TRENDI V INTELIGENTNEM PRIDOBIVANJU PODATKOV IZ HETEROGENIH VIROV

Slavko Žitnik, Bojan Klemenc, Dejan Lavbič, Lovro Šubelj, Štefan Furlan, Aleš Kumer, Aljaž Zrnec, Neli Blagus in Marko Bajec
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška cesta 25, SI-1000 Ljubljana
slavko.zitnik@optilab.net, [ime.priimek]@fri.uni-lj.si

Povzetek

V tem članku predstavljamo trende na področju pridobivanja podatkov in predlagamo nov iskalnik oziroma sistem za priklic podatkov iCORE - intelligent context-aware (multisource) information retrieval. Predlagani sistem skuša narediti iskanje še bolj personalizirano. Zaradi strukture omogoča uporabo na katerikoli napravi in beleženje uporabnikovih aktivnosti. Uporabnika želi spremljati do takšne mere, da mu bo prikazoval strukturirane rezultate poizvedbe, brez da bi mu jih eksplicitno podal. Uporablja lahko več zbirk podatkov ali specializiranih iskalnikov - metaiskalnik, obdela zadetke in iz njih izlušči informacije. Uporabnik kot rezultat sistema iCORE prejme sestavljene informacije, ki so ustrezno prikazane glede na njihovo strukturo, imeti mora možnost ročnega nastavljanja preferenc, usmerjenega iskanja in navigacije med vrnjenimi entitetami, ki so med seboj povezane.

Ključne besede

priklic podatkov, iskalnik, kontekst, personalizacija, iCORE

Abstract

TRENDS IN INTELLIGENT DATA RETRIEVAL FROM HETEROGENEOUS DATA SOURCES

In this paper, we present trends in the area of data retrieval and propose a new search engine and information retrieval system iCORE - intelligent context-aware (multisource) information retrieval. This system personalizes search more than comparable approaches. The structure of the system enables the application on various devices with possibility of tracking the user. iCORE wants to focus on the user and retrieve relevant information even before the user actually tries to give an explicit input. It uses other specialized search engines or data collections. Results are presented as structured information, which are presented appropriately. The user has to have options such as setting his preferences, faceted search and navigating through results.

Key words

information retrieval, search engine, context, personalization, iCORE

1 UVOD

Odkar je človek začel zbirati podatke, je vedno skušal poskrbeti, da so bili čim enostavneje dosegljivi. Značilni primer organizacije podatkov je kazalo v knjigi. Z začetkom tiskanja knjig in ustanavljanjem knjižnic se je uveljavilo urejevanje publikacij po tematskih področjih, naslovu in avtorju.

Podobno se je s podatki dogajalo v računalništvu [6]. Sprva je bilo le malo dokumentov, ki smo jih lahko obvladali. Zaradi potrebe po izmenjevanju virov se je rodil svetovni splet. Tipično se vsak

dokument nahaja na določenem naslovu. Tega moramo eksplicitno poznati, da lahko do njega dostopamo. Pravzaprav so se spletni iskalniki pojavili za reševanje teh težav. Omogočajo uporabo širšim množicam ljudi, izjemno poenostavljajo dostop do dokumentov in pohitrijo uporabo interneta. Vprašajmo se: »*Kam pogledamo, ko želimo pridobiti neke informacije?*« Prepričani smo, da je najpogostejši odgovor enak imenu najbolj popularnega iskalnika v današnjem času.

Čeprav so nam spletni iskalniki najbližje, avtomatski ali polavtomatski sistemi za pridobivanje podatkov že dolgo obstajajo. Najprej so se začeli uporabljati v knjižnicah, kjer je bila potreba po njih največja. Iskalni sistem je vsak sistem, ki poizveduje nad podatkovno bazo. Gotovo smo se že znašli v primeru, ko nas je uradnik vprašal: »*Ali mi poveste vaše ime in priimek, da najdem vaše podatke?«.* V tem primeru mu je sistem omogočal poizvedovanje do podatkov le po imenu in priimku.

Prvotna beseda za iskanje je bila priklic podatkov, ki se je v splošni uporabi preoblikovala v iskanje.

Res je, da smo tekom razvoja izboljševali načine iskanja, personalizacijo rezultatov, dodajali podporo za različne vrste dokumentov, prikazovali relevantne oglase, itd., vendar pa se noben od sistemov ni ukvarjal z informacijami v dokumentih, njihovo obdelavo in prikazovanjem. V raziskovalnih krogih so se porodile zamisli o semantičnem svetovnem spletu (angl. semantic web), a trenutno ne kaže, da bi se začenjal množično uporabljati [8].

V drugem razdelku na kratko predstavimo trende in področja priklica podatkov, nato opišemo sistem iCORE in njegovo uporabo v praksi. V četrtem razdelku podrobneje predstavimo vsako komponento sistema posebej. Sistem smo razdelili na modul za obdelavo poizvedb, metaiskalnik, modul za zlivanje, modul za post-procesiranje in na statično strukturo modelov. V razdelku Evalvacija preverimo, kako lahko naš sistem testiramo in ugotavljamo njegovo uspešnost. V zadnjem razdelku navedemo nekaj primerov uporabe predlaganega sistema v javni upravi.

2 TRENDI NA PODROČJU PRIKLICA PODATKOV

S problemom inteligentnega iskanja se ukvarja veliko število posameznikov in raziskovalnih skupin. Znotraj ene organizacije so informatiki delno rešili problem z uvedbo integrirane podatkovne baze (ERP sistemi), ki združuje podatke iz več funkcionalnih področij. Še vedno pa obstaja velik problem pri povezovanju več zbirk podatkov, uporabi javno dostopnih virov, intranetnih in dokumentnih sistemov ali raznovrstnih tipov nestrukturiranih dokumentov, kar lahko rešimo le z uporabo inteligentnih metod.

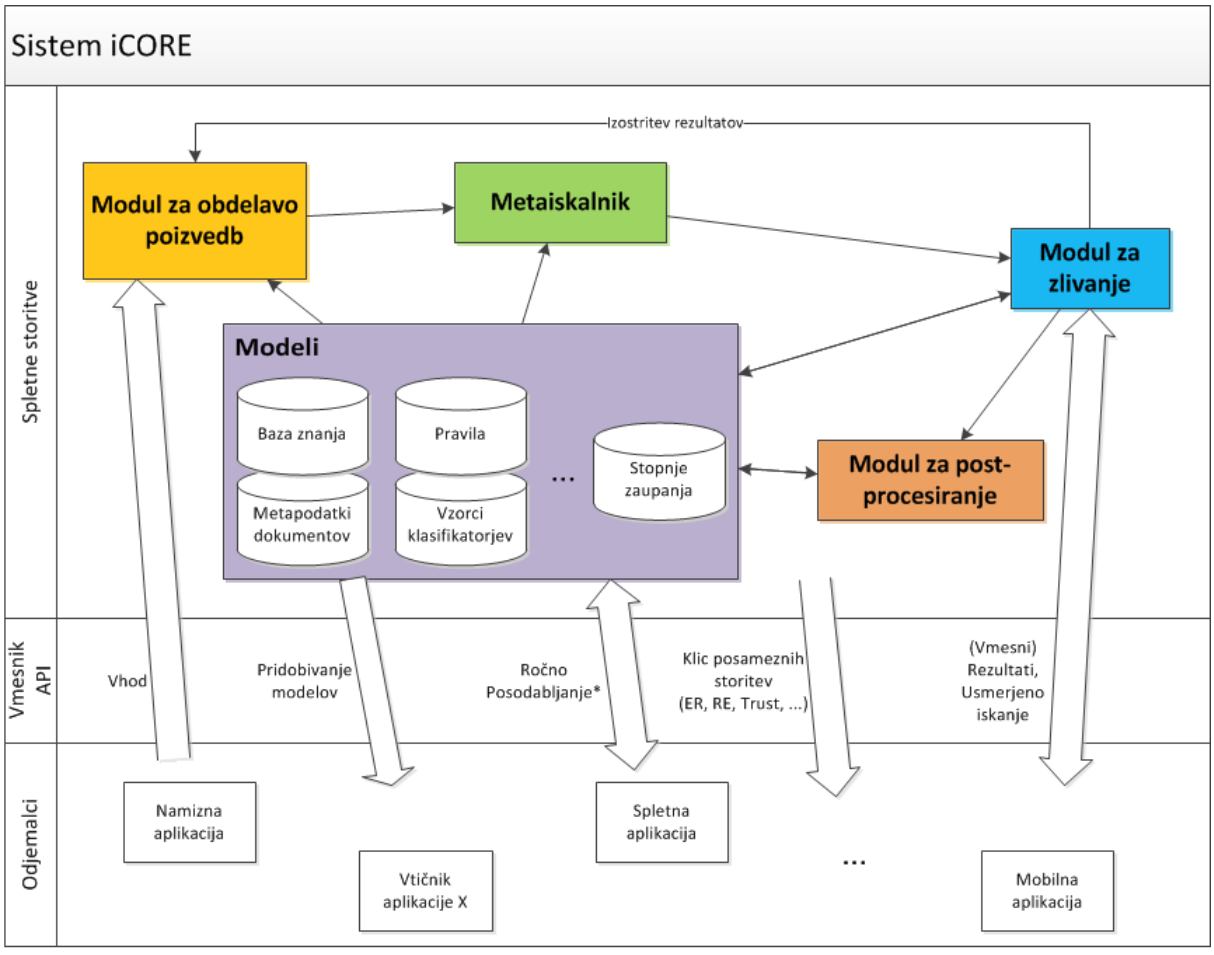
Strokovnjaki poudarjajo predvsem pomen konteksta pri poizvedovanju in prikazovanju rezultatov [14], [15], pomen personalizacije poizvedb [18], vrednost uporabe semantične informacije v različnih delih poizvedovanja [13], [16], ter pomen primerne predstavitve rezultatov uporabnikom [12]. Poleg tega študije kažejo, da se poizvedovanje znatno izboljša s pomočjo združevanja in prepletanja rezultatov različnih iskalnikov [17]. Sicer so bili v preteklosti razviti številni pristopi in metode, ki podpirajo različne dele poizvedovalnega procesa. Izkaže se, da so le-te sami po sebi pogosto nezadostni; ključna je njihova integracija [12]. Poleg omenjenega bodo vse večje količine nestrukturiranih in drugih podatkov v prihodnosti zahtevale prilagojene metode za njihovo obdelavo in integracijo [17], [19]. Tako lahko zaključimo, da bodo prihodnje generacije poizvedovalnih sistemov morale združevati inteligentne pristope v različnih delih poizvedovalnega procesa, pri čimer pa bo obdelava prilagojena poljubni obliki podatkov [12]. Podobni poizvedovalni sistemi že obstajajo [20], a ne upoštevajo vseh omenjenih metod, zato v nadaljevanju predstavljamo svoj sistem.

3 SISTEM iCORE

Sistem iCORE (**i**ntelligent **CO**ntext-aware (multisource) information **RE**trieval) je zasnovan na ideji, da razume uporabnika in njegove namere. Uporabimo ga lahko za pridobivanje informacij iz poljubnih zbirk podatkov, kot so svetovni splet ali zasebne zbirke podatkov. Sistem deluje kot skupina storitev, ki jih lahko uporabljajo odjemalci kjerkoli in kadarkoli. Poleg tega si sistem prizadeva delovati v čim večjem kontekstu glede na posameznega uporabnika, z ozirom na omejitve uporabnikovega odjemalca. Sistem pri izbiri relevantnih dokumentov lahko izkoristi tudi vpliv družbene okolice na uporabnika. Osnovna skica sistema je prikazana na sliki 1.

3.1 Storitve

Osnovna storitev, ki jo sistem ponuja uporabniku, je odgovor na njegovo poizvedbo. Uporabniku so na voljo alternativni načini vnosa poizvedbe. Odgovor ni le seznam najbolj relevantnih dokumentov,



ampak že strukturirano predstavljene informacije.

Slika 1: Sistem iCORE

Sistem poizvedbo najprej analizira. Glede na uporabnikov kontekst poskuša z interno bazo znanja ugotoviti, kaj uporabnik želi. V kontekst se štejejo vsa uporabnikova dejanja, ki jih sistem zajema. Iz konteksta se iščejo vzorci z upoštevanjem dimenzij kot so na primer čas, uporabnikovo razpoloženje, obnašanje. Nato sistem ustrezno preoblikuje poizvedbo in izračuna, katere iskalnike se izplača

uporabiti s pomočjo **modula za obdelavo poizvedb**. Sistem ne vsebuje lastnega iskalnika, ampak njegov iskalni modul deluje kot **metaiskalnik**. Sistem uporablja vnaprej določene iskalnike ali jih samodejno odkriva. Ko pridobi ustrezne rezultate več iskalnikov, ki so navadno predstavljeni v obliki povezav do dokumentov, jih rangira glede na ustreznost. Pri tem si pomaga s statičnimi ocenami in dinamičnimi modeli, ki jih neprestano posodablja. **Modul zlivanja** obdela rangirane dokumente, tako da iz njih izlušči pomembne informacije, s katerimi lahko obogati znanje uporabnika. Sedaj so delni rezultati že pripravljeni za uporabnika. Ključne podkomponente modula za zlivanje so združevanje duplikatov, preverjanje stopnje zaupanja virov v določenem kontekstu in iskanje zakonitosti v spletnih vsebinah (angl. web mining). Če sistem ugotovi, da lahko delne rezultate še izboljša z novimi poizvedbami, samodejno ustvari novo poizvedbo in ponovi celoten postopek. Uporabniku sistem načeloma vrne več rezultatov, od katerih uporabnik izbere pravilnega oz. tistega, po katerem naj bi poizvedoval. Uporabnik lahko izbere tudi bolj natančno nadaljnjo iskanje, če je bila njegova prvotna poizvedba preveč splošna. V zadnjem delu mora sistem še posodobiti modele za nadaljnjo uporabo. To je naloga **modula za post-procesiranje**, ki mu vhodne podatke posredujeta modul za zlivanje in morda tudi odjemalec. **Modeli** vsebujejo ocene virov, metapodatke o virih, stopnje zaupanja, podatke o uporabniku, bazo znanja, itd. Pravilnost in stalna posodobljenost modelov je ključni del za uspešno delovanje sistema.

Ostale storitve, ki so na voljo uporabniku, so še dostop do baze znanja in modelov. Za spreminjanje modelov ali baze znanja potrebujejo uporabniki dodatno avtorizacijo. Uporabnik lahko sistemu doda svoj zasebni iskalnik, če ga le-ta ne doda samodejno. Za enega ali več dokumentov po potrebi uporabi modul za zlivanje, ki mu izlušči entitete in jih obogati.

3.2 Programski vmesnik

Sistem deluje kot storitev, zato ima definiran splošen vmesnik za komunikacijo z odjemalci. Odjemalci uporabljajo varno komunikacijo in prenašajo sporočila v standardiziranih formatih, kot sta XML ali JSON.

3.3 Odjemalci

Posamezni uporabnik značilno uporablja več naprav, zato sklepamo, da bo uporabljal tudi več različnih odjemalcev. Odjemalec na mobilnem telefonu lahko na primer prebira SMS-e, obdeluje telefonske pogovore, medtem ko odjemalec v spletnem brskalniku nadzoruje brskanje po spletnih straneh, uporabo anonimnega načina in s tem ustvarja dodaten uporabnikov kontekst. Uporabniku moramo torej ponuditi, da ga bo sistem prepoznal ne glede na to, kje ga uporablja.

Za odjemalca je pomembno, kakšne možnosti ponuja za vnos uporabnikove poizvedbe. Ali je to na primer le niz znakov, prilaganje celotnega dokumenta, vpis znanja o neki entiteti z določenimi atributi – slika ali zvok. Vnosu nato sledi prikaz in navigacija med rezultati, dodatno filtriranje ali uporaba za nadaljnjo iskanje.

Trenutno sistemi za priklic podatkov še vedno uporabljajo vnosno polje za podajanje poizvedbe. V prihodnosti se bo to verjetno spremenilo, saj se s to težavo ukvarja novo področje o inteligentnih uporabniških vmesnikih. Ali bi nam sistem za priklic podatkov lahko vrnil rezultate, še preden mu podamo poizvedbo? Sistem iCORE bo poskušal uporabiti vse zgornje načine za zajem uporabnikovega obnašanja. Naučil se bo ključnih vzorcev, da bo znal sestaviti poizvedbo in mu predstaviti rezultate, še preden bi poizvedoval sam uporabnik. Četudi uporabnik morda včasih ne želi poizvedovati po določenih stvareh, je lahko zanj koristno, da se mu prikazujejo relevantne informacije glede na njegovo početje, saj so lahko njegove informacije napačne.

Odjemalec je najbližje uporabniku in z beleženjem njegovih dejanj lahko pomaga sistemu povečevati uporabnikov kontekst. V kontekst spadajo vsi podatki, ki jih sistem iCORE lahko uporabi za izboljšanje priklica podatkov. Poleg avtomatskega zajemanja konteksta bo imel uporabnik na voljo tudi ročen vnos in hiter dostop do usmerjenega iskanja. Usmerjeno iskanje (angl. faceted search) omogoča izostritev pogojev poizvedbe, če je za uporabnika na voljo preveč različnih rezultatov.

Vizualizacija

Večina spletnih brskalnikov ponuja rezultate v tekstovni obliki, pogosto linearno in urejeno po prednostnem vrstnem redu. Vendar pa so rezultati iskanj pogosto med seboj povezani in tvorijo smiselne gruče. Če želimo rezultate preiskovati bolj sistematično in si ustvariti celostno sliko, si lahko pomagamo z različnimi vizualizacijami [10], vendar pa je rezultate iskanj zaradi njihove medsebojne povezanosti smiselno predstaviti z omrežji. Omrežja pogosto vizualiziramo v dvodimenzionalnem ali tridimenzionalnem prostoru, vendar pa se pri tridimenzionalnih srečujemo s problemom učinkovite navigacije po celotni množici rezultatov. Vizualizirani rezultati so že obdelana in obogatena množica rezultatov iz metaiskalnika. Zaradi problema navigacije v tridimenzionalnem prostoru je potrebno uporabiti čim več človeških zaznavnih sistemov (multimodalna interakcija) – uporaba celotnega barvnega prostora pri predstavitvi podatkov, uporaba zvoka za označevanje in opozarjanje na potencialno zanimive rezultate, uporaba vmesnikov na dotik (haptični vmesniki) za bolj intuitivno navigacijo [11].

4 KOMPONENTE SISTEMA iCORE

4.1 Modul za obdelavo poizvedb

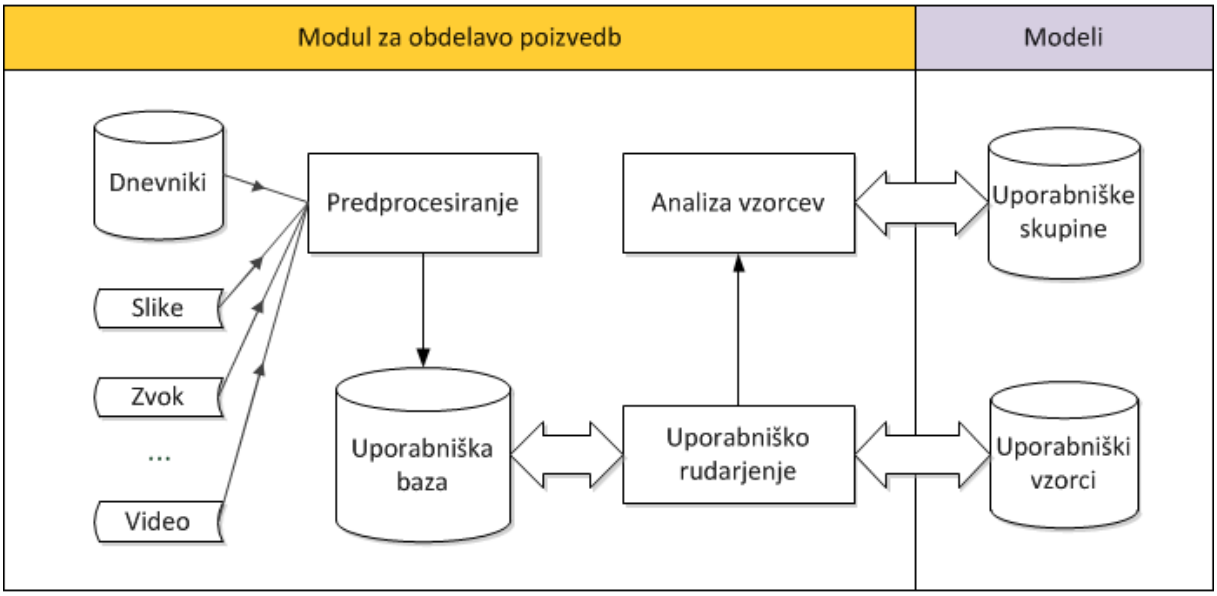
Poizvedba v splošnem predstavlja uporabnikovo izražanje zahteve po informaciji. Preko poizvedbe uporabnik predstavi del svojega trenutnega znanja, ki ga bogati pri konvergiranju do želenega cilja. Sistem iCORE s svojim vmesnikom ponuja odjemalcem več načinov vnosa poizvedb.

Metode, ki se uporabljajo za prilagajanje poizvedb, delimo na lokalne in globalne. Vsak tip metode lahko zahteva interakcijo z uporabnikom ali deluje avtomatsko. Osnovni tehniki, ki jih uporabljajo za spreminjanje poizvedb, so razširjanje poizvedb z novimi besedami ali ustrezno uteževanje delov poizvedbe.

Cilj modula za obdelavo poizvedb v sistemu iCORE je ustrezna prilagoditev poizvedb. Ugotoviti moramo različne pomenske skupine, kamor poizvedba lahko spada in vsako obravnavati posebej. Poizvedbe s pomočjo uporabnikovega konteksta prilagodimo za vsako skupino posebej in jih posredujemo modulu metaiskalnik.

Sistem poskuša spremljati uporabnika tako, da hrani njegove pretekle poizvedbe, izbiro rezultatov ali druge podatke, ki jih lahko odjemalec zajema. S pomočjo teh ob novi poizvedbi poskuša čimbolj natančno klasificirati, kaj uporabnik želi ter mu s tem omogočiti iskanje z upoštevanjem konteksta. To stori s pomočjo baze znanja, ki jo modul za post-procesiranje tudi posodablja. Ta ni koristna le za prilagajanje poizvedbe, ampak tudi za razumevanje uporabnika, ker se lahko na podlagi tega modul metaiskalnik odloča, kateri iskalnik sploh uporabiti.

Sistem iCORE uporablja dve bazi znanja. Prva je namenjena ugotavljanju tipa rezultata uporabnikove poizvedbe in poleg sheme vsebuje podatke [7]. Druga baza znanja hrani le shemo.



Slika 2: Grajenje konteksta

Za grajenje konteksta o uporabniku je za sistem uporabno spremljanje njegove interakcije z računalnikom ali drugo napravo. Osnovni koncept izgradnje je prikazan na sliki 2. Izgradnja uporablja principe uporabniškega iskanja zakonitosti v spletnih vsebinah. Običajno nas pri pisanju nekega besedila ali branja kakega dokumenta zanima določena podrobnost bolj in jo zato hočemo raziskati. Podobno lahko sklepamo tudi z zvokom ali telefonskimi pogovori. Uporabimo lahko tudi uporabnikovo obrazno mimiko, podobno kot obstajajo modeli, ki glede na kupca pred izložbo trgovine prikažejo ustrezen oglas. Prav tako so nam na voljo dnevniški zapisi strežnikov. Iz njih lahko izluščimo uporabnikove seje, čase ogledovanja določenih dokumentov in povezave, ki jih je najprej izbral (angl. clickthrough data). Vse te podatke, ki jih pridobimo, s predprocesiranjem pretvorimo v ustrezen (pogosto tabelaričen) zapis in shranimo v uporabniško bazo. Nad takšnimi podatki nato odkrivamo zakonitosti in jih pomnimo v uporabnikovih vzorcih. Uporabnike po obnašanju razporedimo v skupine. Zaradi tega vzorce enega uporabnika primerjamo z ostalimi in ga poskušamo z nenadzorovanim učenjem uvrstiti v eno ali več skupin. Pri novih klasifikacijah lahko nato uporabimo tudi rezultate skupin, v katere spada, če uporabniški vzorci niso dovolj prepričljivi.

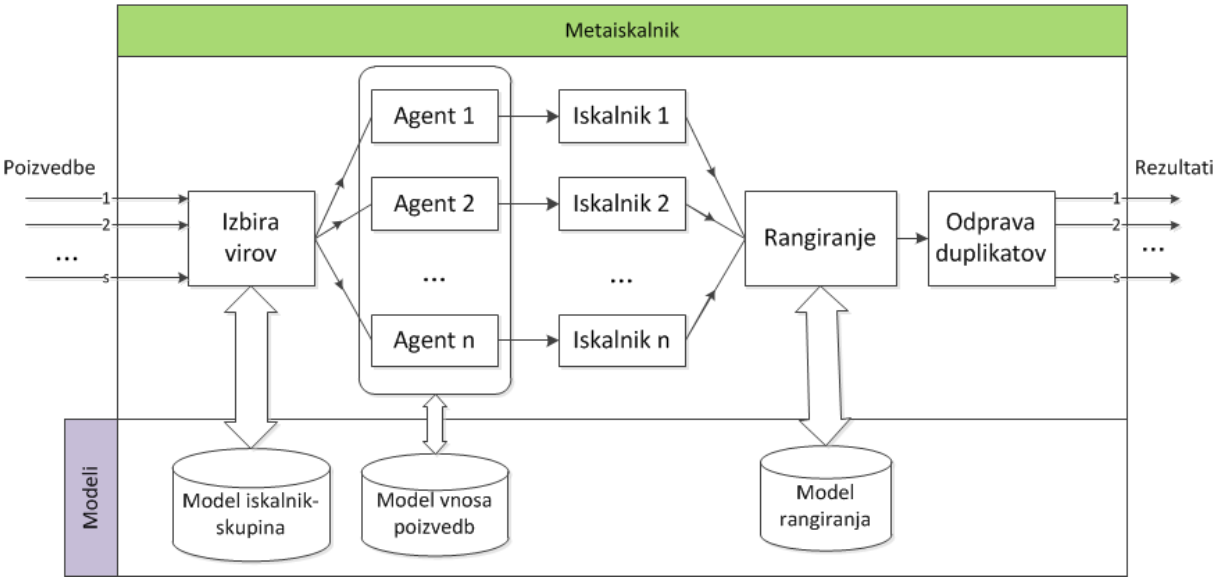
4.2 Metaiskalnik

Metaiskalnik je iskalnik, ki uporablja dva ali več drugih iskalnikov, ter predstavi njihove rezultate. Metaiskalniki se uporabljajo, kadar eden iskalnik ne vsebuje vseh primernih dokumentov [9]. S tem poskušamo uporabniku olajšati ročno iskanje z več iskalniki. Ryen je v raziskavi [1] pokazal, da 36% uporabnikov v kratkem času uporabi več iskalnikov in da je 12% vpisanih poizvedb istih na različnih iskalnikih.

Shemo delovanja metaiskalnika v sistemu iCORE si lahko ogledamo na sliki 3. Komponenta metaiskalnik od modula za obdelavo poizvedb dobi sezname poizvedb za vsako skupino posebej in tudi vrne sezname zadetkov. Potem mora glede na poizvedbo izbrati ustrezne iskalnike, ki naj bi vsebovali relevantne dokumente - izbira virov (angl. source selection). Za to skrbi klasifikator glede na trenutno stanje modelov. Za eno uporabnikovo poizvedbo lahko modul za obdelavo poizvedb oblikuje več specifičnih poizvedb, ki jih lahko prejmejo različni iskalniki. To lahko vrne mnogo dokumentov, ki bi jih moral nato modul za zlivanje obdelati. Zaradi tega pred posredovanjem rezultatov tudi ustrezno rangira pridobljene dokumente in ohrani le njihov omejen nabor. Med zadetki se lahko pojavljajo tudi duplikati, ki jih moramo odstraniti. Povsem globalna razvrstitev dokumentov ni

zadostna za iCORE. Kot smo že omenili, iCORE beleži akcije uporabnika in mu poskuša ponuditi čimbolj personalizirane rezultate. To pomeni, da moramo za dvoumne poizvedbe, kjer so dokumenti v več različnih pomenskih skupinah, za vsako skupino imeti svoj nabor rezultatov. Modul za obdelavo poizvedb že ugotovi ciljno skupino, zato se zanjo pridobi več dokumentov in za ostale manj. Ker klasifikacija uporabnikovega namena ni vedno točna, mu lahko predstavimo tudi alternativne odločitve. Nabori dokumentov za vsako skupino se posredujejo modulu za zlivanje na sliki 4. Med delovanjem ima modul metaiskalnik možnost avtomatskega prepoznavanja in dodajanja novih iskalnikov.

Modul metaiskalnik v sistemu je zelo prilagodljiv. Ker sistem iCORE lahko deluje tudi samostojno znotraj neke organizacije, imajo razvijalci možnost razviti ali uporabiti dodaten iskalnik za lastne namene.



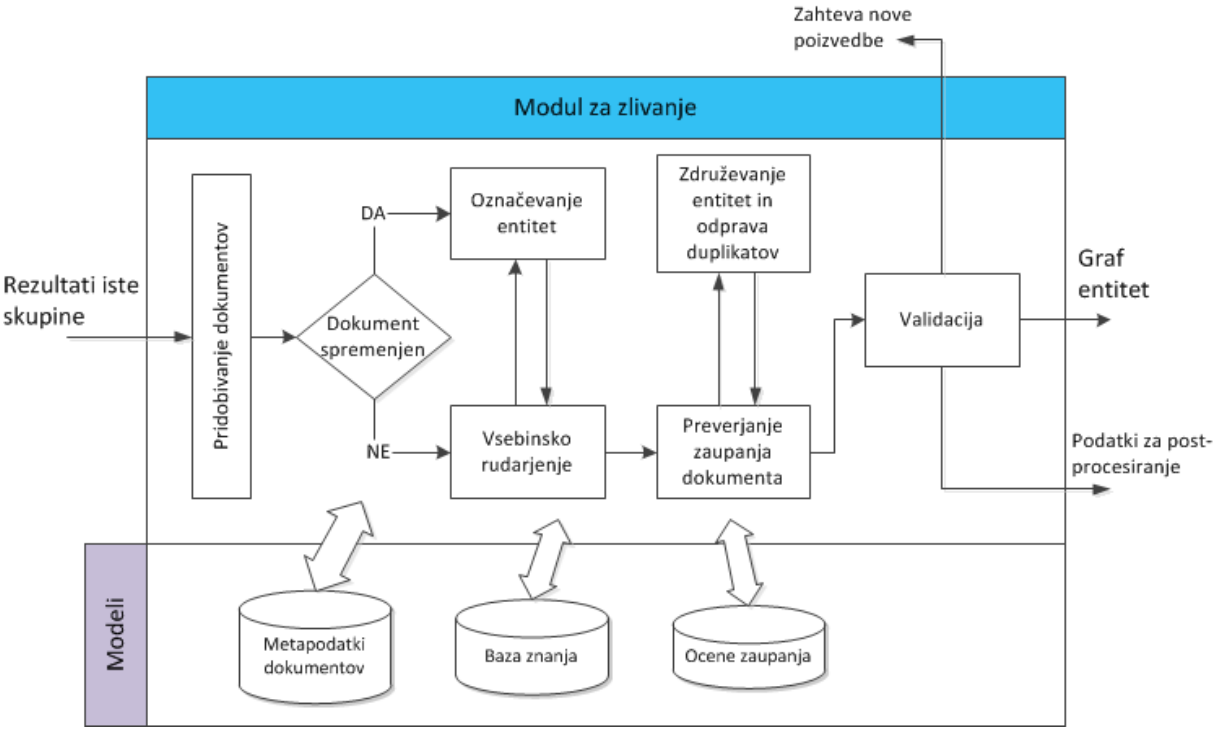
Slika 3: Metaiskalnik

4.3 Modul za zlivanje

Modul za zlivanje v sistemu iCORE predstavlja zadnjo enoto pred posredovanjem podatkov uporabniku. Njegova naloga je, da sprejme metaiskalnikove skupine dokumentov in jih obdela, da pridobi čimveč informacij. Pri tem mora poleg uporabnikovega konteksta upoštevati še kontekst informacij v dokumentih. Če ugotovi, da bi lahko podatke še bolj obogatil, lahko tudi sam oblikuje poizvedbo za modul za obdelavo poizvedb. Kot rezultat naprej posreduje ustvarjene entitete z rezultati metaiskalnika.

Na sliki 4 je prikazano delovanje modula za zlivanje. Pred začetkom obdelave modul najprej prenese dokumente v lokalni pomnilnik sistema iCORE. To se v celotnem obdelovanju poizvedbe v sistemu prvič in edino le zgodi na tem mestu. Pred tem imamo v sistemu le povezave do dokumentov in njihove izvlečke, kot nam jih vrnejo iskalniki v modulu metaiskalnik. Dokument po končani obdelavi še nekaj časa ostane v predpomnilniku zaradi časovne lokalnosti poizvedb. Iz pridobljenih dokumentov s pomočjo **podkomponente za kontekstno iskanje zakonitosti v podatkih** (angl. web content mining) ustrezno zgradimo entitete in jih povežemo glede na medsebojne relacije. Če modeli še ne vsebujejo navodil za pridobitev podatkov iz določenega dokumenta ali se je njegova struktura spremenila, dokument pregledamo z **označevalnikom podatkov/entitet** (angl. data/entity extractor). Ta glede na bazo znanja označi vse dele dokumenta, ki jih razpozna. Posebej se moramo osredotočiti

na entitete, ki naj bi zanimale uporabnika. Uporabnikov kontekst nam je znan že od analize poizvedbe in ga poskušamo vseskozi izkoriščati. Označeni deli dokumenta služijo za ustrezno vsebinsko odkrivanje zakonitosti v dokumentu. Ko imamo dele dokumentov ustrezno označene, iz rezultatov **zgradimo graf**, kjer povezave predstavljajo relacije med označenimi deli. V takšnem grafu je veliko podvajanj ali različnih delov istih entitet. Graf sprejme **podkomponenta za združevanje entitet** (angl. entity resolution) [5] **in odpravljanje duplikatov** (angl. redundancy elimination). Poleg znanega uporabnikovega konteksta moramo na tem mestu razpoznavati še kontekst informacij v dokumentih. Kontekst vsebuje na primer avtorja informacij, časovni okvir, starost podatkov, stopnjo zaupanja, itd. Ko je graf očiščen, **validacijska komponenta** preveri, če je potrebno izvesti še kakšno poizvedbo za bolj natančen graf, sicer podatke predela v splošno obliko (kot je zapis XML) in jih pošlje uporabnikovemu odjemalcu in modulu za post-procesiranje.



Slika 4: Modul za zlivanje

4.4 Modul za post-procesiranje in modeli

Naloga modula za post-procesiranje je, da po vsaki izvedeni poizvedbi zaradi morebitne spremembe posodobi modele. Poleg tega si hrani učne in testne podatke za prilagajanje klasifikatorjev. Omogočati mora tudi ročno vnašanje sprememb v modele za lažjo administracijo sistema. Na voljo ima vse podatke, kot so osnovne in spremenjene poizvedbe, uteži iskalnikov, itd. Po izvedeni poizvedbi ima na voljo tudi vrnjene dokumente in potencialen odziv uporabnika. Rezultati njegovega izvajanja niso časovno kritični, zato lahko za svoje posle uporablja sklad in jih odloženo procesira.

Za vsakega uporabnika posebej si shranjuje zgodovino poizvedb. Te rezultate potem uporabljamo za prilagajanje klasifikatorja za rangiranje dokumentov specifičnega uporabnika (angl. preference learning). Rezultati več uporabnikov služijo učenju globalnega klasifikatorja. Glede na to, kateri iskalniki vračajo za določene poizvedbe bolj relevantne dokumente, določimo tudi uteži za njihovo izbiro v modulu metaiskalnik.

Ponovna gradnja klasifikatorja po vsaki poizvedbi bi bila zelo potratna, zato lahko klasifikatorje posodabljamo, ko pridobimo določen delež novih podatkov. Določimo tudi maksimalno velikost korpusa, iz katerega se učimo. Potem lahko npr., ko pridobimo 30% novih podatkov, stratificirano izberemo 70% starih in skupaj z novimi zgradimo nov klasifikator. Naslednja možnost je, da klasifikator posodobimo, ko pridobimo primer, ki ga trenutni napačno klasificira.

Pri procesiranju dokumentov smo že v modulu za zlivanje pridobili strukturo zapisa. Na tem koraku moramo za dokumente določenega vira posodobiti stopnjo zaupanja. Pri pregledovanju spletnih dokumentov moramo pregledovati spletne obrazce, če vsebujejo nove iskalnike.

Modeli so strukture v sistemu iCORE, ki jih uporabljajo in spreminjajo preostale komponente. Vsebujejo vse znanje, ki ga sistem ima. Hranijo obe bazi znanja (prva namenjena strukturi entitet, druga vsebuje še podatke), kontekste uporabnikov, metapodatke o dokumentih, vzorce za označevanje entitet, seznam in uteži iskalnikov za določeno skupino poizvedb, ocene virov in njihovo zaupanje, naučene modele klasifikatorjev sistema, itd. Vse komponente sistema iCORE uporabljajo modele, spreminjata pa jih le modul za zlivanje in modul za post-procesiranje.

Izgled posameznih modelov je odvisen od metod, ki jih uporabljamo v sistemu.

5 EVALVACIJA

Sistem iCORE ni običajen sistem za priklic podatkov. Običajni sistemi kot rezultat vračajo le relevantne dokumente in krajše izvlečke, medtem ko sistem iCORE iz njih izlušči tudi informacije. Podobne sisteme ocenjujemo glede na časovni odziv na poizvedbo in najbolj pomembno, z uporabnikovim zadovoljstvom. Uporabnik je zadovoljen, če so zanj relevantne informacije prednostno prikazane. Sistem iCORE je sestavljen iz več komponent, zato lahko tudi vsako izmed njih posebej testiramo kot v strojnem učenju.

Za ocenjevanje metod strojnega učenja uporabljamo klasifikacijsko točnost, oceno F, natančnost in priklic. Rangirane rezultate ocenjujemo z mero NDCG (angl. normalized discounted cumulative gain), s pomočjo krivulje natančnost-priklic, iz katere tudi izračunamo mero MAP (angl. mean average precision), ki predstavlja površino pod krivuljo.

Za kakršnokoli testiranje metod ali sistemov potrebujemo nabor označenih učnih in testnih podatkov. Obširni označeni nabori so nastali za namene raziskovalnih skupin ali tekmovanj na konferencah, kjer raziskovalci poskušajo odkrivati boljše algoritme. Zajem takšnih podatkov zahteva veliko časa in ljudi. Pri evalvaciji enostavnejših metod si lahko izgradimo sintetični generator podatkov. V nekaterih primerih lahko uporabimo tudi podatkovne baze, ki že vsebujejo označene podatke in jih samo pretvorimo v želeni format.

Za testiranje iskalnikov obstajajo znani nabori kot sta TREC ali Reuters, za združevanje entitet CiteSeer, ArXiv, BioBase, za označevanje entitet DBPedia, MUC in za rangiranje LETOR ter Yahoo-jev nabor testnih podatkov.

Joachims je svoj sistem za priklic podatkov [2] testiral znotraj svojega raziskovalnega oddelka. Iz dnevnikov je računal odziv uporabnikov glede na to, kateri dokument so izbrali pri določeni poizvedbi. To je tudi uporabil za prilagajanje parametrov klasifikatorja v času delovanja sistema.

Vse podkomponente posebej lahko evalviramo s standardnimi nabori testnih podatkov. Za testiranje celotnega sistema iCORE smo se odločili za nekajmesečno testno obdobje uporabe znotraj

raziskovalne skupine. Če bodo rezultati uspešni, bomo uporabo iCORE-a razširili tudi na nivo organizacije.

6 PRIMERI UPORABE V JAVNI UPRAVI

V zdravstvu: Danes morajo zdravniki ročno iskati pravilne diagnoze in se sami odločati za predpisovanje ustreznih zdravil pacientom. Sistem iCORE bi jim lahko olajšal izbiro in jim tudi predstavil ostale manj relevantne izbire v primerni razdalji. Uporabljal bi lahko svetovne zdravstvene baze podatkov in upošteval nova zdravila ali ugotovljena opozorila, ki jih zdravnik morda še ne pozna. Tudi pri sprejemu novega bolnika bi lahko sistem iCORE uporabili na svetovnem spletu, da bi poskušali ugotoviti, če bolnikov način življenja škoduje zdravljenju ali pa nam hoče prikriti dejstva.

Napredna e-uprava: Sistem iCORE bi lahko uporabili za izgradnjo enostavnejšega oz. bolj preglednega portala e-uprava. Sistem bi pripomogel, da bi uporabnik hitro našel informacije ali ustrezne storitve.

Združevanje več registrov: Pri opravljanju storitev v nekaterih uradih uradnikom predstavljajo težave vpogledi v različne registre, kar tudi prispeva k daljšim čakalnim vrstam in neekonomičnosti poslovanja. Sistem iCORE z avtorizacijo onemogoča zlorabo določenih podatkov. Sistem bi s tem tudi olajšal izpolnjevanje obrazcev tekom delovnega toka in omejil prilaganje potrdil. Pri posodabljanju in povezovanju registrov med seboj bi lahko nudil podporo za identifikacijo podvojitev, ugotavljanje realnih vrednosti z uporabo javno dostopnih virov. Enostaven primer ločenosti baz je urejenost študentske prehrane. Študenti morajo vsako leto ročno prinesti potrdilo o šolanju k okencu, poleg tega sistem ne more odkrivati goljufij, če študentu med letom status preneha.

Poslovna uporaba: Nenapisano pravilo pravi, da je v poslovnih okoljih 80% informacij pridobljenih iz nestrukturiranih virov. Russom je v svoji raziskavi [3] pokazal, da je delež pol- in nestrukturiranih virov v organizaciji 53% in se bo v primerjavi s strukturiranimi še povečeval. Sistem iCORE bo omogočal enostavno pridobivanje in prikazovanje rezultatov iz intraneta, različnih dokumentnih sistemov, e-poštnih sporočil in dokumentov na pomnilnih medijih. Dodana vrednost bodo natančne informacije. Zaposleni bodo lahko hitreje izdelovali poročila, predstavitve, saj jim bo dostop do ustreznih informacij zelo olajšan.

Organi pregona, odkrivanje goljufij: Storilci kaznivih dejanj uporabljajo svetovni splet in priljubljena socialna omrežja. Ne zavedajo se še, da jih lahko s pomočjo tega hitro najdemo. Interpol je letos uspešno izvedel akcijo iskanja »kriminalcev« [4]. Obveščali so prebivalstvo, da so bili na spletu pozorni na sumljive osebe ter statuse na socialnih omrežjih in s takšnim načinom so uspešno našli nekaj iskanih oseb. V prihodnosti so zato napovedali razvoj sistema, ki bo lahko takšna iskanja opravljal sam – to nalogo bi lahko opravljal sistem iCORE. Podoben primer je odkrivanje goljufij, pri katerem lahko bolj kakovostne informacije pomagajo odločiti o potencialni goljufiji. V primeru avtomobilskih nesreč bi sistem iCORE lahko ugotovil, če je osumljenec objavljal kakšne slike, javno komentiral dogodek ali pa se nesreča dejansko sploh ni zgodila.

7 ZAKLJUČEK

Uporabniki postajajo z razvojem tehnologije čedalje bolj zahtevni in od nje pričakujejo vedno več. V današnjem času bi se zdelo nesmiselno graditi nov iskalnik s svojim indeksom, saj so bili v preteklosti razvite številne rešitve, ki jih lahko uporabimo. Trdimo, da trenutnim sistemom manjkajo alternativni načini podajanja poizvedb, ki se lahko kreirajo avtomatsko. Poleg tega smo želeli iz zadetkov, ki nam

jih vrnejo iskalniki, pridobiti čimveč informacij in jih uporabniku predstaviti na način, ki mu je blizu in mu omogočiti, da se lahko po njih enostavno sprehaja. Tako kot smo se privadili, da ima vsak svoj »osebni« mobilni telefon s personaliziranimi nastavitvami, se bomo lahko v prihodnosti privadili na »osebni« sistem iCORE, ki nas bo spremljal na vsakem koraku.

Večji komercialni iskalni sistemi že ponujajo pripravljene vmesnike, s katerimi lahko komuniciramo s svojo programsko opremo. Nekateri imajo že izoblikovano politiko plačevanja v zameno za njihovo uporabo ali pa imajo brez plačila kakšno omejitev, saj je iskalniški marketing eden bolj donosnih. Zaenkrat med iskalniki, ki uporabljajo takšne vmesnike še nismo zasledili podobnega iCORE-u. Večina njih je razvitih kot specializirani iskalnik nad določeno domeno.

V prihodnosti želimo v celoti razviti predlagani sistem. Pri tem bomo tudi testirali, kako se trenutno najboljši algoritmi s posameznih področjih obnašajo pri naših nalogah, in poskušali predlagati njihove izboljšave.

8 VIRI IN LITERATURA

[1] White R.W., Richardson M., Bilenko M., Heath A.P., »Enhancing web search by promoting multiple search engine use,« v zborniku *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ZDA, Redmond, str. 43-50, 2008.

[2] Joachims T., »Optimizing search engines using clickthrough data,« v zborniku *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[3] Russom P., »BI Search and Text Analytics - New Additions to the BI Technology Stack,« v zborniku *TDWI Best Practices Report*, 2007.

[4] (2010) INTERPOL Asks Public To Help Find International Fugitives. Dostopno na: http://www.eurasiareview.com/201007074527/interpol-asks-public-to-help-find-international-fugitives.html.

[5] Bhattacharya I., Getoor L., »Collective entity resolution in relational data,« v zborniku *ACM Transactions on Knowledge Discovery from Data*, 2007.

[6] Bush V., »AsWe May Think,« v zborniku *The Atlantic Monthly*, zv. 176, st. 1, str. 101-108, 1945.

[7] Hu J., Wang G., Lochovsky F., Sun J., Chen, Z., »Understanding user's query intent with wikipedia,« v zborniku *Proceedings of the 18th international conference on World wide web*, China, Beijing, str. 471-480, 2009.

[8] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, R. Lee, »Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections,« v zborniku *The Semantic Web: Research and Applications*, Berlin, Germany, str. 723-737, 2009.

[9] White R.W., Richardson M., Bilenko M., Heath A.P., »Enhancing web search by promoting multiple search engine use,« v zborniku *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ZDA, Redmond, str. 43-50, 2008.

[10] Cugini J. V., Laskowski S., Sebrechts M. M., »Design of 3D visualization of search results: evolution and evaluation,« v zborniku *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, str. 198-210, 2008.

[11] Bowman D. A., Coquillart S., Froehlich B., Hirose M., Kitamura Y., Kiyokawa K., Stuerzlinger W., »3D User Interfaces: New Directions and Perspectives,« v *IEEE Computer Graphics and Applications*, str. 20-36, 2008.

[12] Ashok J., Ramineni K., Rajan E. G., »Beyond Information Retrieval: A Survey,« v *Journal of Theoretical and Applied Information Technology*, zv. 15, št. 1, 2010.

[13] Chen M.-Y., Chu H.-C., Chen Y.-M., »Developing a Semantic-Enable Information Retrieval System.« v *Expert Systems with Applications*, zv. 37, št. 1, str. 322-340, 2010.

[14] Cool C., Spink A., »Issues of Context in Information Retrieval,« v *Information Processing and Management*, zv. 38, št. 5, 2002.

[15] Ingwersen P., Belkin N., »Information Retrieval in Context,« v *ACM SIGIR Forum*, zv. 38, št. 2, str.. 50-52, 2004.

[16] Jimeno-Yepesa A., Berlanga-Llavorib R., Rebholz-Schuhmanna D., »Ontology Refinement for Improved Information Retrieval« v *Information Processing and Management*, zv. 46, št. 4, str. 426-435, 2010.

[17] Manoj M., Elizabeth J., »Information Retrieval on Internet Using Meta-Search Engines: A Review,« v *Journal of Scientific and Industrial Research*, 2008.

[18] Teevan, J. and Dumais, S.T. and Horvitz, E., »Personalizing search via automated analysis of interests and activities,« v *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, str. 449-456, 2005.

[19] Bhattacharya, I. and Getoor, L., »Collective entity resolution in relational data,« v *ACM Transactions on Knowledge Discovery from Data*, 2007.

[20] Jin, H. and Ning, X. and Jia, W. and Wu, H. and Lu, G., »Combining weights with fuzziness for intelligent semantic web search,« v *Knowledge-Based Systems*, zv. 21, št. 7, str. 655-665, 2008.