Aida Slavić

AUTOMATSKO PREDMETNO OZNAČIVANJE : OD RAČUNALNO POTPOMOGNUTOG PREDMETNOG OZNAČIVANJA DO ZNALAČKIH SUSTAVA

Automatic Subject Indexing: >From Computer Subject Indexing to Knowledge-Based Systems

ABSTRACT

This paper deals with developments of automatic indexing. Document indexing is a time consuming and labour-intensive part of document processing and remains the most critical part of the document retrieval process. This paper is an attempt to give a short overview of the subject by answering two questions: what should librarians already know about automatic indexing and what could they expect from it in the fixture? The development of information technology has a major impact on information professionals and determines their future role. The advent of the Intemet and the rapid progress in the digitizing of document collections puts even greater stress on the importance of fast and relevant retrievals in an ever increasing mass of electronic documents. Every day we face a growing number of full text databases which need to be properly indexed in order to be accessed efficiently. Automatic indexing as a tool in document indexing and retrieval has been successfully developed up to the point whereby it should now be taken in to considera6on whenever a full text database is being used. Modern automatic indexing uses techniques from the field of computational linguistics and artificial. intelligence and is moving toward knowledge-based information retrieval. The future role of the information professional will be to find a better way to control, use and improve these systems.

UVOD

Računalni programi koji su zamišljeni da bez čovjekove pomoći uspješno odrede predmet o kojem se u dokumentu govori, razvijaju se već četrdeset godina.(1) Manje ili više uspješni, najčešće su plijenili pažnju informatičara zainteresiranih za razvoj automatske obradbe teksta i umjetne inteligencije. O sustavima automatskog predmetnog označivanja zna se posredstvom projekata i eksperimenata koji su se u prvom redu odnosili na fenomen pretraživanja informacija, a potom i iznalaženje načina da se poboljšaju ručne tehnike označivanja i pretraživanja (potpun pregled ranijih projekata dat je u knjizi *Information Retrieval Experiment, 1981).*

Ručno dodjeljivanje pojmova za identifikaciju sadržaja podrazumijeva da se korisnik u postupku pretraživanja služi istim pojmovima i da pod njima podrazumijeva istovjetne sadržaje kao i predmetni stručnjak koji je dokument obradio. Kod tog načina predmetnog označivanja korisnik često treba pomoć informacijskog stručnjaka posrednika koji bi njegov informacijski problem mogao prevesti na jezik sustava. U sustavima za pretraživanje slobodnog teksta *(free text searching),* međutim, suvremene tehnike pretraživanja omogućile su korisniku da posve slobodno odabere pojam po kojem će pretraživati. Korisnik se može koristiti riječima prirodnog jezika za koje očekuje da se nalaze u tekstu dokumenta, a koje nisu

kontrolirane nikakvim rječnicima ili normativnim popisima.(2) Rafiniranjem spomenutih osnovnih tehnika pretraživanja slobodnog teksta započeo je uspješan razvoj suvremenih tehnika pretraživanja i automatskog označivanja.

Već krajem pedesetih godina počeli su se raditi eksperimenti s ciljem da se poboljša relevantnost pretraživanja dotjerivanjem obrnutih kazala, kontrolom rječnika, razrješavanjem problema sinonimije i homonimije itd. Jedan od poznatijih eksperimenta iz područja pretraživanja informacija jest *SMART* u okviru kojeg je već početkom 60-ih godina započelo ono što danas podrazumijevamo pod nazivom automatsko predmetno označivanje. Računala su se koristila za uslikavanje teksta dokumenata *(scanning)* i dodjeljivanje pojmova označitelja na osnovi učestalosti njihovog pojavljivanja u tekstu *(Salton, I981).* Što se tiče eksperimenata i projekata iz domene automatskog označivanja devedesetih godina, tu svakako treba spomenuti bar neke: europske SIMPR *(Schuegraf, 1997),* IOTA, SPIRIT, NOEMIC *(Mustafa-Elhadi, 1997),* američke *IRIS (Richardson, 1995),* MedlndEx *(Lancaster; Sandore, 1997, Croft, 1989),* CLARIT i indijski PROMETHEUS *(Prasad, 1996).*

Od samog početka korištenja računala u području označivanja i pretraživanja dokumenata ravnopravno se koriste sustavi u kojima posao stvarnog označivanja dokumenata radi isključivo čovjek, potom oni u kojima se računala koriste za kontrolu rječnika, a stvarnu obradbu obavljaju ljudi ili se računala koriste za korekciju ljudskih pogrešaka. Na kraju tu su i manje ili više uspješni sustavi u kojima se nastojalo potpuno isključiti čovjeka u postupku označivanja sadržaja.

Načini kojima se ovi sustavi opisuju u literaturi, kao i terminologija koja se pritom kotisti vrlo su raznoliki i ovise o provenijenciji samog autora. Većina autora bave se u prvom redu automatskom obradbom teksta. Kao znalci na području informacijske tehnologije, označivanju i pretraživanju dokumenata pristupaju nešto drukčije od informacijskih stručnjaka s tradicionalnom knjižničarskom naobrazbom. Da bi se učinio značajan i kvalitetan pomak na području označivanja i pretraživanja, potrebno je zbližiti ova dva različita pristupa. Najmanje što knjižničari, kao informacijski stručnjaci, u tom smislu mogu učiniti jest proučavanje onoga što je na spomenutom području već učinjeno, na čemu se trenutno radi i koji su dosezi pojedinih rješenja. Namjera ovog izlaganja jest kratak pregled razvoja područja automatskog predmetnog označivanja na razini općih informacija i natuknica s ciljem da se potakne zanimanje za odnosnu stručnu literaturu. Kao prilog već navedenim razlozima treba spomenuti i sljedeće riječi G. Saltona:

> "... čini se daje retoričko pitanje g. Kerena "Što ste u posljednje vrijeme vi istraživači učinili za nas praktičare?" promašilo svrhu. Ne postoje nigdje prečaci između istraživanja i primjene. To važi jednako za naše kao i za druga područja intelektualnih napora. Potrebno je **proučavati** literaturu, potrebno je imati dovoljno "know-how" sposobnosti da bi se problem mogao razlučiti i smjestiti u kontekst. S vremenom, djelići će se spojiti u cjelinu, a promatrač će moći uočiti pojedinačne detalje umjesto da se oslanja na površne utiske i uopćavanja. U našem, kao i u svakom drugom području, potrebno je poznavati polje djelovanja kako bi se moglo doprinijeti njegovom razvoju. *"(Salton, 1997, 134)*

NOVA INFORMACIJSKO TEHNOLOŠKA OKOLINA I NOVI PROBLEMI

Za pretraživanje bibliografskih baza podataka postoje već priređeni alati za označivanje i pretraživanje (predmetnice, klasifikacijske oznake, ključne riječi). Na početku razvoja računalne tehnologije računala su se i u informacijskim centrima koristila isključivo za mehaničke, dugotrajne uredske poslove. Obradbu dokumenata, usporedbu i kategorizaciju informacijskih zahtjeva obavljali su isključivo ljudi. Kvaliteta ljudskog rada na ovom području vrlo se rijetko dovodila u pitanje ili znanstveno istraživala. Eksperimenti koji su mjerili efikasnost računalnih programa u obavljanju nekih od složenijih zadataka u području pretraživanja (poput eksperimenta *Cranfield 2)* uspoređivali su rad računala s radom stručnjaka iz područja predmetizacije. Rezultati su pokazali ograničenja i jednog i drugog načina rada i ukazali na kompleksnost problema pretraživanja informacija *(Sparck .lones, 1981). (3)*

Porast broja baza punog teksta i mogućnosti pretraživanja slobodnog teksta *(free text searching)* dala je poticaj razvoju računalnih programa koji su zamišljeni ne samo kako bi obrađivali dokumente jednako dobro kao čovjek, već u cilju da ga u tome nadmaše. Ono što uistinu otvara prostor računalima u području sadržajne obradbe, koja se oduvijek smatrala elitnom, neprikosnoveno intelektualnom domenom informacijskih stručnjaka, jest činjenica da čovjek sam više ne može sadržajno analizirati i označiti nezaustavljivu masu elektroničkih dokumenata koja se velikom brzinom pojavljuje i nestaje u globalnom mrežnom okruženju.

Valja reći da, paralelno s velikom produkcijom dokumenata u elektroničkom obliku, raste i broj tiskanih dokumenata. (4) S druge pak strane, zbog velike potrebe da uspješno upravlja tom rastućom gomilom papira, informacijski servisi pretvaraju papirnate dokumente u, za manipulaciju praktičniji, elektronički oblik. Uslikavanjem *(scanning)* i pretvaranjem slika u tekst pomoću programa za slikovno prepoznavanje slova (OCR *optical character recognition),* primarno "papirnati" tekst postaje elektronički i pridružuje se rastućoj, svima dostupnoj zbirci punog teksta koju treba moći brzo i točno pretraživati.

Zašto trebamo automatsko predmetno označivanje : stari problem u novom ruhu?

Slijedom ovog problema, a gotovo istodobno s nastankom Interneta, počeli su se, kao pomagala za pronalaženje relevantnih dokumenata, na Internetu razvijati programi za pretraživanje *World Wide Web-a (search agents, search engines).* Ovi programi iz teksta dokumenata dostupnih na WWW grade golema strojno generirana kazala (usp. Mulvany, 1996).

Jednostavniji od ovih programa omogućavaju uobičajene tehnike: pretraživanje po svakoj riječi, po čitavom izrazu, kombiniranje više pojmova pomoću Booleovih logičkih operatora (AND, OR, NOT), krnjenje izraza s desna i tome slično.

Složeniji od programa za pretraživanje uposlit će druge jednostavne programe i onda će usporedbom pogodaka dati još bolje rezultate *(meta search engines).* Dodavanjem mogućnosti rafiniranja pretraživanja: redanja po relevantnosti, sužavanja izbora prema jeziku, godini, formatu, lokaciji pa sve do dizajniranja tzv. inteligentnih programa za pretraživanje *(intelligent agents)* s mogućnošću definiranja profila korisnika, dobivaju se još relevantniji odgovori. Nadalje, programima se dodaje sposobnost učenja korištenjem mehanizma povratne sprege za korekciju i uspostavljanje pravilnih korelacija te kao vrhunac - mogućnost da "surađuju" s drugim inteligentnim programima (usp. *Locke, 1997).*

Kao što vidimo, spomenuti programi iz dana u dan postaju sve sofisticiraniji. Još uvijek se, međutim, bore s osnovnim problemom. Veliki broj dokumenata koji se u pretraživanju odaziva na pojedini pojam potpuno je nerelevantan za traženi upit.

Problem je jednostavno u tome što dokumenti na Internetu nisu sadržajno analizirani, označeni niti organizirani bilo na koji način. (5) Pojmovi koje program za pretraživanje traži mogu se nalaziti bilo u kojem kontekstu i bilo kakvom tipu dokumenta. Jasno je da je efikasnost sustava za pronalaženje informacija u potpunosti osigurana jedino onda ako je dokument smišljeno i precizno označen. Označivanje i pretraživanje dokumenata nerazdvojivi su dijelovi jednog te istog procesa čiji je cilj pronalaženje informacija. Nedovoljno relevantan odaziv u slučaju Interneta rezultat je traženja dokumenta čiji sadržaj nije prethodno bio podvrgnut nikakvoj analizi i označivanju s nedvosmislenim i jasnim značenjem *(usp. Buckland et al., 1995).*

Postupak predmetnog označivanja treba osigurati predstavljanje sadržaja i značenja dokumenta. Neovisno o tome je li ovo označivanje obavlja stroj ili čovjek, ono treba osigurati da korisnik svoj problem (zamisao, koncept) formulira u obliku upita, a da sustav može usporedbom tog upita s opisom sadržaja dokumenata u svojoj bazi, dati odgovore. Bez obzira kako je upit (informacijski problem) fonnuliran, najvažniji dio postupka je njegova usporedba s onim popisom pojmova koji predstavljaju sadržaj dokumenata te lociranje relevantnih dokumenata u nekoj zbirci.

Ne postoji, kao što znamo, savršen jezik za označivanje dokumenata, no znamo da onaj koji danas dolazi u obzir mora zadovoljavati neke preduvjete koji se do izravnog računalnog pretraživanja *(online)* nisu postavljali kao imperativ. Danas, kad korisnici dokumentima pristupaju izravno, bez posrednika, u obzir dolazi jedino pretraživanje na prirodnom jeziku. Poznajući sva ograničenja predmetnog označivanja i stotinama godina staru borbu knjižničara da "ukrote prirodni jezik", definiraju popise (rječnik) i propise (sintaksu) jednog umjetnog jezika koji se služi riječima prirodnog – zanimljivo je vidjeti na koji način su posljednjih četrdeset godina to isto, potpuno neovisno, pokušavali učiniti stručnjaci za automatsku obradbu teksta.

RAZVOJ RAČUNALNIH SUSTAVA U PODRUČJU OZNAČIVANJA I PRETRAŽIVANJA

Okupljanje predmeta

I kod predmetnog označivanja i kod klasifikacije dokumenti se okupljaju pronalaženjem i dovođenjem u vezu pojmova koji se nalaze u njihovom sadržaju. Za okupljanje pojmova unutar jednog područja i za odabir riječi kojima ćemo te pojmove izraziti znamo da se može, primjerice, koristiti pojmovnik u obliku tezaurusa. Tezaurus teži opisati i prikazati odnose koji postoje među pojmovima: rod-vrsta, cjelina-dio, discipline-potpodručja, upotreba i djelovanje, i velik raspon manje specifičnih

i manje stalnih asocijativnih veza. Upravo svi ti odnosi omogućavaju šire okupljanje vrsta u rodove, dijelova u cjeline, potpolja u discipline te okupljanje zajedno sudionika iz različitih odnosa i veza. No okupljanje na bazi tezaurusa ili klasifikacije podrazumijeva izuzetan ljudski intelektualni napor, pa je mogućnost da se to obavlja strojno oduvijek bila izazov za informacijske stručnjake.

J. D. Anderson govori o jednom načinu okupljanja na osnovi pojmova u tekstu kao metodi koja se uspješno obavlja i kontrolira strojno te kao primjer daje neka od ispitanih programskih rješenja. To je okupljanje na osnovi, kako ih on naziva, predmetnih osobina *(entity attributs)* može biti vrlo korisno i upotrebljivo u sustavima za pretraživanje *(Anderson, 1989, 74)*. Predmeti-dokumenti, tekstovi ili objekti mogu se grupirati na osnovi pojmova koje sadrže i koji su im zajednički. Takvi pojmovi mogu biti ključne riječi iz naslova, sažetaka ili cjelovitog teksta; dodijeljenih deskriptora, bibliografskih referenci; imena autora, časopisa, nakladnika, jezika ili bilo koje druge imenovane osobine. Jednako tako ti se pojmovi mogu grupirati na osnovi njihovog zajedničkog pojavljivanja ili na osnovi njihovog odnosa prema drugim predmetima. Grupirani pojmovi mogu se potom koristiti za lociranje dodatnih, potencijalno relevantnih predmeta.

Primjer:

1978. g. Tamas Doskocz je predstavio prototip okupljanja pojmova nazvan "asocijativni interaktivni rječnik" *(associative interactive dictionary).* Taj sustav je redao pojmove prema relativnoj učestalosti njihova pojavljivanja u određenoj skupini dokumenata uspoređenoj s čitavom zbirkom. U TOXLINE bazi podataka učinjeno je preliminarno pretraživanje pretraživanjem pojmova *Prenatal i Toxicity.* Asocijativni interaktivni rječnik je potom izlistao sljedeće riječi poredane prema tome koliko se češće pojavljuju u potraživanom segmentu nego u čitavoj zbirci dokumenata:

prenatal, postnatal, gestational, fetus, gestations, teratogenicity.....itd.

Većina ovih pojmova su u jasnim odnosima - sadržajnim i statističkim i jednom ustanovljeni oni bi mogli služiti za sužavanje ili proširivanje određene grupe pojmova.

(Anderson, 1989, 73)

KLASTER ANALIZA KAO NAČIN OKUPLJANJA PREDMETA

Klaster analiza omogućava odabir i grupiranje dokumenata koji dijele neke zajedničke karakteristike. Bilo koji od svojstava pridodanih dokumentu može služiti kao osnova za grupiranje.

Primjer:

Gerard Salton je koristio ovo načelo u sustavu "SMART" 1982. Program je označivao i razlikovao sljedeće odnose među dokumentima koji dijele neke zajedničke osobine:

- "**string**", tipična klasifikacijska struktura gdje se svaki dokument promatra u odnosu na samo jedan drugi dokument (A je u odnosu samo s B, B u odnosu s C itd.)
- "**star**", zvjezdasta struktura: gdje su svi dokumenti stavljeni u odnosu na jedan središnji dokument
- **"clique"** gdje su svi dokumenti stavljeni u odnose jedan naspram drugog
- **"clump"** gdje dokumenti nisu u odnosu na pojedinačne dokumente već na skupinu dokumenata

Umjesto pojmova, deskriptora, kao osnova za grupiranje u ovom sustavu tome može poslužiti popis referenci na kraju dokumenta npr.:

- bibliografsko udruživanje *(bibliographic coupling)* - kad su dokumenti grupirani prema na osnovi zajedničkih referenci;
- su-citiranje *(co-citation)* - kad se dokumenti okupljaju prema njihovom zajedničkom pojavljivanju u bibliografskim referencama. *(Anderson, 1989, 74)*

Strojno potpomognuto predmetno označivanje kao priprema za ručno predmetno označivanje

Nekoliko proizvodača baza podataka koriste se programima za strojno potpomognuto predmetno označivanje. Ovi programski paketi obično analiziraju tekst, najčešće sažetak dokumenta i predlažu pojam označitelj. Predmetni stručnjak potom pregledava dokument i ponuđene pojmove te se odlučuje za one koje drži ispravnima, a po potrebi dodaje i nove. Poznatiji od onih koji se koriste ovom metodom su primjerice American Petroleum Institute, Defence Technical Information Center i

također američki National Aeronautics and Space Administration i Access Inovation Inc (Brown, 1997).

Strojno potpomognuto predmetno označivanje kao dopuna i kontrola ručnom predmetnom označivanju

U ovim se sustavima pojmovi označitelji ili predmetnice koje je čovjek dodijelio kontroliraju i preciziraju pomoću računalnih programa. Time se bitno olakšava čovjekov posao ali i osigurava dosljednost i preciznost u označivanju. To je primjerice metoda koja se koristi u MedIndEx bazi podataka u National Medicine Library (USA)*(Brown, 1997).*

Da bi to bilo moguće, potrebno je izgraditi bazu znanja i precizan sustav pravila na kojima se zasniva označivanje. Kad postoji baza znanja i sustava pravila, moguće je programirati znalačke sustave koji mogu nadzirati i kontrolirati postupke označivanja i ukazivati na propuste. Znalački sustav (6) koji, primjerice, koristi *International Nuclear Information System* kontrolira predmetne oznake na osnovi usporedbe klasifikacijske oznake dokumenta i predmetnica te određuje koji dokumenti nisu pravilno predmetizirani i označuje ih za ponovljeni postupak sadržajne obradbe i ljudsku kontrolu.

Automatsko predmetno označivanje - osnovna pitanja

Obično, tradicionalno predmetno označivanje (indeksiranje) (7) je iskazivanje osobina predmeta ili događaja o kojima se govori u nekom dokumentu, korištenjem jezgrovitih odabranih pojmova označitelja (deskriptora), predmetnica ili šire - indikatora. Te označitelje sadržaja G. Salton opisuje kao surogate dokumenta, a predmetno označivanje postupkom stvaranja surogata *(process of constructing document surrogates) (Salton, 1989, 276).* Pojmovi se "doznačuju" *(assignment)* dokumentu i sam dokument ih ne mora sadržavati, a predmetni stručnjak se za njih odlučuje na osnovi svog znanja, iskustva i svoje subjektivne procjene.

U novije vrijeme, ravnopravno spomenutoj tradicionalnoj metodi predmetnog označivanja, a u direktnoj vezi s računalnom obradbom teksta i bazama punog teksta, rabe se i druge metode. Kao osnova za predmetno označivanje koristi se originalni dokument, a analiza teksta dokumenta izvodi se i kontrolira automatski pomoću specijalnih računalnih programa. Ta vrsta predmetnog označivanja poznata je pod nazivom automatsko ili automatizirano indeksiranje *(automatic indexing, automated indexing).* Automatsko predmetno označivanje počiva na "uzimanju" pojmova označitelja iz teksta dokumenta *(extraction) i* podrazumijeva niz logičkih postupaka koji se mogu lako izvoditi i kontrolirati strojno. Načelno, u postupku predmetnog označivanja dokumentima se dodjeljuju identifikatori sadržaja koji u postupku pretraživanja upućuju korisnika na određene jedinice. Dobro odabrani identifikatori sadržaja mogu se koristiti i za međusobno povezivanje dokumenata - obzirom da jedince kod kojih se preklapa veliki broj identifikatora očigledno pokrivaju isto ili srodno područje.

Osnovna ideja na kojoj počiva automatsko predmetno označivanje jest ta da učestalost pojavljivanja neke riječi u tekstu koincidira s vrijednošću koju ona ima kao pojam označitelj.

"...pisac obično ponavlja određene riječi dok varira ili razvija svoje tvrdnje. Taj način isticanja uzet je kao identifikator značenja."

H. P. Luhn

(prema Automatic indexing, I 997, 1)

U tekstu dokumenta nalazi se velik broj riječi poput članova, veznika, prijedloga i priloga čija je učestalost pojavljivanja konstantna u svim dokumentima u nekoj zbirci. Za te tzv. funkcijske riječi *(function words),* kako ih naziva Salton, karakteristična je velika frekvencija pojavljivanja u običnom tekstu. Za one druge "nefunkcijske riječi " koje prezentiraju sadržaj dokumenta vrlo je neujednačena frekvencija pojavljivanja i ona se potpuno razlikuje od teksta do teksta u nekoj zbirci. Frekvencija pojavljivanja "nefunkcijske" riječi u nekom tekstu može služiti kao indikator važnosti te odredene riječi kao identifikatora nekog sadržaja.

Svaki sustav za automatsko predmetno označivanje logično počinje razlučivanjem riječi koje u tekstu ne nose značenje (spomenuti članovi, prijedlozi, prilozi, veznici itd.) od potencijalnih pojmova označitelja. To se postiže stvaranjem popisa svih pojmova koji se isključuju *(stoplist).* Sljedeći opći postupak je uklanjanje nastavaka (sufiksa) iz popisa pojmova kako bi se dobili osnovni oblici riječi odnosno korijeni, a potom se izračunava učestalost njihove pojave u tekstu.

Učestalost pojavljivanja pojmova

Svako pretraživanje mora ispunjavati dva uvjeta: odaziv i preciznost. Učestalost pojavljivanja nekog pojma garancija je da će se u postupku pretraživanja na taj pojam "odazvati" odgovarajući broj dokumenata. Ako, primjerice, uzmemo riječ "ptica" - koja se umjereno često pojavljuje u nekoj zbirci dokumenata, učestalost je indikacija da određeni dokumenti u toj zbirci govore o pticama i ova riječ će svakako ispuniti funkciju pojma za pretraživanje u danoj skupini dokumenata.

Ako želimo postići veću preciznost u pretraživanju, situacija je nešto složenija. Uzmimo određenu skupinu dokumenata koji govore o istom problemu. Obično nas zanima koji od tih dokumenata govore o spomenutom problemu više od ostalih. To ćemo moći saznati samo ukoliko se broj pojavljivanja neke indikativne riječi znatno razlikuje od dokumenta do dokumenta u toj

skupini. Drugim riječima riječ "ptica" nije precizan označitelj sadržaja u nekoj ornitološkoj zbirci dokumenata iako je njezina učestalost vrlo velika. Preciznost u pretraživanju (za razliku od odaziva) puno se bolje postiže korištenjem riječi koja se u određenoj skupini dokumenata pojavljuju rijetko. Ti će pojmovi dati informaciju o razlici među tekstovima koji se odnose na isti problem.

Ove dvije situacije mogu se kombinirati u jednom jedinom frekvencijskom modelu koji se postavi tako da je najbolji pojam za označivanje (tj. onaj koji ispunjava obje funkcije pretraživanja: odaziv i preciznost) - pojam koji se pojavljuje često u pojedinim dokumentima, ali rijetko u cijeloj zbirci *(Salton, 1989, 275-230).*

AUTOMATSKO PREDMETNO OZNAČIVANJE I UMJETNA INTELIGENCIJA

Razvoj računalne tehnologije, jača i brža računala dala su naročit zamah razvoju automatske obradbe teksta te umjetne inteligencije, odnosno obradbe prirodnog jezika *(natural language processing, NLP) (8)* - na kojem se zasniva suvremeni razvoj automatskog predmetnog označivanja.

Croft razlikuje tri osnovne skupine sustava za automatsko predmetno označivanje odnosno tri osnovna stupnja u razvoju ovih modela *(Croft, 1989):*

1. Jednostavni statistički model

Tehnike označivanje zasnovane na statističkom modelu određuju značenje teksta prema frekvenciji pojavljivanja pojedinih riječi u tekstu. Ovu metodu razvioje H. P. Luhn. Postupak jednostavnog automatskog predmetnog označivanja (statistički model) obično ima sljedeći tok:

- identificirati riječi unutar naslova, sažetka ili cjelovitog teksta dokumenta
- ukloniti ne-propisane termine *(stop words)*
- pronalaženje korijena riječi/osnovnog oblika riječi upotrebom jednostavnih postupnika
- zamjena korijena riječi s brojevima deskriptora • pobrojavanje pojavljivanja korijena
- izračunavanje težine *(weigths)* frekvencije (unutar dokumenta i obrnute frekvencije dokumenta)
- zamjena niske "diskriminacijske vrijednosti" pojmova s pojmovima iz tezaurusa za pojmove s niskim frekvencijama te dodjeljivanje fraza za pojmove s visokom frekvencijom pojavljivanja

(prema Croft, 1989 i prema Salton, 1989)

Nakon definiranja korpusa teksta, uklanjanja ne-deskriptora tj. učestalih i neinformativnih riječi s popisa stop-liste, pronalaze se korijeni riječi (odnosno osnovni oblici kod fleksivnih jezika poput hrvatskog) korištenjem jednostavnog postupnika kojim se uklanjaju sufiksi. Taj postupak je kod nefleksivnih jezika "korjenovanje" *(stemming)* ili kod fleksivnih jezika (kojima se traži osnovni oblik riječi, a ne korijen) lematizacija.

Nakon toga se korijen odnosno osnovni oblik riječi u takvim popisima zamijeni brojevima kako bi se sustav učinio efikasnijim sa stajališta pohrane i pobrojavanja pojavljivanja u tekstu. (9) U tom statističkom dijelu se dakle utvrduje točan broj pojavljivanja pojedine riječi u tekstu. Uz to se po potrebi može koristiti i tezaurus za otklanjanje sinonimije uspoređivanjem riječi iz tezaurusa koje korespondiraju skupinama riječi u tekstu.

Sljedeći korak je izračunavanje "težine" za odabrane pojmove. Te "težine" su kombinacija važnosti pojma izmjerenog prema frekvenciji njegovog pojavljivanja u pojedinom dokumentu, nasuprot njegovoj "težini" u čitavoj zbirci dokumenata. Istraživanja koja su do sada napravljena potvrduju da čak i ova najjednostavnija metoda automatskog označivanja cjelovitog teksta postiže jednako dobre rezultate kao i predmetno označivanje koje izvodi čovjek koristeći kontrolirani rječnik (usp. *Croft, 1989, 90-91).*

Kad je riječ o relevantnosti odaziva prilikom pretraživanja, vrlo je teško dati prednost bilo kojem od ta dva načina pretraživanja. Prema istraživanju koja je proveo Salton, ručno i automatsko označivanje ovog tipa daju podjednak broj relevantnih pogodaka, no među njima vrlo često nema preklapanja (usp. *Croft, 1989., 91).*

Usavršavanje statističkog modela modelima za pretraživanje - model vektorskog prostora, model klaster analize i model vjerojatnosti

Tehnike automatskog predmetnog označivanja u početku svog razvoja bile su zamišljene potpuno neovisno o području na kojem su se primjenjivale. Dok se kod ručnog predmetnog označivanja stvaraju kontrolirani rječnici za pojedino stručno područje, automatsko označivanje oslanjalo se samo na statistiku tekstova iz pojedinog područja koja se kontrolirala samim tekstovima. Značajan napredak učinjen je pronalaženjem dodatnih izvora informacija *(different source of evidence)* koji se

kombiniraju s automatskim označivanjem i formaliziranjem postupka kontroliranja i korekcije pojmova određenih u postupku automatskog označivanja. Kombiniranje modela pretraživanja i uključivanje povratne sprege u korekciju označivanja dokumenta značajno je unaprijedilo razvoj automatskog predmetnog označivanja.

Modeli za pretraživanje koriste različite strategije: model vektorskog prostora, model vjerojatnosti i model klastera. Prva dva modela zasnivaju se na pretpostavkama kako čovjek predstavlja sadržaj. Nova istraživanja nastoje koristiti ove metode kao alternativne izvore podataka uz jednostavnu metodu automatskog označivanja. Prvo se gradi model koji će formulirati informacijski zahtjev. Tzv. inteligentni posrednički sustavi *(intelligent intermediary systems)* dizajnirani su za formuliranje informacijskih zahtjeva. Metodom moguće podudarnosti *(plausible inference)* uspoređuju se upiti i sami dokumenti, a potom se metodom povratne sprege rafiniraju odgovori. Za formaliziranje procesa moguće podudarnosti pri označivanju dokumenta koriste se i svi drugi raspoloživi alati. Ako postoje ručno kontrolirani rječnici, kazalo citata, tezaurusi, oni se uzimaju kao dodatni izvori informacija *(source of evidence)* kojima se provjerava relevantnost nekog dokumenta u odnosu na traženi upit. Na ovaj se način već neko vrijeme u *online* bazama podataka kombinira pretraživanje slobodnog teksta i predmetnica *(Library of Congress Subject Headings* primjerice). Nastoji se uključiti sve raspoložive postojeće forme u združenu strategiju pretraživanja, kako bi se preklapanjem dobila informacija o relevanfiosti.

Statistički model automatskog označivanja povezan s bazom znanja iz područja i postupkom pretraživanja (knowledge based information retrieval)

Sljedeći veliki korak u razvoju automatskog predmetnog označivanja zasniva se na kombinaciji sustava za označivanje i pretraživanje s bazom znanja nekog područja *(domain knowledge base).* Model u cjelini jest zapravo sustav za pretraživanje zasnovan na znanju (knowledge based infromation retrieval). I predstavljanje sadržaja dokumenata i postupak pretraživanja ostvaruje se obradom prirodnog jezika. To omogućuje puno kompleksnije određivanje konteksta pojma nego što je to moguće s predmetnicama ili deskriptorima. Ovakvo predstavljanje sadržaja dokumenta zasniva se na "obrascima" *(case frames)* koji izražavaju odnose predmeta unutar pojedinog područja. Da bi ovo bilo moguće potrebno je prethodno izgraditi bazu znanja iz područja koja će se sastojati od svih pojmova i koncepta unutar područja (tezaurus) i od načina na koji se oni mogu povezivati i odnositi jedan na drugi.

Za ovaj sustav program mora "znati" nešto o području u kojem će se obavlja pretraživanje, nešto o predmetima iz tog područja i odnosima koji postoje među njima. Da bi se tekst mogao obraditi, program treba posjedovati lingvističko znanje kako bi mogao staviti predmete u odnose na način na koji se o tome govori u tekstu ili formulira upit u postupku pretraživanja.

Spomenuta Croftova podjela i evolucija sustava može se prepoznati i kod drugih autora. Obzirom da svaki sustav za automatsko predmetno označivanje nužno počiva na obradbi prirodnog jezika, kategorizaciji ovih sustava može se pristupiti i obzirom na model analize prirodnog jezika. Ocjenjujući neke sustave za francusko govorno područje, W. Mustafa-Elhadi, govoreći o prirodi pristupa kad se radi o lingvističkom aspektu te namjeni ovih sustava, razlikuje četiri kategorije (uz napomenu da većina sustava ipak istodobno ulazi u više kategorija):

1. sustavi za prepoznavanje pojmova *(term extractors)* - baziraju se na sintaktičkim sustavima kojima se ponekad dodaju moduli za statističku obradbu. Ovi sustavi koriste setove gramatičkih riječi i interpunkcijskih znakova u svrhu izolacije pojmova iz teksta. Primjeri: ACABIT, ANA, KES;
2. sustavi za klasifikaciju *(classifying tools).* Ti sustavi stvaraju klase mreža pojmova povezanih u hijerarhiju, sastoje se od čistih statističkih modela i modela za povezivanje. Primjer: CONTERM;
3. sustavi za prepoznavanje semantičkih odnosa *(semantic relations extractors).* Ovi sustavi posebnu pozornost poklanjaju semantičkim odnosima. Primjeri: IOTA (sustav koji identifikaciju pojmova integrira s prototipom sustava za pretraživanje), SEEK i SPIGRAPHE.
4. sustavi koji integriraju pohranu i pretraživanje informacija i mogu se svrstati u kategoriju inteligentnih sustava *(intelligence system).*

Primjer:

NOEMIC - analiza semantičkih odnosa (tezaurus), kojoj ne prethodi morfološka ili sintaktička analiza. Ovaj sustav analizira sadržaj baze punog teksta i uspoređuje je s predefiniranom strategijom pretraživanja.

SPIRIT - sustav za automatsko predmetno označivanje zasnovan na višejezičnoj analizi i statističkom modelu, povezan sa sustavom za pretraživanje i s mogućnošću korekcije relevantnosti odgovora. Sustav omogućava da korisnički upiti o pojedinoj temi na prirodnom jeziku postaju dio sustava. Kao odgovor na upit sustav daje listu relevantnih dokumenata uređenih prema padajućoj relevantnosti. *(Mustafa-Elhadi, 1997)*

ZAKLJUČAK

Uslijed tehnoloških promjena koje su omogućile rad s velikim bazama podataka, stručnjacima u organizaciji i

posredovanju informacija i znanja pridružili su se i znalci s područja računarstva. Istraživački rad koji se odnosio na pretraživanje informacija istom je usmjeren na pretraživanja slobodnog teksta, odnosno pretraživanje na prirodnom jeziku. Pristup računalnih stručnjaka problemu pretraživanja i označivanja usredotočen je od početka na lociranje i pronalaženje informacija u nekoj bazi podataka i na relevantnost dobivenih odgovora. Od samog početka se tražilo rješenje pomoću kojeg bi se velike baze dokumenata mogle uspješno sadržajno pretraživati, da ti dokumenti ne moraju prethodno biti analizirani i označeni od strane čovjeka. Danas, govoreći o sustavima za pretraživanje i označivanje, a napose o automatskom predmetnom aznačivanju, možemo zaključiti da su sva nastojanja u tom pravcu bila opravdana. Brojna literatura iz tog područja govori u prilog silnom znanstvenom-istraživačkom i praktičnom naporu koji je uložen u njihov razvoj i usavršavanje.

U počecima razvoja računalne tehnologije rezultati su bili prilično skromni, a entuzijazam računalnih stručnjaka djelovao je prilično neuvjerljivo (na silno olakšanje knjižničara i dokumentalista starog kova). Računalni znalci koji su se bavili tim područjem često su odbijali uzimati u obzir pomagala za organizaciju znanja, kojima su se informacijski stručnjaci služili stoljećima. Postojeće klasifikacijske sheme i tezaurusi bili su prepuni nedosljednosti, nerazumljivo i nepotrebno kompleksni, a napor koji je potreban da se oni prilagode tadašnjim standardu računalne tehnologije, prevelik i preskup. Na žalost, iznimno rijetko se upuštalo u proučavanje koncepcija na kojima počivaju sustavi jezika za označivanje u svrhu njihove formalizacije i približavanja računalnim sustavima.

S druge strane, tradicionalni informacijski stručnjaci (i knjižničari i dokumentalisti), koji se stoljećima bave posredovanjem znanja i informacija, razmišljali su i dalje u pravcu razvoja svojih ručnih pomagala za organizaciju znanja. Izgradnja tih alata kao i sam posao sadržajne obradbe zahtijeva veliku stručnost i veliki intelektualni napor. To je posao koji je oduvijek slovio za elitni intelektualni posao i knjižničari nisu vjerovali u mogućnost da će na tom području, razvojem tehnologije, moći doći do neke bitne promjene.

Napretkom komunikacijske tehnologije i razvojem globalne informacijske infrastrukture, napose pojavom Interneta, omogućen je pristup velikom broju različitih informacija u digitalnoj formi. Među njima iz dana u dan raste broj baza podataka s punim tekstom. Potreba označivanja i pretraživanja stalno rastuće mase dokumenata već sada nadilazi ljudske mogućnosti. No iskustva posljednjih desetljeća, na području označivanja i pretraživanja, pokazala su da najbolje rezultate daju oni sustavi u kojima se kombinira rad čovjeka i računala *(usp. Milstead, 1996; Croft, 1989).*

Na primjeru razvoja automatskog predmetnog označivanja vidimo kako se posljednjih desetljeća dosta toga promijenilo u pogledu pretraživanja i označivanja dokumenata i kako niti računarski stručnjaci, niti knjižničari nisu bili potpuno u pravu ignorirajući jedni druge. Srećom, u posljednje vrijeme automatsko predmetno označivanje nastoji se kvalitetno združiti s tradicionalnim pomagalima za označivanje poput klasifikacija i tezaurusa *(usp. Schuegraf, 1997 i Ferber, 1997).* Iskustvo i znanje stručnjaka za sadržajnu obradbu te tradicionalna pomagala (tezaurusi, klasifikacije) uzimaju se u obzir pri programiranju znalačkih sustava i mogu osigurati njihovu kvalitetnu primjenu.

Stručnjaci koji se bave umjetnom inteligencijom, naročito obradbom prirodnih jezika, područje označivanja i pretraživanja informacija prepoznaju kao područje idealno za primjenu svojih dostignuća. Neki sustavi koje spominju Croft i Mustafa-Elhadi daju vrlo dobre rezultate i govore u prilog sljedećoj Croftovoj tvrdnji:

"Automatsko, kontrolirano predmetno označivanje koje se zasniva na bazi znanja nekog područja može davati bolje rezultate jer je: dosljednije od Ijudskog predmetnog označivanja, detaljnije predstavlja znanje, jer je stvoreno uzimajući u obzir poces pretraživanja. "

(Croft, 1989, 99)

No valja se ovdje osvrnuti na Croftovo upozorenje o dosegu umjetne inteligencije. Bazu znanja na kojoj se ovi naizgled "savršeni" sustavi zasnivaju kao i tezauruse potrebne da se ovo znanje predstavi još uvijek nije moguće stvoriti bez sudjelovanja čovjeka i pitanje je kad će to biti moguće.

No čak i ne spominjući one krajnje dosege, treba naglasiti da automatsko predmetno označivanje baza s punim tekstom treba imati na umu kao jeftinu i efikasnu metodu uz koju je često korisno i moguće kombinirati i ručno predmetno označivanje, odnosno bilo koju vrstu ljudske kontrole. Za englesko govorno područje postoje i moguće je lako i jeftino nabaviti neke gotove programe za automatsko predmetno označivanje (primjerice *Indexicon,* MACREX, CIDEX) čije su skromne mogućnosti srazmjeme njihovoj cijeni (cijena *Indexicona* je 129$!). Ozbiljni programi s kojima se indeksiraju velike baze poput INSPEC-a i BIOSIS-a pripadaju u posve drugu kategoriju *(Brown, 1997).* Bez obzira na kategoriju, dosege i cijenu, sustavi za automatsko predmetno označivanje su na putu da postanu uobičajena pomagala i dio svakodnevnice u knjižnicama koje imaju građu u elektroničkom obliku, a načini na koji se oni grade i kako funkcioniraju od velikog su utjecaja na označivanje i pretraživanje dokumenata uopće.

BILJEŠKE

1. Počeci automatskog predmetnog označivanja vezuju se uz model koji je osmislio H. P. Luhn i koji je objavljen u radu A statistical approach to the mechanized encoding and searching of literary information objavljenom u IBM Journal of

Research and Development, 28, (1957)1.

2. Pretraživanje se zasniva na tome da sustav za pretraživanje gradi obrnuta (invertirana) kazala riječi svih dokumenata pomoću kojih se u bazi pronalazi točno određeni dokument.
3. Prigovori tradicionalnim metodama označivanja i pretraživanja mogu se svesti na sljedeće: skupoća, nedosljednost i subjektivni pristup u analizi dokumenata (usporedi Croft, 1989, 86).
4. Preko 92 milijuna godišnje dok se broj fotokopija kreće izmedu 300 i 400 milijuna (prema Encyclopedia of library and information science : volume 54, supplement 17., 1994. 98).
5. Ono što bi omogućilo efikasnost ovih programa jest označivanje dokumenata pomoću metapodataka (metadata) koji će, zapisani u izvoru samog dokumenta, osigurati njegovu pravilnu katalogizaciju i klasifikaciju na Internetu.
6. Razvoj umjetne inteligencije omogućava razvoj računalnih programa nazvanih znalačkim sustavima (expert system) koji su sposobni, na osnovi pohranjenih podataka i pomoću postupnika (algoritama) definiranog ponašanja proizvoditi pravilne zaključke i donositi odluke, tj. djelovati inteligentno. Jedna od temeljnih ideja u početku razvijanja ovog područja bila je da će ti programi u potpunosti zamijeniti čovjeka u proizvodnom procesu ili u područjima ljudske djelatnosti gdje nije moguće postići po život sigurne radne uvjete ili je to naprosto ekonomski neisplativo. Danas, međutim, znalački sustav ima vrlo široko polje primjene, najvećim dijelom kao pripomoć čovjeku u donošenju pravilnih odluka ili kontroli od onih manje važnih do vrlo odgovornih profesija. I knjižničarstvo je svakako jedno od područja gdje će znalački sustavi nalaziti sve veću primjenu. *(Lancaster; Sandore 1997)*
7. Treba naglasiti da je izraz "indeksiranje" preuzet iz anglo-američke knjižničarske prakse gdje on ima ponešto drukčije značenje. Naime, izrazom "indexing" u anglo-američkom knjižničarstvu najčešće se označava sadržajna obradba dokumenta tj. predmetno označivanje i klasificiranje zajedno. U računalnom okruženju indeksiranje može imati i bitno šire značenje, pa G. Salton, govoreći o indeksiranju razlikuje "objective identifier" i "nonobjective identifier". Pod "objective identifiers" podrazumijeva formalno popisivanje dokumenta: autor, nakladnik, datum izdavanja, broj stranica itd. napominjući da kataložna pravila daju točno propisane postupke za tu vrstu označivanja. Pod "nonobjective identifiers" on podrazumijeva sadržajno ili stvarno označivanje dokumenata za koje je karakteristično da nema točno propisanih i kodiranih pravila, a razlikujemo ručno od strojnog, postupak označivanja jednostavnim pojmovima i postupak označivanja pojmovima u kontekstu, odnosno prekooridinirano i postkoordinirano označivanje (Salton, 1989). Drugi autori razlikuju još i stvarno označivanje prema tome da li se radi o posve umjetnom jeziku za označivanje (primjerice klasifikacija) ili jeziku za označivanje zasnovanom na prirodnom jeziku (razne vrste predmetnog označivanja koje se koriste rječima prirodnog jezika).
   U hrvatskoj knjižničarskoj praksi posve razdvajamo formalnu od sadržajne obradbe dokumenata, a razlikujemo i u praksi i terminološki dva aspekta stvarne, odnosno sadržajne obrade: klasifikaciju i predmetizaciju. Engleski izraz "indeksiranje" (indexing) također se koristi u našoj literaturi i to najčešće ne u smislu sadržajne ili stvarne obradbe uopće (što je njegovo pravo značenje) već za razlikovanje načina predmetnog označivanja kad se pod njim podrazumijevaju svi stupnjevi i oblici korištenja prirodnog jezika za označivanje sadržaja dokumenta (ključne riječi, razni deskriptorski sustavi, korištenje tezaurusa kao alata za predmetno označivanje i sl.) koji služe pretraživanju u izravnim računalnim katalozima ili bazama podataka. Na taj način razlikuje se tradicionalna predmetizacija dokumenata u svrhu izradbe predmetnog kataloga na listićima od indeksiranja kao modernog koncepta upotrebe prirodnog jezika za označivanje sadržaja dokumenata u širem smislu. Danas bi, možda, trebalo preispitati potrebu za naglašavanjem ove razlike s obzirom na to da je predmetna obradba u suvremenom knjižničarstvu usmjerena na računalne kataloge.
8. Naučiti računalo da razumije prirodni jezik, da njime vlada ili da ga proizvodi oduvijek je zaokupljalo računalne stručnjake. Obradba prirodnog jezika (natural language processing) veže se najvećim dijelom za domenu računarstvu koja se naziva umjetnom inteligencijom. Kako se informacija predstavlja korisniku u obliku prirodnog jezika, nužno svaki sustav za predmetno automatsko označivanje u sebi sadrži ugrađen program za obradbu prirodnog jezika. Okosnica programa za obradbu prirodnog jezika jest semantika. Svaki ovakav suvremeni sustav ovisi o znanju (knowledge dependent) i sadrži znanje rječnika, znanje o sintaksi i znanje o pojedinom području. *(Prasad, 1995.)*
9. Automatsko predmetno označivanje kao i sustavi za pretraživanje uobičajeno kreiraju posebnu datoteku, obično nazvanu obrnuta (invertirana) datoteka koja je zapravo kazalo dodijeljenih pojmova i oznaka njihove lokacije u tekstu:

   Primjer:

pojam	ključ smještaja (locator key)
kabanica	ENCP01L02
katalizator	ENCP01L04
kiša	ENCP01L23
kišobran	ENCP01L05

LITERATURA

1. **Automated indexing.** // Encyclopedia of library and information science : volume 54, supplement 17. - New York : Marcel Dekker, 1994. Str. 98-121.
2. **Automatic indexing.** 1-7. http://pc-sx129.lut.ac.uk/air/air2.html (24.05.1997.)
3. **Browne,** G. Automatic indexing and abstracting. // Online Currents, AuSI Newsletter 20 (1996)6, 4-9. http://www.zeta.org.au/aussi/browneg.htm (24.05.1997.)
4. **Buckland,** M. et al. Partnership in navigation : an information retrieval research agenda. // paper presented at ASIS Annual Meeting, Chicago, October, 1995. 1-9. http://www.asis.org/asis95/papers/norgard.html (20.09.1997.)
5. **Croft,** B. W. Automatic indexing. // Indexing : the state of our knowledge and the state of our ignorance : proceedings of the 20th Annual Meeting of the American Society of Indexers. / edited by Bella Hass Weinberg.- Medford NJ : Learned Information, 1989. Str. 86-100.
6. **Croft,** B. W. What do people want from information retrieval. // From classification to "knowledge organization" : Dorking revisited or "past is prelude" / edited by Alan Gilchrist.The Hague : International Federation for Information and Documentation, 1997. Str. 181-185.
7. **Ferber, R.** Automated indexing with thesaurus descriptors : a co-occurence based approach to multilingual retrieval. // Research and Advanced Technology for Digital Libraries, First European Conference.- Pisa, September 1-3, 1997. Str. 233-252.
8. **Information** retrieval experiment / edited by Karen Sparck Jones.- London : Butterworths, 1981.
9. **Intelligent** help desk systems. // Encyclopedia of library and information science : volume 54, supplement 17.- New York : Marcel Dekker, 1994. Str. 212-224.
10. **Jones**, G. et al. Non-hierarchic document clustering using a genetic algorithm. URL http://http://www.shef.ac.uk/uni/acad...ic/I-M/is/lecturer/paperl.html (03.04.1997.)
11. **Lancaster**, F. W.; Sandore, B. Technology and management in library and information services.- London : Library Association Publishing, 1997. Str. 226-250.
12. **Locke**, C. Intelligent agents create dumb users (?) : paper presented at Libtech `97 (workshop Advanced search techniques). URL http://www.ucl.ac.uk/SLAIS/agents.htm
13. **Milstead**, J. L. Needs for research in indexing. // From classification to "knowledge organization" : Dorking revisited or "past is prelude" / edited by Alan Gilchrist.- The Hague : International Federation for Information and Documentation, 1997. Str. 151-159.
14. **Mulvany**, N. Comments on Steinberg's article about Web indexing. June 4, 1996. http://www.mnsinc.com/curr/nanindex.htm (16.09.1997.)
15. **Mustafa-Elhadi**, W. Natural language processing -based techniques and their use in data modelling and information retrieval : paper presented on the 6th International Study Conference on Classification Research, London, 16-18 June 1997.
16. **Prasad**, A. R. D. Prometheus : an automatic indexing system. // Proceeding of the 4th ISKO conference 1996.- Washington, DC : LC, 1996. Str. 329-335.
17. **Richardson**, J. V. Knowledge-based systems for general reference work: applications, problems, and progress.- San Diego : Academic Press, 1995.
18. **Rowley**, J. Abstracting and indexing. - 2nd ed.- London : Clive Bingley, 1990.
19. **Salton**, G. Automatic text processing.- Reading, MA : Addison-Wesley Publishing Company, 1989.
20. **Salton**, G. Brief communication a note about information science research. // From classification to "knowledge organization" : Dorking revisited or "past is prelude" / edited by Alan Gilchrist.- The Hague : International Federation for Information and Documentation, 1997. Stt. 131-134.
21. **Salton**, G. The Smart environment for retrieval system evaluation - advantages and problem areas. //

Information retrieval experiment / edited by Karen Sparck Jones. - London : Butterworths, 1981. Str. 316-329.

22. **Schuegraf**, E. J. Classification as an aid to automatic indexing : paper presented on the 6th International Study Conference on Classification Research, London, 16-18 June 1997. Sparck Jones, K. The Cranfield tests. // Information retrieval experiment / edited by Karen Sparck Jones.- London : Butterworths, 1981. Str. 256-284.
23. **Sparck** Jones, K. The Cranfield tests // Information retrieval experiment / edited by Karen Sparck Jones. - London : Butterworths, 1981. Str. 256-264.