ISKALNIK ZA SLOVENSKE IN ANGLEŠKE DOKUMENTE NA SVETOVNEM SPLETU

Jure Dimec[*], *Sašo Džeroski*[**], *Ljupčo Todorovski*[*], *Dimitar Hristovski*[*]
[*]Inštitut za biomedicinsko informatiko Medicinske fakultete,
Vrazov trg 2, 1105 Ljubljana, tel. 31 32 33, fax. 311 540,
{jure.dimec|ljupco.todorovski}@mf.uni-lj.si, hristovski@ibmi.mf.uni-lj.si
[**]Inštitut Jožef Stefan, Jamova 39, 1000 Ljubljana
saso.dzeroski@ijs.si

POVZETEK

Predstavljamo razvoj informacijskih orodij za organiziranje in iskanje slovenskih in angleških medicinskih dokumentov, dostopnih na Svetovnem spletu. Orodja, zaenkrat še v testni fazi, omogočajo avtomatsko opisovanje vsebine dokumentov, iskanje z iskalnimi zahtevami v naravnem jeziku in rangiranje zadetkov po izračunani relevantnosti. Iskalnik se zaveda stanja poizvedbe, zato lahko iskalec z iskanjem s povratno zanko postopno izboljšuje kvaliteto iskanja.

ABSTRACT

The development of information tools for the organization and searching of Slovene and English medical documents is presented. The tools, presently in testing phase, provide automatic subject description of documents, searching with natural language queries and ranking of search hits according to their relevance. The search engine is state-full allowing searcher to use relevance feedback in order to perform incremental improvement of search quality.

1 UVOD

Svetovni splet (WWW) je pomemben tudi kot vir informacij, uporabnih pri raziskovalnem in razvojnem delu. V poplavi dokumentov, namenjenih različnim uporabniškim skupinam, postaja vedno pomembnejše njihovo urejanje v digitalnih knjižnicah. Ožji pomen pojma digitalna knjižnica lahko opišemo kot zbirko dokumentov, dostopnih preko omrežja, na katerih je bila opravljena neka vrsta selekcije, ter kot zbirko informacijskih orodij za odkrivanje in pregledovanje teh dokumentov. Med informacijskimi orodji so najpomembnejši iskalniki.

Digitalne knjižnice v današnji rudimentarni obliki le deloma rešujejo problem obilice nepreverjenih dokumentov na Internetu. Dobro informacijsko orodje, namenjeno zahtevnemu iskalcu, bi moralo omogočiti kvalitetno preiskovanje informacijskih virov, zbranih v lokalni digitalni knjižnici in istočasno nuditi individualiziran dostop do množice potencialno koristnih dokumentov na Svetovnem spletu.

V prispevku smo se lotili obeh nalog. Raziskati želimo (a) možnosti gradnje baze z avtomatsko zgrajenimi vsebinskimi opisi nestrukturiranih in strukturiranih dokumentov, (b) razviti iskalnik, namenjen preiskovanju baze z iskalnimi zahtevami v naravnem jeziku, in (c) z metodami strojnega učenja zgraditi uporabniške profile z opisom uporabnikovih informacijskih potreb, v obliki, ki bi omogočala avtomatizirano iskanje relevantnih dokumentov na Svetovnem spletu. Velika večina naravoslovnih dokumentov pri nas je v slovenščini ali angleščini, zato smo se v jezikovno odvisnih postopkih gradnje baze dokumentov in iskalnika omejili na ta jezika.

Prispevek začenjamo z opisom postopkov avtomatskega indeksiranja. V nadaljevanju opisujemo gradnjo iskalnika in predstavimo primer poizvedbe po testni bazi. Uporabo strojnega učenja za učenje relevantnosti angleških in slovenskih medicinskih dokumentov iz primerov predstavljamo na koncu prispevka.

2 BAZA VSEBINSKIH OPISOV SLOVENSKIH IN ANGLEŠKIH MEDICINSKIH DOKUMENTOV

Dinamična narava in obseg informacijskih virov na Internetu praktično onemogoča opisovanje vsebine dokumentov s ključnimi besedami ali deskriptorji, ki bi jih določal informacijski strokovnjak. Naloga je izvedljiva le z metodami avtomatskega indeksiranja, ki jih uporabljajo tudi pri gradnji baz, na katerih slonijo znameniti iskalniki AltaVista, Excite, Infoseek in drugi.

Avtomatsko indeksiranje običajno poteka v treh korakih:

1. *blokiranje*, pri katerem iz postopka izključimo besede z minimalno količino informacije (stop-words),
2. *krnjenje*, pri katerem različne oblike besede poenotimo na skupni krn, in
3. računanje *povedne moči* krna - njegovega deleža v zalogi informacije, ki jo ima dokument.

Pri gradnji baze uporabljamo vse tri korake za dokumente v obeh jezikih, vendar tu opisujemo le postopke za slovenščino. Za avtomatsko indeksiranje angleških dokumentov smo uporabili metode, pogosto opisane v strokovni literaturi [1].

Testno bazo sestavljajo polni dokumenti iz dveh strokovnih revij, dostopnih tudi v elektronski obliki -

slovenska izdaja revije JAMA (Journal of the American Medical Association) in ISIS, glasilo slovenske Zdravniške zbornice) - ter bibliografski zapisi in izvlečki iz nacionalne baze Biomedicina Slovenica.

2.1 Blokiranje in krnjenje slovenskih dokumentov

Vsaka beseda iz dokumenta do neke mere zastopa njegovo vsebino in je zato načeloma kandidat za indeksni termin. Izjema so besede, ki jih imenujemo *blokirane*. To so tiste besede, ki so v bazi dokumentov najenakomerneje porazdeljene in zato tudi v posameznih dokumentih nosijo najmanjšo količino informacije. Gre za predstavnike nekaterih besednih vrst, kot so predlogi, prislovi, zaimki ipd. V seznamu blokiranih besed za slovenščino je trenutno blizu 1600 besed in besednih oblik. Besede, ki jih najdemo v tem seznamu izključimo iz nadaljnje obdelave.

Tudi *krnjenje* (stemming) je jezikovno-odvisen postopek in od njega je v največji meri odvisna kvaliteta avtomatskega indeksiranja in, posledično, kvaliteta iskanja. Pri krnjenju poskušamo najti niz znakov, imenujemo ga *krn*, ki lahko predstavlja vse oblike neke besede in istočasno to besedo loči od vseh ostalih. Pogosto, vendar ne nujno, krn ustreza korenu besede. Krnjenje je še posebej pomembno pri avtomatskem indeksiranju besedil v jezikih z bogato morfologijo, kakršna je tudi slovenščina.

Do sedaj sta nam znana dva poskusa izdelave statističnih algoritmov za krnjenje slovenskih besedil, namenjena gradnji tekstovnih baz [2, 3]. Oba sta temeljila na obsežnih seznamih končnic. Prvi algoritem je bil enostaven: za vsako besedo v postopku je v seznamu 1205 končnic poiskal najdaljšo končnico, ki je ustrezala zaključku besede in na mestu ujemanja razcepil besedo na krn in odbitek. Edino dodatno pravilo je bila najmanjša dovoljena dolžina krna. Drugi, Popovičev algoritem [3], je bil bistveno bolj zapleten. 5276 končnic v seznamu je razdelil na osem skupin, vsako s svojim načinom krnjenja. Poleg enostavnega ujemanja in pravila o najmanjši dolžini krna je moralo končno zaporedje znakov v krnu ustrezati določenemu vzorcu, značilnemu za skupino, v katero je sodila končnica. Algoritem je uporabljal tudi pravila o popravljanju krnov in pravila o izjemah.

Popovičev algoritem je bil relativno uspešen pri krnjenju slovenskih besedil s splošno vsebino. Njegova osnova pomanjkljivost, tudi sicer pogosta pri podobnih algoritmih, je tendenca k premočnemu krnjenju, pri katerem dve ali več podobnih besed prispeva isti krn. Ko smo ga preizkusili pri krnjenju slovenskih medicinskih besedil, se je izkazalo tudi, da je manj primeren za strokovni medicinski jezik, v katerem so zelo pogoste tujke in poslovenjeni izrazi, temelječi na grški ali latinski osnovi.

V okviru projekta smo razvili nov, poenostavljen algoritem za krnjenje slovenskih besedil. Pri analizi rezultatov krnjenja s Popovičevim algoritmom smo ugotovili, da gre največji del primerov premočnega krnjenja na račun brisanja soglasnikov na koncu krnov. Krni podobnih besed se pogosto ločijo le po zaključnih soglasnikih ali soglasniških skupinah, kar je v skladu z dejstvom, da soglasniki v jeziku nosijo večjo količino informacije od samoglasnikov. Analiza pojavljanja soglasniških parov v zaključkih besed je omogočila sestavo enostavnih pravil za selektivno pretvorbo soglasniških parov v posamezne soglasnike ali prazne nize.

Novi algoritem za krnjenje je dvostopenjski in uporablja dva seznama končnic. V prvem seznamu so samo končnice, ki razcepijo besedo med soglasnikom in samoglasnikom, seveda če pozicija ustreza ostalim pogojem. Algoritem odreže najdaljšo končnico iz seznama, ki se ujema z zaključkom besede, nevarnosti premočnega krnjenja pa smo se izognili z novim pravilom o najmanjši dovoljeni dolžini krna. Pravilo postavlja premično mejo odreza in, poenostavljeno rečeno, dovoljuje odrez tem daljše končnice, čim daljši krn pri tem ostane.

Drugi korak poteka v zanki, ki zaporedoma pretvarja končne soglasniške pare. Zanka se zaključi, ko na koncu besede ostane en sam soglasnik, ali pa v bazi končnic ni pravila za končni soglasniški par.

Algoritem temelji izključno na preprosti statistični analizi besed v učni množici slovenskih medicinskih besedil in nima namena modelirati jezikovnih zakonitosti nastajanja besedne oblike. Rezultati zato niso optimalni, vendar menimo, da smo dosegli učinkovito razmerje med kvaliteto krnjenja in računsko potratnostjo postopka.

2.2 Računanje povedne moči besednih krnov

Povedno moč opisujemo kot delež informacije, ki jo besedni krn prispeva k skupni zalogi informacije v dokumentu. Odvisna je predvsem od frekvence krna v dokumentu in bazi dokumentov. Osnovni razmislek je preprost: (a) pogostejša ko je beseda v dokumentu, pomembejša je vsebina, ki jo beseda zastopa, in (b) redkeje ko se beseda pojavlja v bazi, bolj loči dokumente, v katerih je prisotna, od ostalih v bazi. Pomemben vpliv na frekvenco besede v dokumentu ima seveda dolžina dokumenta, kar je treba upoštevati pri računanju njene povedne moči.

Neizogibna lastnost Interneta je dinamičnost, zato je pogostnost besednega krna v bazi nemogoče določiti med postopkom avtomatskega indeksiranja. V tej fazi besednemu krnu lahko določimo frekvenco v dokumentu, dokončno povedno moč pa izračunamo med iskanjem, glede na vsebino baze v tistem trenutku. Do dokumentov, ki jih indeksiramo (zaenkrat članki v obeh revijah in bibliografski zapisi), programi za indeksiranje dostopajo preko Spleta, zato lahko frekvence besed dopolnimo tudi z informacijami, implicitno vsebovanimi v oznakah HTML. Vsaj načeloma je za vsebino dokumenta manj

pomembna beseda, ki jo najdemo v običajnem besedilu, od tiste, ki je poudarjena, ta pa spet manj od besede, ki izvira iz enega od naslovov.

3 ISKANJE

3.1 Rangiranje rezultatov iskanja

Od sodobnega iskalnika dokumentov pričakujemo možnost sestavljanja iskalnih zahtev v naravnem jeziku in razvrščanje rezultatov iskanja po izračunani relevantnosti. Relevantnost dokumenta je definirana kot mera podobnosti med iskalno zahtevo in dokumentom, v splošnem pa jo izračunamo kot seštevek povednih moči besednih krnov, skupnih iskalni zahtevi in dokumentu. Statistične metode računanja povednih moči in relevantnosti sodijo večinoma v eno od dveh skupin - metode vektorskega prostora in probabilistične metode. Dober pregled metod vektorskega prostora najdemo v [4]. Pri našem delu smo uporabili Croftovo probabilistično metodo, objavljeno v [6] in dopolnjeno z vrednostjo oznake v HTML, pri kateri sorodnost med dokumentom *j* in iskalno zahtevo *k* izračunamo kot:

$$\text{sorodnost}\ (\text{j},\text{k}) = \sum_{i=1}^{Q} (C + IDF_i) \times f_{ij}\ ,$$

$$IDF_i = \log \frac{N - n_i}{n_i},$$

$$f_{ij} = K + (1 - K) \frac{frekv_{ij}}{najv\_frekv_j} \times w_{HTML},$$

kjer je

Q	=	število besednih krnov, skupnih dokumentu *j* in iskalni zahtevi *k*,
$frekv_{ij}$	=	frekvenca krna *i* v dokumentu *j*,
$najv\_frekv_j$	=	največja frekvenca katerekoli besede v dokumentu *j*,
N	=	število dokumentov v bazi,
n_i	=	število dokumentov s krnom *i*,
w_{HTML}	=	vrednost oznake v HTML,
C, K	=	konstanti, namenjeni prilagajanju postopka različnim lastnostim baze.

3.2 Iskanje s povratno zanko

Podobne metode razvrščanja zadetkov uporabljajo vsi veliki iskalniki na Spletu. Analize kažejo, da je večina iskalnih zahtev kratkih, z majhnim številom besednih krnov, zato je zelo pogosto kvaliteta rangiranja daleč od optimalne. Iskalec je v takem primeru prisiljen, da iskanje ponovi s spremenjeno iskalno zahtevo, ali pa pregleda velik del rangiranega seznama zadetkov. Pri klasičnih bazah popolnih besedil poznamo učinkovito metodo, ki omogoča postopno zgoščevanje relevantnih dokumentov na vrhu ranžirne vrste. Metoda, imenujemo jo *iskanje s povratno zanko* (relevance feedback), temelji na interaktivnem dialogu z iskalcem [5, 6]. Iskalniki na spletu te metode ne uporabljajo, ker zaradi interaktivnosti zahteva unikatno identifikacijo iskalca ali neprekinjeno (state-full) iskalno seanso, česar z običajnimi CGI programi in HTTP protokolom ni mogoče zagotoviti.

Naš iskalnik ves čas pozna identiteto iskalca, zato smo lahko implementirali tudi iskanje s povratno zanko, ki poteka v treh korakih:

1. iskalec opravi prvo, enostavno iskanje,
2. pregleda nekaj najvišje uvrščenih dokumentov in označi tiste med njimi, ki so relevantni,
3. sistem avtomatsko reformulira iskalno zahtevo in opravi novo iskanje.

Iskanje z reformulirano iskalno zahtevo vključi v seznam zadetkov nove dokumente, pozicije že poiskanih dokumentov pa se spremenijo tako, da relevantni dokumenti splezajo proti vrhu seznama. Po vsakem iskanju iskalec ponovi pregledovanje najbolje uvrščenih dokumentov in označi relevantne, tako da se drugi in tretji korak zanke ponavljata, dokler so na vrhu seznama še pozitivne spremembe, ali pa iskalec ne odneha.

Jedro postopka je reformulacija iskalne zahteve. Pri tem sistem ponovno izračuna povedne moči vseh krnov iz dokumentov, ki jih je iskalec označil kot relevantne. Tako lahko nekateri besedni krni v iskalni zahtevi, ki se pretežno pojavljajo v relevantnih dokumentih, dobijo večjo težo, dodajo pa se tudi nekateri novi. Povedno moč *w* besede *i* v dokumentu *j*, relevantnem za iskalno zahtevo *k*, tako izračunamo kot:

$$w_{ijk} = (C + \log \frac{p_{ij}(1 - q_{ij})}{(1 - p_{ij})q_{ij}}) \times frekv_{ik} \times w_{HTML}$$

kjer je

p_{ij}	=	verjetnost pojavljanja besede *i* v dokumentih, *relevantnih* za iskalno zahtevo *j*,

Simbol  pomeni, da dokument ni bil označen kot relevanten, s simbolom  pa so označeni dokumenti, za katere je iskalec menil, da so relevantni in ki so prispevali podatke za reformulacijo iskalne zahteve.

4 ČEMU ŠE EN ISKALNIK?

Veliki, javni iskalniki, kot AltaVista, Excite in drugi, po navedbah avtorjev pokrivajo pretežni del Svetovnega spleta, s tem pa tudi strani na slovenskih strežnikih. Zakaj se torej lotevamo razvoja novega iskalnika, za katerega je že pred rojstvom jasno, da velikih ne more nadomestiti?

Avtomatsko opisovanje vsebine dokumentov je jezikovno odvisen postopek, ki v največji meri določa kvaliteto iskanja. *Lingua franca* spletnih strani je angleščina, čeprav nekateri iskalniki zmorejo tudi postopke za zelo omejeno število drugih velikih jezikov. Slovenski dokumenti v veliki meri ostajajo nepoiskani.

Slika 1: Rezultati iskanja na iskalno zahtevo “okužbe z virusom HIV”. Med pregledanimi dokumenti sta bila dva označena kot relevantna.

Zaenkrat nam ni znan obstoj delujočega slovenskega iskalnika, ki bi omogočal kaj več kot iskalne zahteve v obliki ročno krnjenih ključnih besed, čeprav so že obstajale testne implementacije, vendar ne v spletnem okolju [3, 7].

Menimo, da naš iskalnik lahko vsaj deloma zapolni to praznino.

Bazo vsebinskih opisov in kazalcev na dokumente, iskalnik in module za odkrivanje dokumentov smo si zamislili kot orodje za uporabo digitalne knjižnice medicinskih dokumentov. V prihodnosti načeloma lahko odpade vsebinska omejitev, kajti večina postopkov, ki smo jih razvili je vsebinsko neodvisna. Zaenkrat se osredotočamo na vključevanje preverjenih dokumentov v slovenščini in angleščini, uporabnih pri strokovnem, študijskem in raziskovalnem delu v slovenskem zdravstvu. Ena od posledic navedenih samoomejitev je tudi relativno majhna baza.

Slika 2: Rezultati iskanja s povratno zanko, sproženim v situaciji na sliki 1. Vidni so tudi novi besedni krni, dodani prejšnji iskalni zahtevi.

Velikost baze vsebinskih opisov nam omogoča izpeljavo postopkov, ki si jih pri velikih iskalnikih zaenkrat še ne morejo privoščiti. Zavedamo se, da dobrega iskanja ni mogoče opraviti v enem koraku, zato smo velik del pozornosti posvetili sposobnosti iskalnika, da se zaveda identitete iskalca in trenutnega stanja poizvedbe. S tem so bili tudi dani pogoji za uvedbo iskanja s povratno zanko.

5 UČENJE RELEVANTNOSTI DOKUMENTOV IZ PRIMEROV

Pri učenju relevantnosti dokumentov iz primerov poskušamo avtomatizirano modelirati informacijske potrebe posameznega uporabnika preiskovalnega sistema. Iz primerov dokumentov, ki so označeni kot relevantni ali nerelevantni, z metodami za strojno učenje zgradimo

model, ki ga lahko uporabimo za ocenjevanje relevantnosti drugih dokumentov. Posamezna informacijska potreba lahko v grobem ustreza iskalni zahtevi, razlika je v tem, da nam zahteve ni treba formulirati eksplicitno temveč za to lahko uporabimo primere relevantnih in nerelevantnih dokumentov.

Strojno učenje so za učenje relevantnosti dokumentov na Spletu uporabili v okviru projekta WebWatcher [7], za modeliranje informacijskih potreb posameznega uporabnika pa v okviru projekta Personal WebWatcher [8]. Kljub številnim raziskavam uporabe strojnega učenja v tovrstne namene nismo zasledili uporabe strojnega učenja za obravnavo dokumentov v slovenščini.

V prispevku smo preizkusili uporabo metod strojnega učenja na problemu učenja relevantnosti angleških in slovenskih dokumentov iz primerov. Uporabili smo zbirko dokumentov omenjeno v drugem razdelku, v kateri so dokumenti označeni glede na relevantnost za vsako od 50 testnih iskalnih zahtev [9]. Vsako od iskalnih zahtev obravnavamo kot posebno informacijsko potrebo in učni problem, kjer je vsak od dokumentov v zbirki primer: pozitiven, če je relevanten za podano iskalno zahtevo in negativen, če ni.

Zbirka vsebuje 335 izvlečkov v angleščini ter njihove slovenske prevode. Pri vsakem učnem problemu smo vseh 770 dokumentov obravnavali naenkrat. Na začetku smo uporabili zelo enostavno predstavitev dokumentov kot množic besed, nato pa smo dokumente obravnavali kot množice krnov.

Za učenje smo uporabili sistem za strojno učenje relacij TILDE [10]. Za razliko od sistemov za atributno strojno učenje, ki uporabljajo predstavitev dokumentov s fiksno dolžino, sistemi za učenje relacij lahko uporabijo manj restriktivne oz. bolj bogate predstavitve. Sistem TILDE generira logična odločitvena drevesa, ki jih v danem primeru lahko prevedemo na urejene sezname pravil.

Oglejmo si kot primer iskalno zahtevo št. 10, ki se glasi: "Nastanek, diagnostika in zdravljenje ulkusa, še posebej razjede želodca in dvanajstnika" oz. "The origin, diagnosis and treatment of ulcer, especially duodenal and gastric ulcerations". Iz primerov za to zahtevo (od 770 je vsega 8 relevantnih dokumentov, 4 v slovenščini in 4 v angleščini), TILDE generira naslednje drevo:

```
ulcer ?
+--yes: rel_10
+--no:  ulkusa ?
        +--yes: rel_10
        +--no:  not_rel_10
```

Pomen drevesa je naslednji: "Če v dokumentu nastopa beseda `ulcer` potem je dokument relevanten. Dokumenten je tudi relevanten če v njem beseda ulcer ne nastopa, vendar nastopa beseda `ulkusa`, sicer je pa dokument nerelevanten."

Pri zahtevi 14, ki se glasi: "Kirurško (operativno) zdravljenje zlomov kosti" oz. "Surgical (operative) treatment of bone fractures" se TILDE nauči drevesa z naslednjim pomenom: "Dokument je relevanten, če v njemu nastopa kakšna od besed `fracture`, `zlome`, `calcaneal`, `zlomih` ali `zlomov`". Drevo napačno klasificira štiri relevantne dokumente kot nerelevantne in en relevanten dokument kot nerelevanten.

Oglejmo si še zahtevo št. 31, ki se glasi: "Uporaba ultrazvoka v diagnostiki" oz. "Use of ultrasound in diagnosis". Iz primerov za to zahtevo (od 770 je 22 relevantnih dokumentov, 11 v slovenščini in 11 v angleščini), TILDE generira drevo s pomenim: "Dokument je relevanten, če v njemu nastopa kakšna od besed `ultrazvočni`, `sonography`, `ultrazvokom`, `ultrasound`, `echocardiographic`, `ehokardiografija`, `hoechst`, `laser`, ali `ultrazvočno`, sicer je nerelevanten." Drevo napačno klasificira tri relevantne dokumente kot nerelevantne in dva relevantna dokument kot nerelevantna.

Pri drevesu za zahtevo št. 14 se pojavi več besed, ki izvirajo iz besede zlom (zlome, zlomih, zlomov). Pri drevesu za zahtevo št. 31 pa se pojavi več besed, ki izvirajo iz besede ultrazvok (ultrazvočni, ultrazvokom, ultrazvočno). To nakazuje smiselnost uporabe predstavitve dokumentov s krni, ki bi predvidoma dala kot rezultat manjša in bolj zanesljiva drevesa za ocenjevanje relevantnosti dokumentov.

Pri uporabi predstavitve dokumentov s krni dobimo pri iskalni zahtevi 14 drevo s pomenom: "Dokument je relevanten, če v njemu nastopa krn `zlom` ali krn `calcaneu`. Sicer pogledamo če v dokumentu nastopa krn `fractur`: če nastopa in v dokumentu ni ne krna `bas` ne krna `manag`, je dokument relevanten. Sicer je dokument relevanten če v njemu nastopata krna `spin` in `oper`." Drevo s krni napačno klasificira le en nerelevanten dokument kot relevanten.

Pri iskalni zahtevi 31 se TILDE nauči drevesa s pomenom: "Dokument je relevanten, če v njemu nastopa kakšen od krnov `ultrazvoč`, `digit`, `ultrasound`, `echocardiograph`, `sonographi`, `ehokardiograf`, ali lh. Če nastopa v dokumentu krn `cist`, potem je dokument relevanten če v njem ni krna `premer`." Drevo s krni napačno klasificira dva nerelevantna dokumenta kot relevantna.

Po pričakovanju drevesa naučena iz dokumentov prestavljenih s krni bolje ločijo med relevantnimi in nerelevantnimi dokumenti. V zgornjih dveh primerih zagrešijo drevesa s krni manjše stevilo napak. Poleg tega noben od relevantnih dokumentov ni bil klasificiran kot nerelevanten (kar ni bilo primer pri drevesih z besedami). Slednje je pomembno zaradi majhnega števila relevantnih dokumentov.

5. NAČRTI ZA NADALJNJE DELO

Informacijsko orodje, ki ga predstavljamo, je v zgodnji testni fazi. Med najočitnejšimi pomanjkljivostmi sta dokaj okoren uporabniški vmesnik in slabo pregledna predstavitev rezultatov iskanja, predvsem pa počasnost,

kar določa prioritetne naloge. V nadaljevanju se nameravamo posvetiti tudi segmentaciji dokumentov, ki bi omogočala selektivno iskanje krajših, vsebinsko bolj homogenih delov dokumentov. V ta sklop nalog sodi tudi avtomatsko odkrivanje jezika posameznih segmentov.
Pri učenju relevantnosti dokumentov iz primerov z metodami strojnega učenja bi kazalo raziskati možnosti vključevanja postopka učenja v iskanje s povratno zanko. Zanimiva bi bila tudi uporaba predznanja v obliki semantične mreže medicinskih pojmov. Končno bi bilo treba v prihodnjih poskusih učenja relevantnosti dokumentov upoštevati še težo oz. povedno moč posameznih krnov.

ZAHVALA

Delo, predstavljeno v prispevku, je v okviru podprojekta "Inteligentno shranjevanje in iskanje slovenskih in angleških medicinskih dokumentov na Internetu" . Podrojekt sodi v projekt INCO COPERNICUS 960 154 Cooperative Research in Information Infrastructure (CRII), ki ga financira Evropska Unija.

ACKNOWLEDGEMENT

The work presented in this paper was supported by the subproject "Intelligent information storage and retreival of Slovene and English Medical Documents on the Internet" of the INCO COPERNICUS Project 960 154 ooperative Research in Information Infrastructure (CRII), funded by the European Union.

6 VIRI

1. Porter MF. An algorithm for suffix stripping. Program. 14. 1980:130-7.
2. Dimec J. Računalniška analiza slovenskega informacijskega jezika v biomedicini. Magistrsko delo. Ljubljana: Medicinska fakulteta, 1989; 77.
3. Popovič M, Willett P. Processing of documents and queries in a Slovene language free text retrieval system. Literary and Linguistic Computing. 5. 1990:183-90.
4. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management. 24. 1988:513-23.
5. Harman D. Relevance feedback and other query modification techniques. In: Frakes WB, Baeza-Yates, editors. Information Retrieval. Data Structures & Algorithms, Englewood Cliffs: Prentice hall, 1992; 241-63
6. Croft WB. Experiments with representation in a document retrieval system. Research and Development in Information Technology, 1983; 2(1)1-21.
7. Joachims T, Freitag D, Mitchell TM. Web Watcher: A tour guide for the WWW. Proc. IJCAI-97, 15th Intl. Joint Conference on Artificial Intelligence, 1997.
8. Mladenič D. Personal WebWatcher: Implementation and design. Delovno poročilo IJS-DP 7472, 1996.
9. Dimec J. Združevanje informacij z analizo povedne moči različnih vrst slovenskih medicinskih besedil in možnosti njihovega iskanja z ne-Boolovimi metodami. Doktorsko delo. Ljubljana: Medicinska fakulteta, 1995; 108.
10. Blockeel H, De Raedt L. Experiments with Top-down Induction of Logical Decision Trees. Artificial Intelligence, 1998 (v tisku).