UDK 025.036

# Razvoj i izazovi metoda automatskog predmetnog označavanja

****

# Development and challenges of automatic indexing

**Lejla Hajdarpašić**

Filozofski fakultet Univerziteta u Sarajevu

Odsjek za komparativnu književnost i bibliotekarstvo

Sarajevo, Bosna i Hercegovina

lejla.hajdarpasic@ff.unsa.ba

***Sažetak***

*Sadržajna obrada kao neupitno važan vid konsolidacije informacija posljednjih je desetljeća obogaćena brojnim automatskim metodama predmetnog označavanja koje koegzistiraju sa manualnim metodama predmetnog označavanja te imaju velike potencijale ali istovremeno i neka nerazriješena pitanja glede kvaliteta označavanja rastućeg broja elektronskih izvora informacija. Sa automatski generiranim indeksima u vezi, ovaj će rad kroz prizmu reprezentativnih metoda automatskog predmetnog označavanja i njihova razvoja pokušati osvijetliti neke od izazova procesa automatskog predmetnog označavanja te zaključno primijetiti da su u principima metoda automatskog predmetnog označavanja sadržane i naznake budućeg razvoja u ovoj domeni. S unapređenjima metoda automatskog predmetnog označavanja u vezi, kombinacija automatskog predmetnog označavanja sa bazom znanja pojedinog područja se treba prepoznati kao posebno obećavajuća.*

***Ključne riječi**: automatsko indeksiranje, statističke metode indeksiranja, obrada prirodnog jezika, ekspertni sistemi*

***Abstract***

*Content analysis as unquestionably important aspect of information consolidation in recent decades was enriched with many automatic indexing methods that coexist with manual indexing methods and have great potential but at the same time some unsolved questions concerning the quality of electronic resources indexing. In regard to automatically generated indexes and through the prism of the representative automatic indexing methods and their development this article will try to illuminate some of the challenges underlying the process of automatic indexing and notice that the underlying principles of the automatic indexing methods contain indications of future developments in this area. Concerning the improvement of automatic indexing methods the combination of automatic indexing and knowledge base of individual field should be identified as particularly promising.*

***Keywords:** automatic indexing, statistical indexing methods, natural language processing, expert systems*

## Uvodna razmatranja

Automatsko predmetno označavanje,[1] pod kojim se popularno podrazumijeva bilo koja metoda kojom se dokument podvrgava algoritamskim operacijama s ciljem izvlačenja termina i/ili fraza koje prezentiraju predmet, temu ili osobine dokumenta,[2] jednako kao i manualno, ima za cilj unaprijediti proces pretraživanja informacija, udovoljiti onim informacijskim potrebama koje se artikuliraju predmetnim pretraživanjem materijala i zadovoljiti temeljna načela indeksiranja koja se tiču principa iscrpnosti, specifičnosti te dosljednosti. Iako je s tim u vezi još od 1950. godine, kada su se kao najjednostavniji koncept automatskog predmetnog označavanja pojavili KWIC (*engl.* Keyword in Context) indeksi, automatsko predmetno označavanje imalo tendenciju postati izuzetno sofisticirano rješenje, ono ni danas (a u kontrastu sa manualnim indeksiranjem, to jest rezultatima predmetnog označavanja koji nastaju ljudskom intervencijom) nije odveć fleksibilno, sporo se prilagođava novoj terminologiji, pa se otuda i njegovim kapacitetima koristi u različite svrhe. Automatsko predmetno označavanje se često koristi kao pomoć / priprema manuelnom predmetnom označavanju ili kao pomoć / kontrola manuelnom predmetnom označavanju, odnosno, automatsko predmetno označavanje se u ponekim slučajevima u potpunosti realizira (dakle, neovisno od manuelnog predmetnog označavanja). Takva primjena automatskog predmetnog označavanja, osobito koegzistira-

1 Prerađeno poglavlje iz završnog diplomskog rada: Hajdarpašić, Lejla. Metode automatskog predmetnog označavanja (indeksiranja): završni diplomski rad. Sarajevo: Filozofski fakultet, 2011.

2 Wellisch, Hans. H. Glossary of terminology in abstracting, classification, indexing, and thesaurus constructions. The American society of Indexers, 2000. Str. 10

nje manualnih i automatskih metoda označavanja, nije iznenađujuća jer automatsko predmetno označavanje pokriva širok spektar tehnika koje se kreću u rasponu od jednostavnih statističkih metoda do izuzetno zahtjevnih lingvističkih metoda (kojima se sadržaji dokumenata mogu posve automatski označavati, dakle bez referiranja na neki eksterni izvor znanja kakav je, recimo, tezaurus ili se mogu na kontroliran način označavati, putem uspostavljanja konceptualnih veza između termina pojedinog dokumenta i jedinica tezaurusa). Pomenute metode različito pristupaju problematici automatske predmetne obrade, njihovom se primjenom postiže različita kvaliteta označavanja, a u njihovim su principima sadržane i naznake budućeg razvoja u ovoj domeni.

## Statistički (jednostavni) model automatskog predmetnog označavanja

Automatsko se predmetno označavanje u određivanju sadržaja dokumenta koristi raznolikim pristupima u vezi s kojima razlikujemo nekoliko temeljnih: statistički model automatskog predmetnog označavanja (te sve metode kojima se ovaj model pokušao unaprijediti), sintaksički model predmetnog označavanja te semantički model automatskog predmetnog označavanja. Automatsko predmetno označavanje koje je utemeljeno na statističkom modelu određuje značenje teksta na temelju frekvencije pojavljivanja pojedinih riječi u tekstu. Ovaj se jednostavni postupak automatskog označavanja odvija u nekoliko glavnih etapa iliti normalizacijskih postupaka. Nakon identifikacije riječi u naslovu, sažetku ili cijelom tekstu (proces identifikacije mogu otežati interpunkcijski znaci, riječi koje se pišu sa pojedinim znakovima, brojčane oznake i sl.) uklanjaju se nepropisni termini, neinformativne riječi sa popisa stop-liste (npr. veznici, prijedlozi i sl). Nakon eliminacije nedeskriptorskih riječi, pronalaze se korijeni riječi i to tehnikama korjenovanja (*engl.* stemming) i lematizacije (*engl.* lematization). U prvom se slučaju uklanjaju sufiksi, a u drugom, umjesto na korijen riječi, riječ se svodi na njen osnovni oblik – lemu. U konačnici se korijen i/ili osnovni oblik riječi zamjenjuje s brojevima deskriptora, dakle, u ovoj se etapi izračunava ukupan broj pojavljivanja pojedine riječi u dokumentu, a usto se, po potrebi, može koristiti i tezaurus za otklanjanje sinonimije (uspoređivanjem riječi iz tezaurusa koje korespondiraju skupinama riječi u tekstu).[3] H. P. Luhn, začetnik ove metode, cijeni kako je frekvencija pojavljivanja pojedine riječi u kauzalnoj vezi s njenom relevantnošću, tj. da se u odnosu na broj pojavljivanja pojedine riječi u dokumentu, u odnosu na IF (*engl*. term frequency) riječi, frekvenciju riječi, može mjeriti značaj pojedine riječi u dokumentu. Na tragu mišljenja da koristan indeksni termin mora u cjelosti zadovoljiti dvije funkcije: reprezentacijsku - indeksni termin mora opisati sadržaj dokumena (funkcija odziva), diskriminacijsku - indeksni termin mora razlikovati pojedini dokument od drugih dokumenata u zbirci (funkcija preciznosti)[4], cilj jednostavnog statističkog automatskog predmetnog označavanja jeste, dakle, uočiti riječi koje imaju visoku diskriminacijsku vrijednost iliti sposobnost odražavanja sadržaja dokumenta. Visoka diskriminacijska vrijednost upućuje i na to da su pojedine riječi sposobne razlikovati i dokumenete međusobno, stoga, nasuprot jednostavnom statističkom automatskom predmetnom označavanju, sofisticiranije automatsko predmetno označavanje teži ustanoviti relativnu važnost termina u dokumentu te relativan značaj istog termina u različitim dokumentima. Otuda se IF riječi vrlo brzo počinju kombinirati sa IDF-om, inverznom frekvencijom dokumenta, te nizom drugih vrijednosti kako bi se povećao kvalitet automatskog predmetnog označavanja[5].

## Unapređenja statističkog modela automatskog predmetnog označavanja

Riječi ili termini koji se pojavljuju u nekoliko dokumenata smatraju se značajnijim za sadržaj tih dokumenata nego riječi koje se frekventno pojavljuju u nekoliko dokumenata (npr. iz, u, a i sl.), a IDF (*engl.* Inverse Document Frequency) ili nerijetko u literaturi imenovani CFW (*engl.* Collection Frequency Weight) metod se koristi ovim fenomenom da izvuče riječi koje mogu najpotpunije opisati sadržaj dokumenta. Nasuprot IF-u, IDF inverzna frekvencija dokumenta je mjera za broj pojavljivanja termina u zbirci dokumenata[6]. S tim u vezi, preduslov za utvrđivanje vrijednosti IDF-a je poznavanje DF (*engl.* Document Frequency) vrijednosti, a kojom se može ustanoviti ukupna količina dokumenata u korpusu koji sadrže pojedinu riječ (paralelno sa IF-om, koji

3 Ibid, str. 107/108

4 Lahtinem, Tino. Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods: academic dissertation. Faculty of Arts at the University of Helsinki, 2000. URL: http://ethesis.helsinki.fi/julkaisut/hum/yleis/vk/lahtinen/automati.pdf (30.10.2010.)

5 Najreprezentativniji primjeri statističkog automatskog predmetnog označavanja su projekti SMART i SYNTOL. Iako se danas koristi isključivo statističkom metodom, u početku je projekt SMART (ali i projekt SYNTOL) pokušao usvojiti sintaksičku i semantičku analizu u automatskom predmetnom označavanju. Svaka rečenica dokumenta parsirala se uz pomoć Hardvard Predictive Analayser-a koji bi analizirao osnovnu strukturu rečenice i parsiranjem izvlačio substrukture, subjekat-predikat te imenica-pridjev veze.

6 Lauc, Tomislava. Pretraživanje obavijesti: pristupi automatskom indeksiranju dokumenata / Modeli znanja i obrada prirodnog jezika. Zagreb: Zavod za informacijske studije, 2003. Str. 178.

se koristi za mjerenje pojedine riječi u dokumentu, koristi se i normalizator duljine, u protivnom, dužina dokumenta može bitno uticati na IDF vrijednost). Produkt IF-a i IDF-a, IF/IDF, se dakle koristi za mjerenje diskriminacijske vrijednosti pojedinog dokumenta u zbirci dokumenata i pritom se svi frekventni termini tumače tj. prepoznaju kao značajni, ali se u konačnici samo riječi sa visokim stupnjem TF/IDF-a selektiraju kao indeksni termini (Salton et al.)[7]. Otuda, za automatsko predmetno označavanje ovoga tipa, izvjesni frekventni pragovi su potrebni npr. minimalni TF (ili *informetrička vrijednost*) kako bi se osiguralo da su odabrani indeksni termini podesni deskriptori sadržaja dokumenta**.**

Sa evolucijom statističkog modela automatskog predmetnog označavanja u vezi treba spomenuti još nekoliko hibridnih pristupa predmetnom označavanju. U tom smislu, pažnje vrijedna je kombinacija morfološke analize i statističkih metoda, spoj koji se za potrebe postizanja generalne deskripcije sadržaja dokumenata opravdava obećavajućim rezultatima jer: morfemi su općenitiji od drugih deskriptora, morfemi su pogodni za grupiranje i razvrstavanje dokumenata, dakle, uz njihovu općenitost, sposobni su diskriminirati dokumente različitih semantičkih klasa[8]. Eksperiment s ovim hibridnim pristupom u vezi potvrdio je prednosti morfološke analize pokazavši primjerice kako statistička analiza ponekada kao indeksne termine predlaže lična imena što nije slučaj s morfološkom analizom.

Statističke siromašne metode se vrlo brzo počinju usavršavati i modelima za pretraživanje: probablističkim modelom, modelom vektorskog prostora te modelom klaster analize. U prva se dva hibridna pristupa, težina odabranih deskriptora dokumenta povezuje i uspoređuje sa korisničkim upitima, a potom se metodom povratne sprege rafiniraju odgovori[9]. Uključivanje povratne sprege u predmetno označavanje dokumenta u probablističkim metodama se smatra važnim jer je povratna sprega najpotpunija informacija o relevatnosti dokumenta u odnosu na postavljeni upit. S tim u vezi, jedan od načina kojim se na korisnički upit može odgovoriti setom relevantnih dokumenata jeste da se korisnički upit uspoređuje prije sa grupama (*engl.* cluster) termina nego sa pojedinačnim terminima. Griffiths i drugi su komparirali različite klaster modele i ustanovili kako za korisnički upit najrelevantnije rezultate daju baze podataka koje su podijeljene na mnoštvo malih klastera. Stoga autori cijene da je optimalni klaster onaj koji se sastoji od dvije jednice koje su međusobno najsličnije (najbliži susjedi)[10]. Značajan je dakle pomak u automatskom predmetnom označavanju načinjen sa pronalaženjem dodatnih izvora informacija (*different source of information*) koji se kombiniraju sa automatskim označavanjem[11]. Pritom, treba primijetiti da normalizacijske postupke koji se realiziraju u statističkom modelu automatskog predmetnog označavanja i pravila na kojima su utemeljeni ne treba poistovjećivati sa dodatnim izvorima informacija, kakav je npr. tezaurus, niti sa konceptnom normalizacijom. Uključivanje tezaurusa kao dodatnog izvora informacija u proces automatskog predmetnog označavanja ima za cilj unaprijediti pretraživanje kroz omogućavanje npr. kombiniranja pretraživanja slobodnog teksta i predmetnica. Konceptna noramalizacija, s druge strane, je postupak dodavanja standardnog termina ili fraze pojedinom sinonimu koji se utvrdi u dokumentu. Očito da ovakva normalizacija zahtijeva prisutnost tezaurusa, a kada se konceptna normalizacija kombinira sa mapiranjem (*engl.* mapping) dobivamo, po mišljenju nekih autora, najkorisniji vid automatskog predmetnog označavanja jer se korisnika oslobađa da promišlja o sinonimima ili načinima na koji će izaziti svoj upit[12].

Statističke metode automatskog predmetnog označavanja očito se koncentriraju skoro pa isključivo na selekcijske tehnike dok je generalizacija u pojedinim hibridnim metodama ovakvog automatskog predmetnog označavanja ograničena samo na morfološka skraćivanja. Stoga se može zaključiti da su ključne osobine jednostavnih statističkih metoda označavanja dokumenata te statističkih hibridnih pristupa sljedeće:

-visoko frekventni i izuzetno nisko frekventni termini smatraju se nepodesnim kandidatima za indeksni termin,

-jednostavna statistička metoda je prilično efikasna u izvlačenju indeksnog termina koji se sastoji od samo jedne riječi,

-jednostavna statistička metoda je relativno jednostavna za mašinsko izvođenje jer je njena osnovna operacija računanje,

7 Prema: Hfriedrichwitschel, Hans. Terminology extraction and automatic indexing: comparison and qualitative evaluation of methods. URL: http://wortschatz.uni-leipzig.de/~fwitschel/papers/TKEIndexing.pdf (12.11.2010.)

8 Ibid

9 Slavić, Aida. Automatsko predmetno označavanje: od računalno potpomognutog predmetnog označavanja do znalačkih sustava / Predmetna obradba: Ishodišta i smjernice. Zagreb: Hrvatsko knjižničarsko društvo, 1998. Str. 108

10 Lancester, F. W. Elliker, Calvin. Connell, Tchera H. Subject analysis / Annual Review of Information Science and Technology (ARIST), Volume 24, 1989. Str. 35

11 Slavić, A. str. 108

12 Automatic indexing today. Kaim Associates. Inc, 2003. URL: http://www.kaim.com/site/literature/auto_white.pdf (10.11.2010.)

-jednostavna statistička metoda je osjetljiva na sadržaj i

-osjetljiva je na dužinu dokumenta,

-statističkom metodom se može mjeriti vrijednost, valjanost riječi kao indeksnog termina, težina riječi se uobičajeno mjeri:

binarno (0 i 1) [prisutna, nije prisutna] ili

rangiranjem (0 <=x=>y) [prisutna je na neki način, nije prisutna]

-vrijednosti se, dakle, mogu ustanoviti na razne načine i kombinirati sa:

1. TF-om ili

2. IDF-om,

3. a u praksi se ponajviše kombinacija TF*IDF pokazala kao efikasna.[13]

## Lingvistički pristup automatskom predmetnom označavanju

Automatsko predmetno označavanje se danas sve češće koristi lingvističkim strategijama, semantičkom te sintaksičkom analizom u utvrđivanju sadržaja dokumenata. Cilj je sintaksičke obrade prirodnog jezika za danu rečenicu putem odgovarajućeg formalnog opisa odrediti njezinu sintaksičku strukturu[14] (dakle, sintaksička analiza se fokusira na gramatičke strukture u rečenici), a semantička razina obrade jezika uključuje analizu značenja rečenice neovisno o njezinom puno širem kontekstu u tekstu ili diskursu. Ova razina obrade jezika usredočena je na pitanje kakav formalizam uporijebiti za prikaz znanja (*knowledge representation*) da bi se smisleno interpretirale rečenice sa više značenja[15] (dakle, fokusira se na značenje riječi). U kontrastu sa statističkim metodama automatskog predmetnog označavanja koje operiraju s frekvencijom termina na različite načine, osnovna strategija sintaksičke analize, tj. na ovom pristupu utemeljenog automatskog predmetnog označavanja, je parsiranje rečenica dokumenta, a uobičajeni generički koraci ovoga metoda su: sintaksička analiza dokumenta, selekcija imeničkih sveza, mjerenje težine termina i u konačnici selekcija indeksnih termina. Rečenice ili naslovi se razbijaju na dijelove i svaka se odvojena komponenta izolira i morfološki analizira, gramatički opisuje – imenica, glagol, pridjev itd. a računar se pritom koristi raznim pravilima da automatski prepozna sekvence riječi. Na ovom se jednostavnom nivou parsiranja komponente rečenice identificiraju kao jedinice koje su potencijalni indeksni termini. Na višem se nivou strategije parsiranja računar programira da raspoznaje i interpunkcijske znakove, tj. da sve komponente koje se nađu između pojedinih interpunkcijskih znakova izdvaja kao fraze, da bi na tom tragu analizirao i povezanost između pojedinih fraza u rečenici[16]. U konačnici se identificirani termini (u prvobitnom koraku) koriste za grupiranje dokumenta na konceptualan način i tako se korisnicima omogućava sintetička reprezentacija znanja sadržanog u dokumentima[17]. Dakle, statistički pristup automatskom predmetnom označavanju se bazira na statistici termina u dokumentu, a lingvistički na sintaksičkoj analizi tekstova koja omogućuje utvrđivanje imeničkih sveza[18] i otuda ova metoda ima veliki potencijal glede poboljšanja kvaliteta automatskog predmetnog označavanja. Ipak, osnovni problem automatskog indeksiranja pomoću imeničnih sveza predstavlja moguća raznolikost prikaza određenog pojma koji u tekstu može biti iskazan različitim svezama. Stoga ga je potrebno prikazati određenom normaliziranom formom koja sjedinjuje sve jezičke varijacije u danom tekstu. Za pojedinačne riječi normalizacija se u najjednostavnijem slučaju može provesti korjenovanjem. Jednu od alternativa u identificiranju sveza riječi koja se često koristi u praksi predstavlja uporaba sintaktički raščlanjenih rečenica teksta na temelju kojih se pronalaze oni parovi riječi koji će se koristiti kao indeksne jedinice. Konceptualno je ovakav pristup negdje između indeksiranja pojedinačnim riječima i indeksiranja svezama onako kako se one pojavljuju u tekstu. Dobiveni parovi riječi podudaraju se s lingvističkim ovisnostima što ih čine glava sveze (*head*) i njezini uobličitelji (*modifiers*). Indeksiranjem parovima riječi na opisani način, zaobilaze se problemi podudarnosti sveza kakve se nalaze u upitu sa svezama u tekstu dokumenta jer dobiveni parovi predstavljaju jedan oblik normalizirane, pojednostavljene veze[19]. U kontrastu sa statističkim

13 Primjer: CRF (*engl.* Conditional Random Fields) novi probablistički model koji se koristi za ekstrakciju ključnih riječi iz dokumenata, a ujedno podržava i model vektorskog prostora. Više o CRF-u: Zhang Chengzhi et al. Automatic keyword extraction from documents using conditional random fields / Journal of Computational Information Systems 4:3(2008) 1169-1180. URL: http://www.JofCI.org (10.11.2010.)

14 Lauc, Tomislava, str. 183

15 Ibid, str. 183

16 Automatic indexing and the linguistics connections. URL:http://www.garfield.library.upenn.edu/essays/v5p031y1981-82.pdf (08.11.2010.)

17 Jaquemin, Christian. Daille, Beatrice. In vitro evaluation of a program for machine-aided indexing. URL: http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?nfpb=true&_&ERICExtSearch_SearchValue_0=EJ656196&ERIExtSearch_SearchType_0=no&accno=EJ656196 (18.11.2010.)

18 Hutchins, John. Some problems and methods of text condensation / UEA Papers in Linguistics. URL: http://www.hutchinsweb.me.uk/UEAPIL-1983.pdf (01.11.2010.)

19 Lauc, Tomislava. Pretraživanje obavijesti: pristupi automatskom indeksiranju dokumenata / Modeli znanja i obrada prirodnog jezika.

metodama automatskog predmetnog označavanja, sintaksičke i semantičke metode automatskog predmetnog označavanja su očito podesne za indeksiranje fraza, to jest sposobne prevazići nejasnoće i dvosmislenosti s kojima se statističko indeksiranje fraza u tom smislu suočava.[20]

### Ekspertni sistemi

Uočeni nedostaci statističkih metoda automatskog predmetnog označavanja, tj. predloženi principi automatskog predmetnog označavanja koji se bore sa kompleksnoću i bogastvom prirodnog jezika, motivirali su izgradnju ekspertnih sistema koji će pomoći procese predmetnog označavanja dokumenata. Metode automatskog predmetnog označavanja koje su kombinirane sa bazom znanja iz pojedinog područja (*engl.* knowledge based indexing system) okupljene su oko ideje da je znanje simbolično pa se otuda na neki način može predstaviti iliti kodirati.[21] Drugim riječima, ove metode automatskog predmetnog označavanja teže razumijevanju sadržaja dokumenta, a ne samo površnoj analizi koja treba da rezultira dodjeljivanjem (odabranih) indeksnih termina (pritom i ovdje razumijevanje varira od jednostavne analize strukture dokumenta do duboke semantičke analize) dokumentima. Ovakvo predstavljanje sadržaja dokumenata zasniva se na „obrascima" (*engl.* case frames) koji izražavaju odnose predmeta unutar pojedinog područja. No, da bi ovo bilo moguće potrebno je predhodno izgraditi bazu znanja iz područja koja će se sastojati od svih pojmova i koncepata unutar područja i od načina na koji se oni mogu povezivati i odnositi jedan na drugi[22]. Zato se u razvoju sistema za automatsko indeksiranje sve više i češće ugrađuju tradicionalni bibliotečki alati poput tezaurusa i klasifikacijskih shema. Iskustva i znanja bibliotekara kao i korištenje klasifikacije i tezaurusa uzimaju se kao važna pomagala i u ekspertnim sistemima jer osiguravaju njihovu kvalitetnu primjenu[23]. Automatsko predmetno označavanje koje se zasniva na obrascima, tezaurusu, omogućava visoku preciznost koju automatsko predmetno označavanje koje ne referira na eksterni izvor znanja ne može garantirati. Dakle, bilo koje softversko rješenje ovoga tipa posjeduje izvjesno lingvističko znanje (u protivnom ne bi moglo uspostavljati veze između pojedinih predmeta) koje se uobičajeno ugrađuje u nekoliko modula:

1. baze znanja koje obuhvataju najmanje dvije vrste znanja: znanje eksperta za određeni domen (nauku) i znanje o strategijama i pravilima za izbor termina i formulisanje upita tokom PI;

2. mehanizam zaključivanja ili rezonovanja koji omogućava sistemu da iz baze znanja pronađe relevantno znanje i primijeni ga na rješavanje problema do zaključka o rješenju ili odgovarajućeg savjeta;

3. baza činjenica koja opisuje tekući status problema koji se rješava, a koje uglavnom daje korisnik u interakciji sa sistemom;

4. korisnički interfejs koji omogućava komunikaciju/interakciju sistema i korisnika kojim je često obuhvaćen i modul za obrazlaganje preporuka ili zaključaka sistema[24].

Najveći problem sa automatskim predmetnim označavanjem koje je povezano s bazom znanja uočava se u njegovoj nemogućnosti da funkcioniše preko granice svog definiranog područja to jest u nemogućnosti da identificira dokumente koji ne pripadaju definiranoj domeni[25].

---

Zagreb: Zavod za informacijske studije, 2003. Str. 187/188

20 Primjeri: SOAF (*engl.* Semantic Indexing based on collaborative tagging), sistem za semantičko automatsko označavanje objekata učenja (*engl*. e-learning objects) u repozitorijima ili NASA-in program MAI. Kada se MAI programu učitaju naslovi ili sažeci članaka, program obilježava potencijalne nazive tražeći interpunkciju kojom se završava misao, kao što su tačka, zarez i tačka sa zarezom. Zatim, upotrebom posebno izrađenog rječnika za provjeru, pojedine se riječi koje čine potencijalne nazive dodjeluju sintaktičkim kategorijama. Nizovi se riječi potom ispituju na osnovi dopuštenih sintaktičkih formata. Proces nije u potpunosti automatiziran jer se rečenični nizovi koji prođu ispitivanje podvrgavaju ljudskom pregledu. Prema: Svenonius, Elaine. Intelektualne osnove organizacije znanja. Lokve: Naklada Benja, 2005. Str. 141. Više o SOAF-u u: Cernea, Doina Ana et al. SOAF: semantic indexing based on collaborative tagging. Interdisciplinary Journal of e-learning and learning objects. Volume 4, 2008. URL: http://ijklo.org/Volume4/IJELLOv4p137-149Cernea.pdf (05.11.2010.)

21 Bradshaw, Edward Charles. The use of automated document structuring and classification methods in the legal domain: academic thesis. 1995. https://circle.ubc.ca/handle/2429/3936 (05.11.2010.)

22 Slavić, A. str. 109

23 Dizdar, S. str. 152

24 Matić, Milena. Istraživanje inteligentnih sistema za pretraživanje informacija. URL:http://scindeks-clanci.nb.rs/data/pdf/1450-8915/2000/1450-89150001067M.pdf (24.11.2010.)

25 Bradshaw, Edward Charles. The use of automated document structuring and classification methods in the legal domain: academic thesis. 1995. URL: https://circle.ubc.ca/handle/2429/3936 (05.11.2010.) Primjeri: MedIndex sistem koji se koristi za predmetno označavanje medicinske literature, kao pomoć stručnjacima u selekciji indeksnih termina. ILIAD program za mašinski potpomognuto predmetno označavanje, uključuje: modul za automatsku lingvističku analizu dokumenata, modul za proširenje novih terminima te modul za rudarenje podataka. ILIAD lingvistički inžinjerski program može predstaviti područje znanja i u formi mape. Kao reprezentativan primjer mašinski pomognutog predmetnog označavanja ovoga tipa svakako treba izdvojiti MAI softversko rješenje koje je razvijeno za potrebe indeksiranja EPOQUE baze podataka Evropskog parlamenta. MAI ekspertni sistem predlaže indeksne termine iz baze znanja koja se sastoji od pravila za prepoznavanje podesnog indeksnog termina, a koja su iscrpljena iz strukture EUROVOC višejezičnog hijerarhijskog tezaurusa koji sadrži 5359 deskriptora. Testiranje MAI programa pokazalo je kako isti, bez ikakve ljudske intervencije, može proizvoditi dosljedne i podesne indeksne termine, odnosno, da ovaj ekspertni sistem može omogućiti brže, konzistentnije, ekonomičnije i čak kvalitetnije indeksiranje

## Zaključna razmatranja

Počevši od jednostavnih KWIC i KWOK indeksa, preko TF/IDF metoda i drugih statističkih hibridnih pristupa, ponajprije je uključivanje linvističkog znanja u proces automatskog predmetnog označavanja počelo rezultirati relativno kvalitetnim indeksnim terminima, a kao posebno sofisticiranu metodu automatskog predmetnog označavanja treba prepoznati kombinaciju automatskog predmetnog označavanja i baze znanja pojedinog područja. Kvalitetno automatsko predmetno označavanje (a pogotovo ono povezano s bazom znanja) zahtijeva naprednu tehnološku podršku ali je ista u odnosu na način i napor kojima će se razumijevati jezički procesi, vršiti obrada prirodnog jezika, od sekundardnog značaja. Otuda automatsko predmetno označavanje može profitirati tek združivanjem znanja informatičara, lingvista, bibliotekara i drugih strana. Tek se takvom sinergijom mogu unaprijediti metode automatskog predmetnog označavanja, osigurati prihvatljivi automatski generirani indeksni termini, to jest podržati procesi konsolidacije ogromne i rastuće količine elektronskih izvora informacija.

## Literatura

**Aluri,** R. Kemp, D.A. Boll, J.J. Subject analysis in online catalogs. Englewood: Libraries Unlimited Inc, 1991.

**Andersson**, Linda. Performance of two statistical indexing methods, with and without compound-word analysis. URL:

http://www.ifs.tuwien.ac.at/~andersson/LindaAndersson_Compound.pdf (10.11.2010.)

**Automatic** indexing and the linguistics connections.

URL:http://www.garfield.library.upenn.edu/essays/v5p031y1981-82.pdf (08.11.2010.)

**Automatic** indexing today. Kaim Associates. Inc, 2003.

URL: http://www.kaim.com/site/literature/auto_white.pdf (10.11.2010.)

**Bradshaw,** Edward Charles. The use of automated document structuring and classification methods in the legal domain: academic thesis. 1995.

URL: https://circle.ubc.ca/handle/2429/3936 (05.11.2010.)

EPOQUE baze podataka. Više o tome: Bradshaw, Edward Charles. The use of automated document structuring and classification methods in the legal domain: academic thesis. 1995.
URL: https://circle.ubc.ca/handle/2429/3936 (05.11.2010.)

**Cernea,** Doina Ana et al. SOAF: semantic indexing based on collaborative tagging. Interdisciplinary Journal of e-learning and learning objects. Volume 4, 2008.

URL: http://ijklo.org/Volume4/IJELLOv4p137-149Cernea.pdf (05.11.2010.)

**Dizdar**, Senada. Sistemi za označavanje i pretraživanje informacija u bibliotečko-informacionim sistemima i servisima: doktorska disartacija. Sarajevo, 2007.

**Fugmann**, Robert. Subject analysis and indexing. Frankfurt; Main: Indeks Verlag, 1993.

**Hfriedrichwitschel**, Hans. Terminology extraction and automatic indexing: comparison and qualitative evaluation of methods.

URL: http://wortschatz.uni-leipzig.de/~fwitschel/papers/TKEIndexing.pdf (12.11.2010.)

**Hutchins**, John. Some problems and methods of text condensation / UEA Papers in linguistics.

URL: http://www.hutchinsweb.me.uk/UEAPIL-1983.pdf (01.11.2010.)

**Jaquemin**, Christian. Daille, Beatrice. In vitro evaluation of a program for machine-aided indexing. URL:http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=EJ656196&ERICExtSearch_SearchType_0=no&accno=EJ656196 (18.11.2010.)

**Lahtinem,** Tino. Automatic indexing: an approach usingusing an index term corpus and combining linguistic and statistical methods: academic dissertation. Faculty of Arts at the University of Helsinki, 2000.

URL: http://ethesis.helsinki.fi/julkaisut/hum/yleis/vk/lahtinen/automati.pdf (30.10.2010.)

**Lancester**, F. W. Elliker, Calvin. Connell, Tchera H. Subject analysis / Annual Review of Information Science and Technology (ARIST), Volume 24, 1989.

**Lauc,** Tomislava. Pretraživanje obavijesti: pristupi automatskom indeksiranju dokumenata / Modeli znanja i obrada prirodnog jezika. Zagreb: Zavod za informacijske studije, 2003.

**Matić**, Milena. Istraživanje inteligentnih sistema za pretraživanje informacija.

URL: http://scindeks-clanci.nb.rs/data/pdf/1450-8915/2000/1450-89150001067M.pdf (24.11.2010.)

**Rowley**, J. The controlled versus natural indexing languages debate revisited revisited: a perspecti-

ve on information retrieval practice and research / Journal of information science. 20, 2(1994)

**Shields**, Ginger. What are the main differences between human indexing and automatic indexing. URL: http://www.shieldsnetwork.com/LI842_Shields_Automatic_Indexing.pdf (30.10.2010.)

**Slavić,** Aida. Automatsko predmetno označavanje: od računalno potpomognutog predmetnog označavanja do znalačkih sustava / Predmetna obradba: Ishodišta i smjernice. Zagreb: Hrvatsko knjižničarsko društvo, 1998.

**Svenonius**, Elaine. Intelektualne osnove organizacije znanja. Lokve: Naklada Benja, 2005.

**Taylor**, Arlene G. The organization of information. Englewood, Colorado: Libraries Unlimited Inc, 1999.

**Wellisch**, Hans. H. Glossary of terminology in abstracting, classification, indexing, and thesaurus constructions. The American society of Indexers, 2000.

**Zhang** Chengzhi et al. Automatic keyword extraction from documents using conditional random fields / Journal of Computational Information Systems 4:3(2008) 1169-1180. URL: http://www.JofCI.or (28.10.2010.)

***

***Mr. Lejla Hajdarpašić*** *(Sarajevo, 1983.) je magistrica bibliotekarstva. Na Odsjeku za komparativnu književnost i bibliotekarstvo Filozofskog fakulteta u Sarajevu je birana 2009. u zvanje asistenta za oblast bibliotekarstvo. Na istom fakultetu je diplomirala (2006.) i stekla akademsku titulu i stručno zvanje magistar (2011.). Nosilac je Zlatne značke Univerziteta u Sarajevu. Pohađa poslijediplomski znanstveni doktorski studij informacijskih i komunikacijskih znanosti (smjer: Bibliotekarstvo) na Filozofskom fakultetu u Zagrebu. Objavila je zbirku poezije i niz priloga u različitim časopisima (Odjek, Most i dr.) Sa referatima je učestvovala na međunarodnim i domaćim naučnim skupovima (SEEDI International Conference, BAM i dr.). Članica je Sekcije za fakultetske i specijalne biblioteke unutar Asocijacije informacijskih stručnjaka – bibliotekara, arhivista i muzeologa (BAM). Učestvuje u realizaciji TEMPUS projekta pod nazivom Developing information literacy for lifelong learning and knowledge economy in Western Balkan countries. Područje njena interesovanja su: informacijska pismenost, digitalne biblioteke i institut obaveznog primjerka.*

***Lejla Hajdarpašić,*** *MA of librarianship (Sarajevo, 1983). In 2009 she got a professional rank of assistant of librarianship domain at the Department for Comparative Literatures and Librarianship at the Faculty of Philosophy. She graduated at the same faculty (2006) and has got the academic title and professional rank of MS (2011). She was awarded by Golden badge of University of Sarajevo. She attends postgraduate study from information and communication sciences (Department of Librarianship) at the Faculty of Philosophy in Zagreb. She made public poetry collection and papers in different journals (Odjek, Most etc.) She participated with her papers in numerous international and internal symposiums (SEEDI International Conference, BAM etc.). She is a member of the Section for Faculty and Special Libraries within Association of Information Professionals – Librarians, Archivists and Museologists. She participated in realisation of TEMPUS project with the title Developing information literacy for lifelong learning and knowledge economy in Western Balkan countries. Information literacy, digital libraries and legal deposit are the spheres of her profesional orientations.*