi.*

6. *İstifadəçilərin təsnifatı və müəyyənləşdirilməsi*. Bu, xüsusilə fərdiləşdirilmiş axtarış üçün faydalıdır .

Axtarış sistemləri üçün maşın təlimi metodlarının işlənməsi: qısa icmal

AS-i məşhur edən odur ki, o, istifadəçilərə intuitiv, sadə şəkildə, yəni sorğu sahəsinə bir neçə açar söz göndərməklə veb-interfeyslə sorğu etməyə imkan verir. Lakin elə hallar olur ki, İnternet sorğuları ilə bağlı iddia edilən sadəliyə baxmayaraq, genişmiqyaslı AS istifadəçilərin informasiya ehtiyaclarını ödəyə bilmir və nəticədə sorguların yenidən formalaşmasına çox vaxt sərf edilir. Yəni, məşhurluqlarına baxmayaraq, AS-lər çox vaxt istifadəçilərin informasiya tələbatını, xüsusilə də pis formalaşdıqda, ödəmirlər. Məsələn, AS istifadə olunan terminologiyanı aydın şəkildə bilməyən istifadəçilərə az kömək edir. Baxmayaraq ki, tədqiqatçılar lazım olan məlumatları vaxtlarını sərf etməklə tapmağa müyəssər olurlar.

AS-in əsasını istifadəçilərin axtarış niyyətlərini nəzərdə tutan sorğuların qurulması təşkil etməlidir. İnformasiya axtarışının konkret rejimlərində AS-lər tam və xüsusi sorğulara baxa bilmirlər. Bu problemin əsas səbəblərindən biri, axtarış qurğusu ilə müvafiq AS arasında bir neçə qarşılıqlı əlaqədən ibarət, ilk və son veb-axtarış seansı arasındakı semantik boşluğun olmasıdır. Qrafik interfeys baxımından, AS-də bu problemin həlli giriş pəncərəsinin funksionallığını genişləndirməklə həyata keçirilir. Hazırda İnternetdə aparıcı AS, məsələn, *Google* və *Yahoo!*, eyni hərflərdən başlayaraq, giriş sahəsindəki axtarış sorğularının tipləri üçün cümlələr təklif edir. Belə cümlələr adətən populyarlıqlarına görə qiymətləndirilirlər.

Qeyd etmək lazımdır ki, istifadəçilərin yüksək tələbatlarını ödəmək və axtarışın intellektual səviyyəsini inkişaf etdirmək üçün elmi ədəbiyyatda müxtəlif MT metodlarına əsaslanan AS təklif edilir. Buna görə də aparılan tədqiqatlar, istifadəçilər də daxil olmaqla, veb- axtarış prosesi üzrə aparılır. Məqsəd AS və istifadəçi sorğuları arasındakı uyğunsuzluğu azaltmaqdır. İstifadəçi sorğularının fərdi şəkildə emalı, istifadəçi haqqında mövcud biliklərin yenilənməsi, informasiya axtarışının effektivliyi, axtarışın istifadəçinin niyyətinə uyğunluğu və istifadəçi məmnunluğu bu sahədə həll edilməli əsas problemlərdir.

[23]-də təkilf olunan metod istifadəçilərin informasiya ehtiyaclarını müəyyənləşdirmək üçün dəqiq olan terminlərin seçilməsində kömək edir. Təklif olunan yanaşmanın faydalılığı qiymətləndirilir. Bir çox tədqiqatlarda semantik AS sahəsində son nailiyyətlər, biliklərin toplanması üsulları və informasiyanın axtarış metodları müzakirə olunur. Eyni zamanda

informasiya axtarışında problemləri həll etmək üçün bilik agenti əsasında AS-dən hansının daha yaxşı olması təhlil edilir. Hazırkı AS-də bilik agentlərinin problemləri ətraf mühiti kifayət qədər qavrayıb, onu anlamaq, bilikləri toplamaq və böyük həcmli verilənlərdən informasiya axtarmaqdır. Ümumilikdə, məqsəd axtarış təcrübəsini yaxşılaşdırmaq, təkrarlanan sorğuların sayını azaltmaq və istifadəçilər üçün onların maraqlarına, niyyətinə uyğun yüksək nəticələrin əldə edilməsini asanlaşdırmaq, AS-in səmərəliliyini artırmaqdır [24].

Elmi ədəbiyyatda biliklərin toplanması, əldə edilməsi, informasiyanın axtarılması üçün Sİ-dən istafadə etməklə AS-i yaxşılaşdıra bilən bəzi problemlərin həlli göstərilir. Təbii dillə ifadə olunan istifadəçi suallarını, yəni AS-ə verilmiş sorğuları avtomatik olaraq anlamaq, cavablamaq informasiya axtarışı və Sİ tədqiqat sahələrində vacib və çətin problemə çevrilib. Adi interaktiv veb-axtarışda, məsələn, seans axtarışında istifadəçi uyğun informasiyanı əldə etmək üçün adətən AS ilə bir neçə dəfə müxtəlif formalarda, məsələn sorğu formalarında, klikləmə və s. şəklində qarşılıqlı təsir göstərir. Bu təsirlər adətən qarşılıqlı olaraq mürəkkəb şəkildə bir-birilə bağlıdır.

İdeal məqsəd üçün ağıllı AS bu təsirlərdən istifadəçiyə lazımlı informasiyanı əldə edə bilən Sİ agenti kimi görünə bilər. Ancaq hələ də texnikanın indiki vəziyyəti ilə bu məqsəd arasında böyük bir boşluq mövcuddur. [25]-də boşluğun aradan qaldırılması, sorğular, kliklənmiş sənədlər arasındakı qarşılıqlı asılılıqların nəzərə alınması, sorğuların avtomatik genişlənmələri üçün Markov metodu əsasında yanaşma təklif olunur. Empirik qiymətləndirmə genişmiqyaslı veb-axtarış verilənlər yığımı üzərində aparılır və nəticələr təklif olunan modellərin effektivliyini göstərir.

Mobil axtarışların artan populyarlığı və səs tanıma texnologiyalarının inkişafı veb-axtarış istifadəçilərinin sorğularının yazıdan danışığa (səsə) keçidinə yol açdı. Bu cür səsli axtarışların ilkin inkişaf mərhələsində olmasına baxmayaraq, onlar tədricən daha geniş yayılmaqdadır. [26]-da kommersiya veb-axtarış sisteminin mobil tətbiqində sorğuların səsli axtarışla təhlili təqdim olunur. Sorğuların semantik və sintaktik xüsusiyyətlərinə xüsusi diqqət yetirilərək səs və mətn axtarışı müxtəlif aspektlərdə müqayisə edilir. Tədqiqatda göstərilir ki, səsli sorğular daha çox audio-vizual məzmunlara və sorğu cavablarına (sual-cavaba) fokuslanır və daha az sosial şəbəkə üçün istifadə edilir. Bundan əlavə, səsli sorğular daha çox birbaşa təqdim olunur, yəni göndərilir. Burada, həmçinin səsli sorğuların dilinin mətn sorğularının dilindən daha çox təbii dilə yaxın olduğunu göstərən empirik qiymətləndirmə aparılır. Təhlillər səs və mətn axtarışları arasındakı fərqlilikləri göstərir. Gələcək səsli veb-axtarış vasitələrinin dizaynı üçün bu fərqlərin nəticələrinin müzakirəsi aparılır.

Araşdırmalar göstərir ki, uzun sorğulardan istifadə etməklə lazımlı nəticələrin alınmaması AS-in əsas problemlərindən sayılır. Uzun sorğular cari veb-proqramlarda, eləcə də ədəbiyyat axtarışlarında, xəbər axtarışlarında və s. geniş istifadə olunur. Lakin uzun sorğular tez-tez açar sözlər kimi deyil, təbii dil mətnləri kimi ifadə olunur. Cari açar sözlərə əsaslanan *Google* kimi AS qısa sorğularla müqayisədə uzun sorğular üçün pis nəticə göstərir.

[27]-də uzun sorğulu veb-axtarışlarda sorğunun qurulması və sorğu nəticələrinin dəyişdirilməsi üçün yanaşma təklif edilir. Bu yanaşmada, ilk olaraq istifadəçilərin sorğu tarixindən uzun sorğulara əsasən bir neçə qısa sorğu alınıb. Daha sonra, qısa sorğuların klasterləri qurulub və əsl (orijinal) uzun sorğunu dəyişdirmək və əvəz etmək üçün ən uyğun olan sorğu seçilib. Bununla yanaşı, göstərilir ki, müvafiq qısa sorğuların axtarışları əsl uzun sorğudakı məzmun və şərtləri görməyə bilər və nəticədə müxtəlif nəticələr və məlumatlar əldə edilə bilər. Ona görə də əsl uzun sorğuların məzmunları ilə axtarış nəticələrinin məzmunları müqayisə edilib və uyğun olmayan nəticələr ayırılaraq filtrlənib. Bu yanaşma uzun sorğulu veb axtarışlar üçün eksperimentlər vasitəsi ilə göstərilmişdir.

Bəzi tədqiqatlarda səsli və yazılı sorğuların keyfiyyət və axtarış effektivliyi baxımından müqayisəsi aparılıb. [28]-də müəlliflər səsli (bəzi ədəbiyyatda səsli sorğu, danışıq sorğuları kimi verilir) və yazı sorğuları arasında fərqlər barədə eksperimental araşdırma haqqında məlumat verirlər. İstifadəçilərin yazılı mövzuları yazılı və səsli sorğular yığımından yaranır.

Bəzi tədqiqatlarda isə bu iki sorğu dəsti keyfiyyət və onların axtarış effektivliyi, uzunluq,

müddət və nitq hissəsi baxımından müqayisə edilir. Əlavə olaraq müqayisədə səsli sorğuların mükəmməl transkripsiyası nəzərə alınaraq, yazılı və səsli sorğulara uyğun sənədləri təsvir etmək xüsusiyyətinə baxılır. Səsli və yazılı sorğuların axtarış effektivliyi üç müxtəlif informasiya axtarışı modeli ilə müqayisə edilir. Nəticələr göstərir ki, danışıq (nitq) öz informasiya ehtiyacını formalaşdırmaq üçün ona daha təbii şəkildə ifadə etmək imkanı verir və daha uzun sorğuların formalaşdırılmasına kömək edir. Buna baxmayaraq, uzun səsli sorğular yazılı sorğularla müqayisədə axtarışın effektivliyini əhəmiyyətli dərəcədə yaxşılaşdırmırlar.

Nəticə

Bu gün verilənlərin həcminin çox böyük sürətlə artması, onların real vaxtda analizi, AS-də sorğuların səmərəli emal edilməsinin zəruriliyi yeni alətlərin yaradılması və tətbiqini tələb edir. Verilənlərin həcminin, xüsusi halda qrafik prosessorlarda hesablama gücünün, artması ilə Sİ (maşın təlimi metodları) böyük həcmli verilənlər üçün intellektual analitik həllərin təqdimatında həlledici rol oynamağa başlayır.

Son onillikdə AS-də informasiyanın axtarılması üçün Sİ-yə əsaslanan alqoritmlər geniş tətbiq edilir. Bu alqoritmlər axtarışın ranqlaşdırılması, sorğuların təsnifatı, axtarışın istifadəçinin niyyətinə və ehtiyaclarına uyğunluğunun təmin edilməsi, əhval-ruhiyyələrin analizi, istifadəçilərin təsnifatı və müəyyənləşdirilməsi, sorğunun genişləndirilməsi və s. kimi məsələlər üçün istifadə olunur. Bunu nəzərə alaraq, məqalədə axtarışın keyfiyyətinin yaxşılaşdırılması, verilənlərin həcminin böyüklüyü, istifadəçi sorğularının axınının intensivliyi ilə bağlı məsələlərin həlli zamanı Sİ-yə əsaslanan alqoritmlərdən istifadənin zəruriliyi göstərilmişdir.

Ədəbiyyat

1. Qasımova R.T. “Rəqəmsal irs: problemlər və perspektivlər”. Ekspress-informasiya. İnformasiya cəmiyyəti seriyası, Bakı: “İnformasiya Texnologiyaları” nəşriyyatı, 2018, 148 s.
2. Хартия о сохранении цифрового наследия // Библиотековедение, 2004, №6, с.40–43.
3. Brian R. Digital Access to Cultural Heritage and Scholarship in the Czech Republic // Slavic & East European Information Resources, 2008, vol.9, no.1, pp.12–29.
4. Tallova L. Copyright aspects of disclosure of works within the European Digital Library / Proceedings of the International Multidisciplinary Scientific Conferences on Social Sciences and Arts, 2014, vol.1, pp.561–568.
5. Qasımov V.Ə. İnformasiya axtarışı üsulları və sistemləri. Dərslik. Bakı: MTN-in Maddi-texniki Təminat Baş İdarəsinin Nəşriyyat-Poliqrafiya Mərkəzi. 2015, 288 s.
6. Reinsel D., Gantz J., Rydning J. Data Age 2025: The Digitization of the World – From Edge to Core, November 2018, IDC White Pape. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf
7. Рост объема информации – реалии цифровой вселенной // Журнал Технологии и средства связи, №1, 2013, с.24. http://lib.tssonline.ru/articles2/fix-corp/rost-obema-informatsii--realii-tsifrovoy-vselennoy
8. Alguliyev R.M., Gasimova R.T., Abbaslı R.N. The Obstacles in Big Data Process // International Journal of Modern Education and Computer Science (IJMECS), 2017, vol. 9, no.3, pp.28–35. DOI: 10.5815/ijmecs.2017.03.04.
9. Qasımova R.T. Big data analitikasi: mövcud yanaşmalar, problemlər və həllər // İnformasiya Texnologiyaları Problemləri, 2016, №1, s.75–93.
10. Madden S. From Databases to Big Data // IEEE Internet Computing, 2012, vol.16, no.3, pp.4–6.
11. By Research Voicebot and PwC. Smart speaker consumer adoption report. March 2018. https://voicebot.ai/wp-content/uploads/2018/03/smart_speaker_consumer_adoption_report_2018.pdf

12. A guide to the security of voice-activated smart speakers An ISTR Special Report Analyst: Candid Wueest. https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-security-voice-activated-smart-speakers-en.pdf
13. Balayev R.Ə., Əlizadə M.N., Musayev İ.K. İntellektual sistemlər və texnologiyalar. Dərs vəsaiti, Bakı: "MSV NƏŞR" nəşriyyatı, 2016, 256 s.
14. The history of search engines. https://www.wordstream.com/articles/internet-search-engines-history
15. Anderson A., Semmelroth D. Statistics for Big Data For Dummies, 2015, 384 pages, e-book: http://www.dummies.com/programming/big-data/data-science/big-data-and-search-engines/
16. Касумов В.А. Методы информационного поиска в Internet на основе нечетких отношений предпочтения // Автоматика и вычислительная техника, 2003, №4, с.71–78.
17. Касумов В.А. Методы построения информационно-поисковых систем на базе иерархической модели информационного пространства Интернет // Автоматика и вычислительная техника, 2002, №1, с.40–51.
18. Big Data Search Tools. https://datafloq.com/big-data-open-source-tools/os-big-data-search Qiu J., Wu Q., Ding G., Xu Y., Feng S. A survey of machine learning for big data processing // EURASIP Journal on Advances in Signal Processing, 2016, pp.1–16.
19. Aliguliyev R.M. Analysis of hyperlinks and the ant algorithm for calculating the ranks of web pages" // Automatic Control and Computer Sciences, 2007, vol.41, no.1, pp.44–53.
20. Cambazoglu B.B., Aykanat C., Baeza-Yates R. A machine learning approach for result caching in web search engines // International Journal of Information Processing and Management, 2017, vol.53, no.4, pp.834–850.
21. Chen H. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms // Journal of the American Society for Information, 1995, vol.46, no.3, pp.194–216.
22. Papadakis I., Stefanidakis M., Stamou S., Andreou I. Semantifying queries over large-scale Web search engines // Journal of Internet Services and Applications, 2012, vol.3, no.3, pp.255–268.
23. Meenakshi S. P., Agarwal G., Bakshi S., Bhatter S., Sivakumar P. Cognitive Agents for Web Based Search Engines: A Review / Proceedings of the Second International Conference on Recent Trends and Challenges in Computational Models, 2017.
24. Xiaozhao Z., Peng Z., et al. Modeling multiple interactions with a Markov random field in query expansion for session search // Computational Intelligence, 2018, vol.34, no.1, pp.345–362.
25. Guy I. The characteristics of voice search: comparing spoken with typed-in mobile Web search queries // ACM Transactions on Information Systems, 2018, vol.36, no.3, pp.1–28.
26. Chen Y., Zhang Y.Q. A query substitution-search result refinement approach for long query web searches / Proceedings of the International Joint Conferences On Web Intelligence (Wi) And Intelligent Agent Technologies (Iat), IEEE/WIC/ACM, 2009, vol.1, pp.245–251.
27. Crestani F., Du H. Written versus spoken queries: A qualitative and quantitative comparative analysis // Journal of the American Society for Information Science and Technology, 2006, vol.57, no.7, pp.881–890.

УДК 004.02

Касумова Рена Т.[1], Аббаслы Рахим Н.[2]
[1]Институт Информационных Технологий НАНА, Баку, Азербайджан
[2]Thinking capital, Монреаль, Квебек, Канада
[1]renakasumova@gmail.com, [2]rahim.abbasli@gmail.com

Анализ поисковых алгоритмов, используемых в больших данных

Цифровые материалы включают постоянно растущие текстовые документы, базы данных, структурированные и неструктурированные изображения, звуковые и графические материалы, программное обеспечение и веб-страницы. Увеличение темпов создания цифровой информации привело к необходимости анализа структуры входных файлов и более быстрого поиска и обработки. С этой целью, исследованы свойства алгоритмов искусственного интеллекта в анализе нетрадиционно структурированных больших данных. Было установлено, что необходимо использовать алгоритмы на основе искусственного интеллекта для решения проблем, связанных с улучшением качества поиска, увеличением объема данных и интенсивности пользовательских запросов. Также анализируются алгоритмы поиска, их недостатки и возможные варианты использования для их применения с целью максимизации их преимуществ.

***Ключевые слова**: цифровое наследие, цифровые данные, большие данные, аналитика больших данных, поисковые системы, поиск информации, искусственный интеллект, машинное обучение.*

Rena T. Gasimova[1], Rahim N. Abbaslı[2]
[1]Institute of Information Technology of ANAS
[2]Thinking capital, Montreal, Quebec, Canada
[1]renakasumova@gmail.com, [2]rahim.abbasli@gmail.com

Analysis of the search algorithms utilized in big data

Digital materials include continuously growing text documents, databases, structured and unstructured image, sound and graphic materials, software and web pages. Increasing pace of the generation of digital information has brought a need to analise the structure of the input files and create relevant and meaningful output faster. The article explores the features of search algorithms, their shortcomings and potential use cases for their application in order to maximize their advantages. It is found that it is necessary to use the algorithms based on artificial intelligence to solve problems associated with improving the quality of the search, increasing the amount of data and the intensity of user queries. The article analyzes the search algorithms, their shortcomings and potential use cases for their application in order to maximize their advantages.

***Keywords**: digital heritage, digital data, Big data, Big Data Analytics, search engines, information retrieval, Artificial Intelligence, machine learning.*