Rəqəmsal Kitabxanalarda Elmi Verilənlərin Çıxarılması Problemləri

Məkrufə Hacırəhimova[1], Aybəniz Əliyeva[2]
[1,2]İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
[1]*makrufa@science.az,* [2]*aybeniz63@rambler.ru*

***Xülasə* — Məqalədə rəqəmsal kitabxanalar, kitabxanalarda böyük həcmli verilənlərin toplanması, axtarış sistemləri üçün vacib olan metaverilənlərin və sənədlərdən elmi informasiyanın çıxarılması məsələsinə baxılmışdır. Həmçinin onların istifadəsi zamanı meydana çıxan problemlər nəzərdən keçirilmişdir.**

Açar sözlər — Big Data; rəqəmsal kitabxanalar; elmi verilənlər; informasiyanın çıxarılması; metaverilənlər

I. GİRİŞ

Elm, texnika və texnologiyaların geniş istifadəsi elm sahəsində də böyük verilənlərin artmasına və verilənlərin intensiv istifadəsinə əsaslanan elm (*data-intensive science*) sahələrinin inkişafına səbəb oldu [1]. Elmi tədqiqatların 4-cü paradiqmasına uyğun olaraq, elmi tədqiqatlar müasir elmi kəşflərin əsas mənbəyi kimi çıxış edən elmi verilənlərdən daha da asılı olmağa başladı [2]. Elmi verilənlər jurnal məqalələri, konfrans materialları, tezis, kitab, patent, təqdimat slaydları və s. kimi elmi verilənlərin böyük kəmiyyətini özündə birləşdirir. Böyük ölçülü elmi verilənlər (*scolary big data*) Big Data-nın həcm (*volume)*, sürət (*velocity*), müxtəliflik (*variety*), dəyər (*value*) və həqiqilik (*veracity*) xüsusiyyətlərinə malikdir. Elmi verilənlərin həcm və sürəti durmadan artır. 2010-cu ildə Microsoft Academic şəbəkəsində elmi sənəd qeydlərinin sayı 50 milyondan çox olmuş, 2008-2011-ci illər ərzində dərc olunmuş məqalələrin orta hesabla 43%-i İnternet vasitəsilə sərbəst girişə malik olmuşdur [3]. 2013-cü ildə İnternet şəbəkəsində yerləşən 114 milyondan çox ingilis dilli elmi sənəd və ya onların qeydlərinin 24%-i, yəni 27 milyonu istifadəçilər üçün əlyetərli olmuşdur [4]. Bu verilənlər maliyyə, təhsil və idarəetmə müəssisələrində qərarların qəbul edilməsində iştirak edən qruplarla yanaşı elm, biznes qurumlarının və geniş ictimaiyyətin maraq dairəsindədir. Elmi böyük verilənlərin ölçüsünü, həmçinin onlara girişin əldə edilməsi və istifadəsiylə bağlı maraqları nəzərə alaraq bu verilənlərin toplanması, analizi və onlara əlyetərliyin təmini ilə bağlı bir sıra xidmətlər yaranmışdır. Böyük ölçüdə elmi verilənlərin toplanmasını, analizini və istifadəçilər üçün əlyetərliyini təmin edən Google Scholar, PubMed, ArXiv və CiteSeerX kimi rəqəmsal kitabxanalar və arxivlər yaranmışdır [5].

Bu xidmətlər üçün vacib məsələlər böyük ölçülü verilənlərin toplanması, bir neçə mənbədən yığılmış informasiyanın inteqrasiyası və verilənlərdən əhəmiyyətli informasiyanın çıxarılmasıdır. Microsoft Academic Search və Google Scholar kimi bazalar və DBLP (*Digital Bibliography & Library Project*) kimi rəqəmsal arxivlər nəşrləri, biblioqrafik kolleksiyaları və metaverilənləri istifadəçilərə təqdim edir [6]. ArXiv (arXiv.org) kimi bəzi bazalar İnternetdən toplanmış sənədləri və metaverilənləri birbaşa insanlara təqdim etməyə imkan verir. 4 milyondan çox elmi sənəddən ibarət CiteSeerX isə avtomatik olaraq İnternetdən sənədləri toplayır və informasiyanın avtomatik çıxarılmasını həyata keçirir. Bu avtomatlaşdırılma prosesinin üstünlüyü yeni elmi verilənlərin toplanması və emalı baxımından daha yaxşı miqyaslanma və böyük verilənlərə tətbiq edilmə imkanına malik olmasıdır [4].

Bəşəriyyətin üzləşdiyi bir çox problemlərin həlli yalnız böyük elmi verilənlərdən lazımlı olan informasiyanın, biliyin və məlumatın əldə edilməsi və mübadiləsi sayəsində mümkünür. Bu cəhətən rəqəmsal kitabxanalarda toplanmış elmi sənədlər və metaverilənlər elmi-tədqiqatların səmərəliliyini təmin edə bilər. Lakin kitabxana kolleksiyalarının ölçüsünün daim genişlənməsi elmi verilənlərin toplanması, lazımi elmi informasiyanın çıxarılması, elmi-tədqiqat və birgə əməkdaşlıq sahəsində bir sıra problemlərin (informasiyanın çıxarılması, sənədlərin surətinin çıxarılması, qarşılıqlı effektiv əlaqənin təmini və s. bağlı) yaranmasına səbəb olur. Bu problemlərin araşdırılması aktual məsələlərdəndir.

Tədqiqat işində rəqəmsal kitabxanalar, kitabxanalarda böyük həcmli verilənlərin toplanması, elmi sənədlərdən metaverilənlərin və elmi informasiyanın çıxarılması, istifadəsi zamanı meydana çıxan problemlər nəzərdən keçirilir.

II. E-KİTABXANALARDA ELMİ VERİLƏNLƏRİN TOPLANMASI

İnternetdəki bütün elmi məqalələrin toplanması və vahid şəkildə emalı rəqəmsal kitabxanaların əsas problemlərindəndir. Elmi kitabxanaların məqsədi yalnız elmi materialların toplanmasından ibarət olduğundan, qeyri-elmi sənədlərin filtrlənməsi (filtrasiyası) tələb olunur. Kitabxanalarda elmi verilənlərin toplanması müxtəlif üsullarla həyata keçirilir.

Nümunə üçün, 2008-ci ildən ABŞ-ın Pensilvaniya Ştat Universitetində fəaliyyət göstərən CiteSeerX rəqəmsal kitabxanasında elmi verilənlərin toplanması məqsədilə PDF-fayllarından və istifadəçi məlumatlarının təqdim olunduğu URL-ünvanlardan ibarət siyahı yaradılır. Bu siyahı skan edilmə tarixi və elmi PDF fayllarının URL-ünvanları əsasında yaradılır. Gün ərzində kitabxanada orta hesabla 50-100 minlərlə PDF-faylı skan edilir. Skan edilmiş PDF sənədlərinin elmi olduğunu müəyyən etmək üçün onların filtrlənməsi həyata keçirilir. Nəticədə bu sənədlərin yalnız 40%-ə yaxını elmi verilənlər kimi qəbul edilir və verilənlər bazasına daxil edilir [3,4,7].

III. İNFORMASİYANIN AXTARILMASI

Elmi sənələrdən metaverilənlərin və digər informasiyaların çıxarılması rəqəmsal kitabxanalar, axtarış sistemləri və sənədlərin idarəetmə sistemləri üçün ümumi məsələdir. İnformasiyanın çıxarılması formaları kitabxana xidmətlərinin mühüm məsələlərdən olub, verilənlərin axtarışı və toplanması üçün istifadə olunan metaverilənlərin avtomatik çıxarılmasınn əlverişliliyinə və keyfiyyətinə təsir edir. Hal-hazırda rəqəmsal kitabxanalar informasiyanın çıxarılmasının bir neçə müxtəlif modulunu özündə birləşdirir.

A. Başlığın çıxarılması

Məqalələr haqqında olan metaverilənlər rəqəmsal kitabxana tərəfindən çıxarılan informasiyanın ən mühüm növlərindən biridir. Xüsusən də rəqəmsal kitabxanalarda filtrlənmədən keçən hər bir sənədin adı, müəllifləri, xülasəsi, yeri, həcmi və buraxılışı (jurnal üçün), səhifə nömrələri, nəşriyyat və nəşriyyatın ünvanı ilə bağlı informasiyaların (metaverilənləri) çıxarılmasına böyük dəyər verilir [4].

Hazırda elmi sənədlərdən metaverilənlərin çıxarılması üçün bir çox xidmətlər mövcuddur. Nümunə üçün, CiteSeerX rəqəmsal kitabxanasında başlığın çıxarılması vasitələri içərisində daha dəqiq hesab olunan SVM HeaderParser alətindən istifadə olunur. Bu model mənaca asılı sözlər üçün klasterizasiya üsulunun qaydalarını sözün spesifik xüsusiyyətlərinin (*word-specific feature*) yaradılması üçün istifadə etməklə PDF sənədlərinin mətn məzmununun səciyyəvi xüsusiyyətlərini çıxarır [4]. O mətn məzmununu başlığın metaverilənlərinə (məs., ad, müəllif və s.) uyğun bir neçə sahə üzrə təsnif edir. Bütün proses xüsusiyyətin çıxarılması, sətrin (xəttin) təsnifatı və metaverilənlərin təsnifatı kimi mərhələdən ibarətdir. Bu mətn çıxarıcısının (*extractor*) dəqiqliyi 92,9% təşkil edir.

Belə xidmətlərin digər bir nümunəsi GROBID yüksək məhsuldarlıqlı proqram təminatıdır. O, başlığın, istinadların metaverilənlərini çıxarmaq və analiz etmək imkanına malikdir. Bu alət müxtəlif sənədlərdən məzmunun avtomatik çıxarılması üçün CRF (*Conditional Random Fields*) maşın təlimi üsulundan istifadə edir [8].

B. İstinadların çıxarılması

İstinadlar elmi sənədlər üçün mühüm rol oynayır. ParsCit proqram paketi sətrin sintaktik analiz alətindən istifadə etməklə rəqəmsal kitabxanada tələb olunan hər bir məqalə üçün istinadlar çıxarıla bilir. Bunun üçün istinaddan ibarət mətn hissəsi əvvəlcə normal ifadələr əsasında mətndən müəyyənləşdirilir, sonra çıxarılan hər bir sitat təhlil olunur və qeyd olunur (işarələnir). Digər xidmət interfeysi isə istifadəçilərə sənədin mətnini təqdim etməyə imkan verir, sonra isə müəyyən edilmiş istinadlarla istifadəçiyə geri qaytarır [4, 5].

Başlığın və istinadların çıxarılması rəqəmsal kitabxanalarda informasiyanın çıxarılmasının əsas modulları hesab olunur. Bundan əlavə cədvəlin, şəklin, fiqurun və onunla bağlı metaverilənlərin, alqoritmlərin çıxarılması da informasiyanın çıxarılmasında istifadə olunan modullarandır. Bu metaverilənlərin çıxarılması üçün də müxtəlif üsullar mövcuddur [3, 7, 9, 10].

IV. ELMİ İNFORMASİYANIN ÇIXARILMASI PROBLEMLƏRİ

Elmi informasiyanın çıxarılması dəqiqlik, əhatəlilik və miqyaslanma (*accuracy, coverage and scalability*) ilə bağlı problemləri özündə birləşdirir [4]. Bu problemlərin ilk ikisi ümumilikdə elmi informasiyanın çıxarılmasına aiddir. Sonuncu problem isə əsasən böyük ölçülü elmi informasiyanın çıxarılması üçün xarakterikdir.

Dəqiqlik dedikdə, çıxarılmış informasiyanın doğruluğu nəzərdə tutulur. Lakin bəzən informasiya itkisi və ya informasiya çıxarılmanın özünün səhvləri ilə bağlı problemlər yarana bilər. Bir çox hallarda rəqəmsal kitabxanalar metaverilənlərin yaxşılaşdırılması üçün əlavə mənbələrin verilənlərinə inteqrasiya etməklə bu problemləri həll etməyə çalışırlar. Bir çox kitabxanalarda metaverilənlərin yaxşılaşdırılması üçün klaster istinadlarının istifadəsi təklif olunmuşdur [3]. Böyük həcmli elmi verilənlərin emalı ilə bağlı problemlərin aradan qaldırılması üçün effektiv alqoritmlərdən və ya verilənlərin paralel və paylanmış emalından istifadə olunur.

V. ELMİ VERİLƏNLƏRİN MÜBADİLƏSİ PROBLEMLƏRİ

Rəqəmsal kitabxanalar tərəfindən toplanmış verilənlərin əməkdaşlıq və elmi tədqiqatlara yardım məqsədilə mübadiləsinin böyük əhəmiyyəti vardır. Lakin kitabxanalarda toplanmış verilənlərin böyük həcmi, həmçinin müəlliflik hüquqlarının pozulması ehtimalı bu verilənlərin birgə istifadəsi üçün problemlər yaradır. Hətta bəzi hallarda müəlliflik hüquqlarının qorunmasına baxmayaraq, zaman keçdikcə müəllif və nəşriyyat ona məxsus materialın kitabxanadan silinməsi ilə bağlı sorğu göndərə bilər. İntellektual mülkiyyət və müəlliflik hüququ ilə bağlı məsələlər verilənlərin surətinin çıxarılmasını və müxtəlif qruplar arasında mübadiləsini məhdudlaşdıra bilər [3].

Rəqəmsal kitabxanalarda elmi verilənlərin mübadiləsi və birgə istifadəsi üçün müxtəlif üsullardan istifadə olunur. Rəqəmsal kitabxana verilənlərinin mübadiləsinin Açıq arxiv metaverilənlərin yığılması üçün açıq arxivlər protokolu üzrə təşəbbüs (*Open Archives Initiative Protocol for Metadata Harvesting*) çərçivəsində həyata keçirilməsi ən asan üsul hesab olunur. Bu halda kitabxanada olan bütün sənədlər (məqalələr) üçün metaverilənləri yükləmək mümkündür [3-7].

Kitabxana verilənlərinin paylanmasında əsas problem faktiki olaraq PDF sənədlərdən və çıxarılmış mətnlərdən ibarət əsas arxivin (anbarın) paylanması ilə bağlıdır. Artıq qeyd olunduğu kimi müəlliflik hüquqları məsələləri ilə yanaşı, hazırda 6 terabaytdan böyük verilənlərin paylanması ilə bağlı problemlər də mövcuddur. Bundan əlavə, verilənlər anbarı gün ərzində 10-20GB verilənlərin yığılmasıyla sürətlə böyüyür. Bu arxivin saxlanması da digər problemlər yaradır [7].

CiteSeerX və onunla bağlı layihələr (məsələn informasiyanı çıxarma modulları kimi) effektiv qarşılıqlı əlaqənin təmini üçün bir qayda olaraq açıq mənbə kodundan istifadə edirlər.

NƏTİCƏ

Rəqəmsal kitabxanalarda toplanmış elmi sənədlər və metaverilənlər elmi-tədqiqatların səmərəliliyini təmin etdiyindən istifadəçilər elektron formatda olan elmi məqalələrə

daha çox üstünlük verirlər. Eyni zamanda aparılan elmi-tədqiqatların hesabına kitabxana kolleksiyalarının həcminin sürətlə artması, verilənlərin çıxarılması və birgə istifadəsi zamanı problemlər yaradır.

Müxtəlif üsullardan, effektiv alqoritmlərdən və ya verilənlərin paralel və paylanmış emalından istifadə etməklə bu problemləri həll etmək olar.

ƏDƏBİYYAT

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

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

[3] K.Williams, J. Wu., S. R. Choudhury, M. Khabsa, C. L. Giles, “Scholarly big data information extraction and integration in the CiteSeer digital library,” Proc. of the 2014 IEEE 30th International Conference on Data Engineering Workshops (ICDEW), 2014, pp. 68–73.

[4] K. Williams., L. Li., M. Khabsa, J. Wu., Shih P.C., C. L.Giles, “A Web Service for Scholarly Big Data Information Extraction,” Proc. of the 2014 IEEE International Conference on Web Services, 2014, pp. 105–122.

[5] A.G. Ororbia II, J Wu, C.L. Giles, “CiteSeerX: Intelligent Information Extraction and Knowledge Creation from Web-Based Data,” Proc. of the 4th Workshop on Automated Knowledge Base Construction, 2014.

[6] A.G. Ororbia II, J Wu, M. Khabsa, K. Williams., C. L. Giles, “Big Scholarly Data in CiteSeerX: Information Extraction from the Web,” Proc. of the 24th International Conference on World Wide Web Companion (WWW 2015), 2015, pp. 597-602.

[7] J Wu, C.L. Giles, “Information Extraction for Scholarly Document Big Data,” www.isi.edu

[8] P. Lopez, “GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications,” Research and Advanced Technology for Digital Libraries, 2009, pp. 473–474.

[9] Y. Liu., K. Bai., P. Mitra., C.L. Giles, “Tableseer: automatic table metadata extraction and searching in digital libraries,” Proc. of the 7th annual international ACM/IEEE joint conference on Digital libraries (JCDL), 2007, pp. 91–10.

[10] S.R. Choudhury., P. Mitra., A. Kirk, S.Szep, D. Pellegrino, et.al. “Figure metadata extraction from digital documents,” Proc. of the 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 135–139.