MOKSLAS – LIETUVOS ATEITIS
SCIENCE – FUTURE OF LITHUANIA
ISSN 2029-2341 print / ISSN 2029-2252 online
http://www.mla.vgtu.lt

ELEKTRONIKA IR ELEKTROTECHNIKA
ELECTRONICS AND ELECTRICAL ENGINEERING

2014 6(2): 151–156

Sistemų inžinerija, kompiuterių technologija T 120
System Engineering, Computer Technologies T 120

http://dx.doi.org/10.3846/mla.2014.20

PAIEŠKOS ELEKTRONINIUOSE KATALOGUOSE VEIKLOS PROCESO TOBULINIMAS

Titas Savickas
Vilniaus Gedimino technikos universitetas
El. paštas titas.savickas@dok.vgtu.lt

Santrauka. Straipsnyje analizuojama paieška elektroniniuose kataloguose. Pasirinkta dalykinė sritis – Lietuvos universitetų akademinių bibliotekų elektroninis katalogas. Straipsnyje pateikiamos esamos paieškos sistemos ALEPH problemos ir vartotojų poreikių tyrimo rezultatai. Išanalizuojamas MARC21 formatas ir galimi alternatyvūs paieškos būdai, naudojami bibliotekoje. Atlikus analizę siūloma sistemos architektūra ir paieškos procesas, kuriais bandoma padidinti paieškos elektroniniame kataloge efektyvumą ir užtikrinti vartotojų poreikių patenkinimą.

Reikšminiai žodžiai: elektroninis katalogas, biblioteka, semantinė paieška, MARC21, ALEPH, ontologijos.

Įvadas

Augant žiniatinklio populiarumui vis sudėtingiau vartotojams pateikti naują ir dinamišką turinį, nes vartotojams sunku jį pasiekti. Taip yra todėl, kad vartotojai nežino, kaip tai padaryti.

Efektyvios paieškos sistemos – vienas svarbiausių elementų, padedančių gauti informaciją. Šiuo metu yra sukurta daug skirtingų paieškos sistemų, kurios turi savo funkcijas ir tikslus. Paieškos sistemos, priklausomai nuo įgyvendinimo, turi ir privalumų, ir trūkumų (Mangold 2007).

Šį straipsnį sudaro penkios dalys – probleminės srities tyrimas, susiję problemos sprendimo būdais, Lietuvos akademinių bibliotekų katalogams pritaikytas problemos sprendimo būdas, pagal siūlomą būdą įgyvendinto prototipo eksperimentinis tyrimas, išvados.

Probleminės srities tyrimas

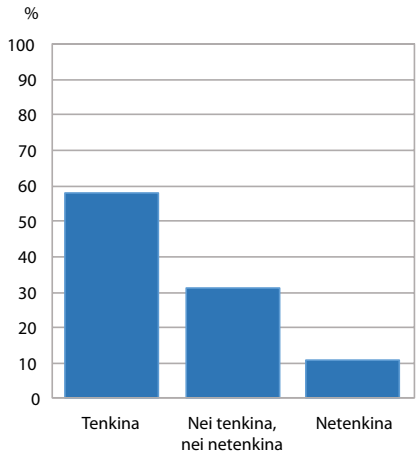
Lietuvos universitetų bibliotekos naudoja ALEPH sistemą. Ji yra skirta visiems bibliotekos procesams valdyti ir yra sudaryta iš trijų dalių – serverio, darbuotojų aplinkos ir Web OPAC (angl. OPAC – *Online Public Access Catalogue*) elektroninio katalogo. Visi leidiniai elektroniniame kataloge aprašyti naudojant MARC21 (angl. MARC – *MAchine-Readable Cataloging*) formatą.

Elektroniniame kataloge vykdomos paieškos yra paremtos reikšminiais žodžiais. Vartotojai įveda reikšminius žodžius, o sistema randa leidinius, kurių aprašymuose yra pateikti reikšminiai žodžiai. Šios paieškos sistemos paskirtis – suteikti vartotojams galimybę peržiūrėti katalogo turinį ir susirasti aiškiai apibrėžtus leidinius. Norėdami lankstesnės paieškos, vartotojai privalo patys naudoti loginius operatorius (AND, OR) ir reguliariąsias išraiškas („*“, „?“). Visgi net ir šie operatoriai nesuteikia pakankamai lankstumo, todėl vartotojai privalo tiksliai žinoti, ko ieško.

Siekiant išsiaiškinti studentų paieškos tikslus, buvo apklausti atsitiktiniai į biblioteką atėję studentai, kurie naudojosi kompiuteriais atlikdami paiešką kataloge. Iš viso buvo apklausta 90 studentų, tačiau iš jų sutiko atsakyti į klausimus tik 48. Apklausa buvo atlikta žodžiu. Kartu buvo stebima respondentų paieška kataloge, siekiant detaliai išsiaiškinti esamus trūkumus. Kadangi 2012–2013 m. m. VGTU studijavo apie 10 tūkst. studentų, gauti apklausos rezultatai turi 13 % paklaidą su 95 % patikimumu. Gauti apklausos rezultatai turėtų būti vertinami kaip tendencijos, o ne tikslus įvertinimas.

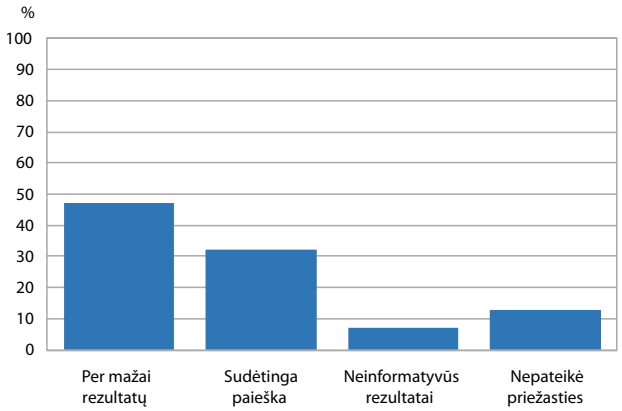
Pirmiausia vartotojų buvo klausiama, ar juos tenkina esama paieškos sistema (1 pav.). Iš visų apklaustųjų daugiau nei pusė (58 %) buvo patenkinti esama sistema, tačiau 11 % vartotojų paieška netenkino, o 31 % neturėjo nuomonės. Vartotojų buvo teiraujamasi, kokie yra pagrindiniai paieškos tikslai kataloge ir kaip jie subjektyviai vertina paieškos sistemą.

Paklausus vartotojų, kas sistemoje kelia daugiausia problemų (2 pav.), paaiškėjo, kad pagrindiniai sistemos trūkumai yra nepakankamas grąžinamų rezultatų skaičius



1 pav. Vartotojų paieškos tenkinimas

Fig. 1. User satisfaction with search engine



2 pav. Nepasitenkinimo priežastys

Fig. 2. Reasons for dissatisfaction

(47 % apklaustųjų) ir sudėtingas paieškos procesas (32 % apklaustųjų). Likusi respondentų dalis teigė, kad juos netenkina informacija, pateikiama kartu su rezultatais (7 %), arba nepateikė priežasties (13 %).

Paskutinis klausimas – ko kataloge ieško vartotojai. Paaiškėjo, kad paieškos kataloge gali būti suskirstytos į dvi pagrindines kategorijas:

- Tikslinė paieška. Vartotojai ieško leidinio, kurio visa informacija žinoma, ir tenori pasiekti tam tikro leidinio kortelę ir jį užsakyti.
- Abstrakti paieška. Vartotojai nežino, ko tiksliai ieško, jų tikslas – rasti literatūros tam tikra tema ar tam tikram studijų programos moduliui.

Tikslinės paieškos kategorijai esama paieškos sistema puikiai tinka, nes sistema grąžina tik visiškai užklausą tenkinančius rezultatus. Antros kategorijos paieškos poreikių sistema visiškai netenkina, nes grąžinamas nepakankamas rezultatų skaičius. Atliekant apklausą buvo pastebėta, kad atliekant abstrakčią paiešką vartotojams sunku atsirinkti, kokius reikšminius žodžius reikia įvesti, ne visi vartotojai moka naudoti loginius operatorius ir reguliariąsias išraiškas. Dėl to, kad vartotojai tiksliai nežino, ko ieško, paieška tampa neefektyvi.

Paieška, paremta reikšminiais žodžiais, yra tiksli, tačiau neveiksminga, nes gaunami tik tie rezultatai, kurie visiškai atitinka užklausoje pateiktus reikšminius žodžius. Papildomų problemų sukelia ir kalbos gramatikos sudėtingumas – lietuvių kalboje yra linksniai, asmenuotės, priesagos. Taip pat kalbose galimi ir sinonimai, homonimai, hiperonimai, kurie apsunkina reikšminiais žodžiais paremtą paiešką, nes panašūs leidiniai gali būti aprašyti vartojant gramatiškai skirtingus, tačiau semantiškai panašius reikšminius žodžius.

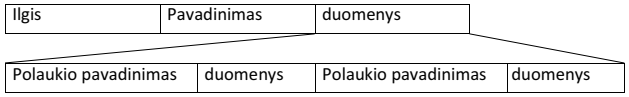
Vienas iš bandymų paieškas padaryti efektyvesnes – semantinė paieška. Semantinė paieška išteklius susieja tarpusavyje pagal tam tikrus atributus, o tai leidžia panaikinti griežtą priklausomybę nuo reikšminių žodžių.

Bibliotekos katalogo duomenys

Bibliotekos kataloge saugoma informacija apie leidinius MARC21 formatu. Šis formatas yra struktūrizuotas ir pritaikytas bibliografinei informacijai saugoti. MARC formatą sukūrė Kongreso biblioteka 1971 m. Šiuo metu yra išleista 21 versija. Lietuvos akademinių bibliotekų tinklas MARC21 specifikaciją išvertė ir pradėjo naudoti katalogo duomenims 2002 m.

MARC21 formato aprašai saugomi kaip individualios eilutės, kurias kompiuteris gali apdoroti automatiniu būdu ir kiekviena eilutė aprašo vieną bibliografinį įrašą. Bibliografinio aprašo eilutę sudaro laukai, kurių kiekvienas prasideda skaičiumi, nurodančiu eilutės ilgį baitais. Po šio skaičiaus eina lauko pavadinimas (pvz., 620 ar 100) ir galiausiai eina lauko duomenys. Lauke esantys duomenys taip pat turi savo struktūrą. Duomenys sudaryti iš polaukių, kurių pradžia žymima $$ simbolių seka. Bendras įrašų formatas pateikiamas 3 pav., o 4 pav. – MARC21 formato įrašo pavyzdys.

Pagal specifikaciją pirmieji įrašo laukai yra kontroliniai, o po jų eina laukai su bibliografine informacija (MARC21 dokumentacija 2012). Visi laukai turi savo paskirtis, juose rašoma tik atitinkama bibliografinė informacija. Pavyzdžiui, 100 lauke saugoma informacija apie autorius, 245 lauke – leidinio antraštė, 080 lauke – universalus dešimtainės koduotės kodas. Ne visa bibliografinė informacija (pvz., kontrolinių laukų turinys, šifras, klasifikaciniai numeriai) yra aktuali vartotojui.



3 pav. MARC21 įrašų struktūra
Fig. 3. Structure of MARC21 records

```
0008FMT LBK0030LDR L01544nam^a2200433^i^450 00
021001 LVGT01-0000000230014003 LLI-ViGTU 00220
05 L20050707151547.00046008 L001127s1998 ^^^^l
i^||||e|^|||^001^0^lit^d0019020 L$$a9986133424
0023040 L$$aLI-ViGTU$$blit00180410 L$$alit$$ae
ng00 11044 L$$ali0022080 L$$a37.013(474.5)0024
080 L$$a37.014.5(474.5)016224500L$$aEdukologij
os idėjos Lietuvos švietimo sistemos moderniza
vimui:$$ bmonografija /$$credaktorė Palmira Ju
cevičienė; Kauno technologijos universitetas.0
041260 L$$aKaunas :$$bTechnologija,$$c1998 .00
34300 L$$a539 p. :$$ blent., brėž.0031504 L$$a
Bibliogr.: p. 506-532.00526 50 7L$$aŠvietimas
$$xFilosofinis aspektas
```

4 pav. MARC21 įrašo fragmentas
Fig. 4. Fragment of a MARC21 record

Toks griežtai struktūrizuotas tekstas yra lengvai apdorojamas ir gali būti nesunkiai transformuotas į kitas formas – RDF formatą (Styles *et al.* 2008), DublinCore ontologiją (Kruk *et al.* 2009).

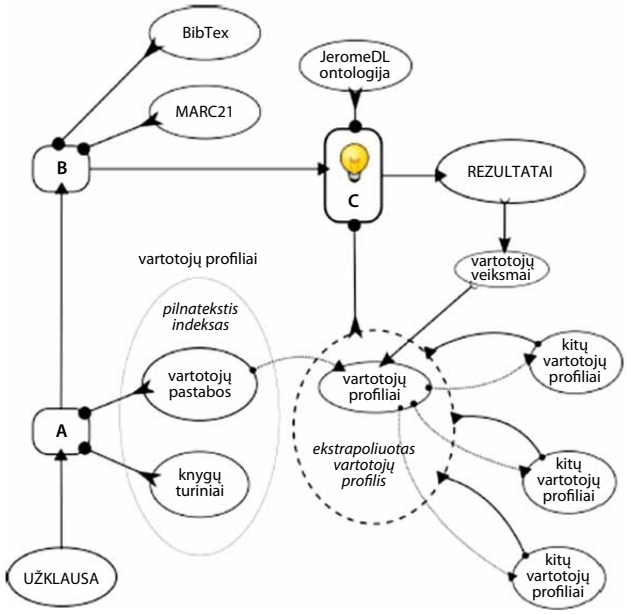
Susiję darbai

Bibliotekų kataloguose taikomi įvairūs bibliografinių įrašų aprašymo metodai – BibTex, MARC, DublinCore ir pan. Šie formatai yra griežtai aprašyti ir juose yra mažai semantikos. Siekiant supaprastinti paieškos mechanizmą elektroniniuose kataloguose ir suteikti struktūrizuotą prieigą prie bibliotekų elektroninių katalogų išteklių, bandoma šiuos išteklius semantiškai aprašyti.

MarcOnt iniciatyva siekė sukurti ontologiją bibliotekoms (Synak, Kruk 2005). Dėl šios iniciatyvos buvo sukurta JeromeDL skaitmeninė biblioteka, kurioje leidiniai turi ne tik įprastus bibliografinius aprašus, bet ir jų transformaciją į JeromeDL ontologiją (Kruk *et al.* 2009). Paieška šioje sistemoje (5 pav.) suskirstyta į šešis žingsnius:

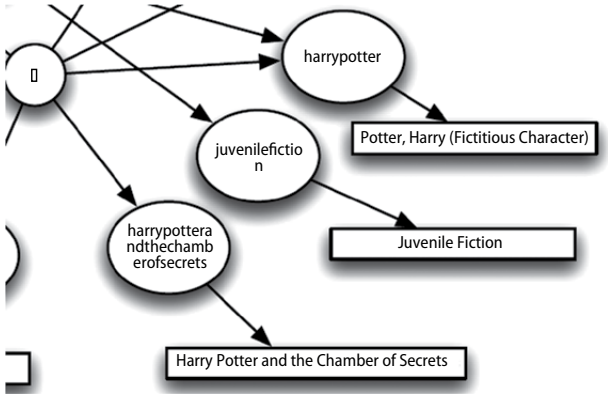
1. Užklausos priėmimas.
2. Visatekstė paieška knygose.
3. Visatekstė paieška vartotojų anotacijose.
4. Paieška bibliografiniuose aprašuose.
5. Paieška pagal JeromeDL ontologiją ir vartotojų profilius.
6. Rezultatų pateikimas.

JeromeDL paieškos algoritmo alternatyva yra MARC leidinių transformavimas į RDF įrašus ir jų susiejimas tarpusavyje naudojant maišos (angl. *Hash*) funkcijas, kuriose panašus tekstas konverguoja į tam tikrą maišos funkcijos rezultato reikšmę (Styles *et al.* 2008).



5 pav. Paieškos algoritmas JeromeDL skaitmeninėje bibliotekoje (Kruk *et al.* 2009)
Fig. 5. Search algorithm in JeromeDL Digital Library (Kruk *et al.* 2009)

Šioje sistemoje visų leidinių tradicinis MARC21 ar BibTex bibliografinis aprašas transformuojamas į RDF aprašus, o nuoroda į išteklių sugeneruojama pagal šį aprašą naudojant maišos funkciją. Gautos nuorodos semantiškai panašius leidinius leidžia pasiekti per tą pačią nuorodą. Tačiau ši sistema veikia tik anglų kalbai, kitoms kalboms algoritmas nėra pritaikytas. 6 pav. pateikiami sistemoje esančių išteklių semantinis susiejimas ir jiems pasiekti vartojami adreso reikšminiai žodžiai.



6 pav. Semantiškai susietų išteklių tinklo fragmentas ir jų pasiekimo reikšminiai žodžiai (Styles *et al.* 2008)
Fig. 6. Fragment of semantically interconnected resources network and their hashed keywords (Styles *et al.* 2008)

Siūloma paieškos sistemos architektūra

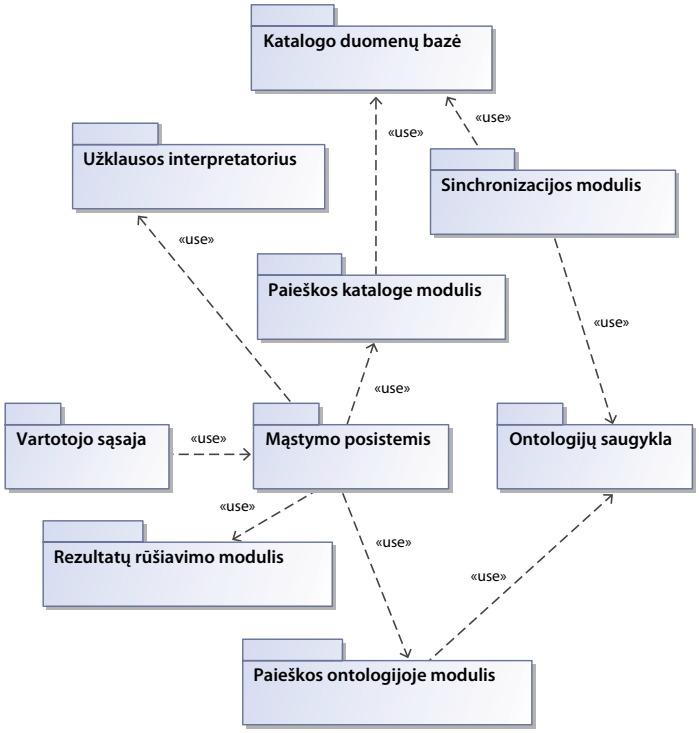
Susijusiuose darbuose siūlomos paieškos sistemos atlieka paieškas, kuriose stengiamasi suprasti vartotojo užklausos prasmę ir pagal ją atlikti paiešką. JeromeDL naudojasi vartotojų informaciją paieškai atlikti. Taip pat esamos sistemos nėra pritaikytos lietuviškiems elektroniniams katalogams. Siekiant išspręsti esamas problemas, siūloma naudoti semantinės paieškos sistemą, sudarytą iš dviejų etapų. Tokios sistemos, pritaikytos elektroniniam katalogui, architektūra pateikta 7 pav.

Šią sistemą sudaro šie moduliai:

- ontologijų saugykla. Joje saugoma leidinių ontologija su semantiškai aprašyta leidinių informacija;
- katalogo duomenų bazė. Duomenų bazėje saugomi MARC21 leidinių aprašai;
- sinchronizacijos modulis. Modulis atsako už tai, kad katalogo ir ontologijų duomenų bazės saugotų sinchronizuotus duomenis;
- užklausų interpretatorius. Šis modulis išanalizuoja užklausą, atrenka vertingus reikšminius žodžius, grąžina jų pradinę formą ir suteikia papildomos informacijos apie tai, kokia sakinio dalis yra reikšminiai žodžiai, o tai leidžia labiau atsižvelgti į užklausos prasmę;
- paieškos kataloge modulis. Modulis atsako už pradinio rezultato rinkinio išgavimą, kuriuo paskui remiamasi ieškant semantiškai panašių leidinių;
- paieškos ontologijoje modulis. Modulis atsako už rezultatų rinkinio išplėtimą pagal pateiktą leidinių rinkinį;
- mąstymo posistemis. Šis modulis atsako už paieškos vykdymą. Modulis priima užklausą iš vartotojo sąsajos, naudoja kitus modulius rezultatams išgauti ir rūšiuoti bei grąžina rezultatą vartotojui;
- rezultatų rūšiavimo modulis atsako už visų paieškos rezultatų rūšiavimą pagal jų atitiktį vartotojo užklausai.

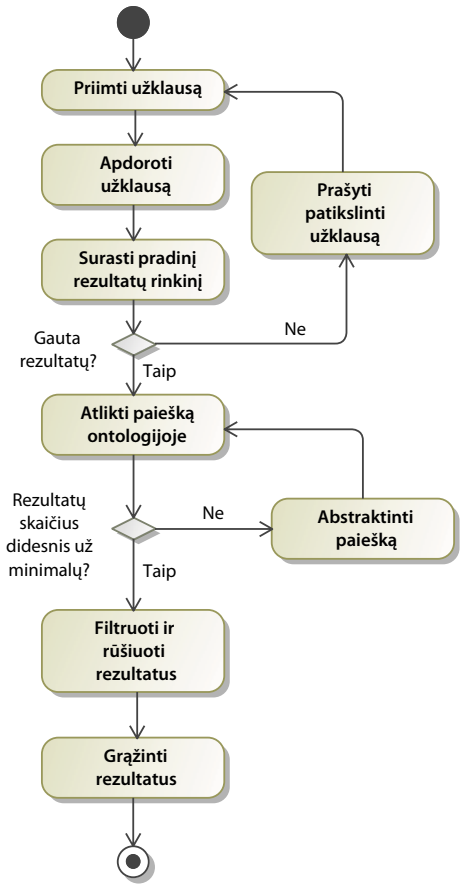
Siūlomos sistemos paieškos procesas (8 pav.) sukurtas siekiant patenkinti visus vartotojų poreikius.

Vartotojams pateikus paieškos užklausą, kurią sudaro reikšminių žodžių rinkinys ir papildomi filtrai, sistema pirmiausia atlieka paiešką katalogo duomenų bazėje ir suranda gramatiškai panašius leidinius. Šiuo žingsniu gaunamas rezultatų rinkinys, kuris tenkina tikslinės paieškos vartotojų poreikius. Jei pradinis rezultatų rinkinys tuščias, vartotojo prašoma performuoti paiešką su siūlomais reikšminiais žodžiais.



7 pav. Semantinės paieškos elektroniniame kataloge sistemos architektūra

Fig. 7. Architecture of a system for semantic search in an electronic catalogue



8 pav. Sistemos paieškos procesas

Fig. 8. Search process

Kitas sistemos žingsnis – rezultatų rinkinio išplėtimas. Sistema atlieka paiešką bibliotekos katalogo ontologijoje ir randa semantiškai panašius leidinius. Jei gaunamas mažas rezultatų rinkinys, sistema pakartoja paiešką naudodama abstraktesnius paieškos kriterijus.

Galiausiai rezultatų rinkinys surūšiuojamas pagal vartotojų užklausos atitiktį, ir pagal užklausoje pateiktus reikalavimus rezultatai filtruojami.

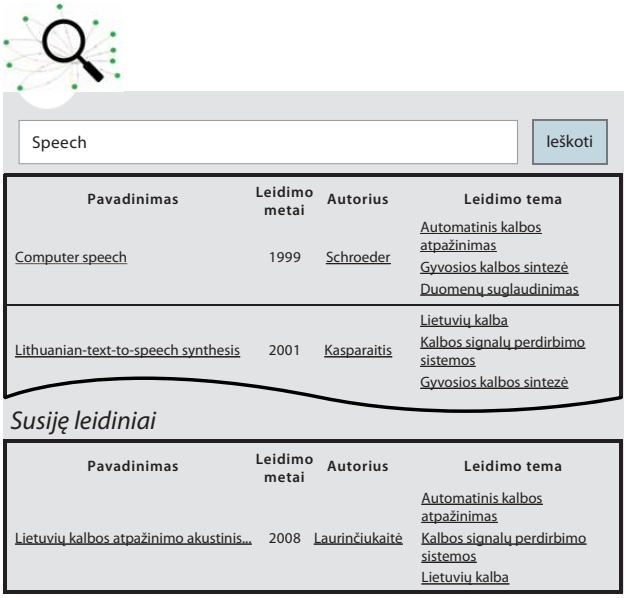
Eksperimentinis tyrimas

Siūlomos architektūros eksperimentiniam tyrimui sukurtas prototipas, kuriuo siekta praktiškai ištirti siūlomo sprendimo rezultatus. Prototipas sukurtas pagal vieno langelio principą – naudotojas visą jam aktualią informaciją mato tame pačiame lange (9 pav.).

Naudotojas pagrindiniame lange esančiame teksto lauke įveda norimą užklausą ir spaudžia paieškos mygtuką. Sistema grąžina rezultatus, suskirstytus į dvi dalis – gramatiškai panašiausius vienoje dalyje ir semantiškai susijusius rezultatus antroje dalyje.

Kartu, siekiant padaryti paiešką paprastesnę, kartu su rezultatais pateikiama ir naudotojams aktuali informacija – turinys, viršelis, leidimo metai, autoriai ir leidiniui priskirta tema MARC apraše.

Siekiant ištirti sistemos efektyvumą, buvo pasirinktas eksperimentinio lyginimo metodas – palyginti gaunami paieškos rezultatai su įgudusio naudotojo standartinės reikšminiais žodžiais paremtos paieškos, naudojant Bulio operatorius ir reguliariąsias išraiškas, rezultatais. Eksperimento metu buvo atlikta 60 paieškos užklausų ir įvertintas paieškos tikslumas ir gaunamų rezultatų skaičius.



9 pav. Prototipo lango fragmentas

Fig. 9. Fragment of the prototype window

Paieškos tikslumo vertinimas (Büttcher *et al.* 2010):

$$T = \frac{m}{n} \cdot 100 , \quad (1)$$

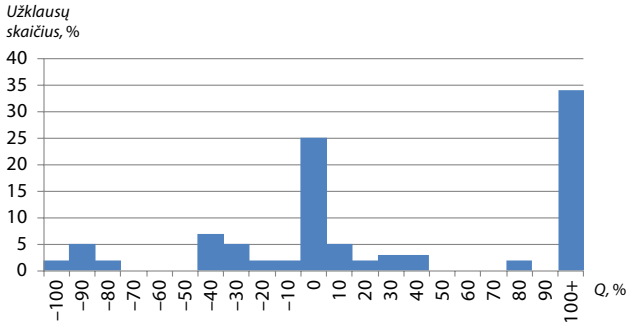
čia m – užklausą atitinkančių rezultatų skaičius, vnt.; n – bendras rezultatų skaičius, vnt.

Naudingų rezultatų skaičiaus palyginimas (Büttcher *et al.* 2010):

$$Q_p = \left(\frac{m_\mathrm{p}}{n_\mathrm{a}} - 1 \right) \cdot 100 , \quad (2)$$

čia m_p – prototipo naudingų rezultatų skaičius, vnt.; n_a – įprastos paieškos naudingų rezultatų skaičius, vnt.

Pirmiausia buvo lyginamas paieškos rezultatų santykis (10 pav.). Iš gautų rezultatų matyti, kad 34 % atvejų prototipas grąžino daugiau nei du kartus didesnį naudingų rezultatų skaičių. Įgudusio naudotojo paieška 25 % atvejų grąžino daugiau rezultatų nei prototipas ir 2 % atvejų grąžino rezultatų, kai prototipas negrąžino nieko. 25 % atvejų rezultatų skaičius sutapo su prototipo. 25 % užklausų grąžino tą patį rezultatų skaičių visoms trims paieškos grupėms, tad galima teigti, kad toms užklausoms daugiau tinkamų rezultatų kataloge ir nėra. Likusiais 16 % atvejų prototipas grąžino 10–80 % daugiau rezultatų nei įgudusio naudotojo paieška.

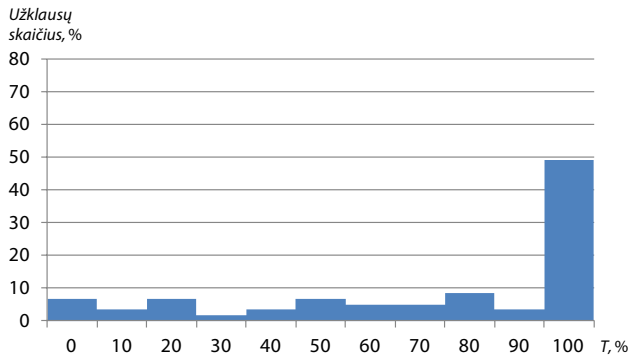


10 pav. Prototipo ir įgudusio naudotojo paieškos rezultatų santykio palyginimas

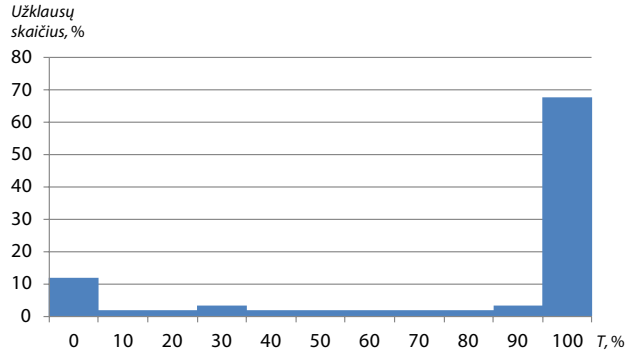
Fig. 10. Relative comparison of search results by prototype and skilled user

Vien tik paieškos rezultatų skaičius neapibūdina paieškos sistemos naudingumo, todėl papildomai buvo lyginamas įgudusio naudotojo (11 pav., b) ir prototipo (11 pav., a) paieškų tikslumas. Įgudusio naudotojo reikšminiais žodžiais paremtoje paieškoje 18,61 % atvejų paieška grąžino nenaudingų rezultatų. Tikslumas tais atvejais kito 10–90 %. 62,71 % atvejų įgudusio naudotojo paieška grąžino tik visiškai su užklausa susijusius rezultatus. Bendras įgudusio naudotojo paieškos tikslumo vidurkis yra 78,8 %.

a)



b)



11 pav. Paieškos tikslumo pasiskirstymo palyginimas: a – prototipo; b – įgudusio naudotojo

Fig. 11. Distribution of the search result precision for: a – prototype; b – skilled user

Prototipas veikia ne taip tiksliai kaip įprasta reikšminais žodžiais paremta paieška. 90–100 % tikslumas pasiektas 49,15 % užklausų. 15,25 % atvejų prototipo tikslumas buvo mažesnis nei 50 %. Toks mažas tikslumas buvo tų užklausų, kurioms įprasta paieška iš viso negrąžino rezultatų arba grąžino labai mažą rezultatų skaičių, todėl įsivėlė „triukšmas“.

Išvados

1. Atlikta esamos bibliotekos katalogo paieškos sistemos vartotojų apklausa parodė, kad esama sistema yra neveiksminga ir neatitinka jų poreikių. Bibliotekos katalogo duomenų ir su problema susijusių darbų analizė parodė, kad ezistuojantys sprendimai nepanaikina visų problemų ar netinka dalykinės srities problemoms spręsti.
2. Pasiūlyta semantinės paieškos elektroniniame kataloge sistemos architektūra, kuria siekiama patobulinti paieškos procesą ir pasiekti geresnių rezultatų, patenkinant abiejų paieškos kategorijų reikalavimus.
3. Įgyvendinus prototipą ir jį eksperimentais ištyrus nustatyta, kad jo veikimas yra efektyvesnis nei įprasta paieška, nes 74 % užklausų grąžino tiek pat arba daugiau rezultatų, o tikslumas išlaikytas panašus (atitinkamai įprasto naudotojo paieškai 78,8 % ir prototipo atveju 72,54 %).

Literatūra

Büttcher, S.; Clarke, C. L. A.; Cormack, G. V. 2010. *Information retrieval: implementing and evaluating search engines*. Cambridge: MIT Press. 606 p.

Kruk, S. R., *et al.* 2009. JeromeDL: the social semantic digital library, in *Semantic Digital Libraries*. Berlin, Heidelberg: Springer, 139–150.

Mangold, C. 2007. A survey and classification of semantic search approaches, *International Journal of Metadata, Semantics and Ontologies* 2(1): 23–34.

MARC21 Dokumentacija. 2012. *The Library of Congress* [interaktyvus], [žiūrėta 2013 m. kovo 20 d.]. Prieiga per internetą: http://www.loc.gov/marc/status.html

Synak, M.; Kruk, S. R. 2005. *MarcOnt Initiative – the Ontology for the Librarian World* [interaktyvus], [žiūrėta 2012 m. sausio 26 d.]. Prieiga per internetą: http://www.marcont.org/marcont/pdf/ms_eswc2005marcont.pdf

Styles, R.; Ayers, D.; Shabir, N. 2008. *Semantic MARC, MARC21 and the Semantic Web* [interaktyvus], [žiūrėta 2013 m. kovo 20 d.]. Prieiga per internetą: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper02.pdf

IMPROVEMENT OF SEARCH PROCESS IN ELECTRONIC CATALOGUES

T. Savickas

Abstract

The paper presents investigation on search in electronic catalogues. The chosen problem domain is the search system in the electronic catalogue of Lithuanian Academic Libraries. The catalogue uses ALEPH system with MARC21 bibliographic format. The article presents analysis of problems pertaining to the current search engine and user expectations related to the search system of the electronic catalogue of academic libraries. Subsequent to analysis, the research paper presents the architecture for a semantic search system in the electronic catalogue that uses search process designed to improve search results for users.

Keywords: electronic catalogue, library, semantic search, MARC21, ALEPH, ontology.