Mr Franjo Žerjav, predavač
Fakultet organizacije i informatike
Varaždin

UDK: 002
Stručni rad

RELEVANTNOST INFORMACIJA I KRITERIJI ZA OCJENJIVANJE EFIKASNOSTI SISTEMA ZA PRETRAŽIVANJE INFORMACIJA

Pojam relevantnosti ima ključnu ulogu kod vrednovanja sistema za pretraživanje informacija. U radu se detaljno raščlanjuje problem relevantnosti, definiraju se njezini pojavni oblici, naznačuju se metode i pokušaji objektivizacije i ukazuje se na buduće pravce razvoja istraživanja. Prikazani su kriteriji za ocjenjivanje efikasnosti sistema za pretraživanje informacija u kojima se kao jedan od parametara pojavljuje relevantnost.

1. UVOD

Opća efikasnost informacijskog sistema mjeri se kao sveukupnost triju pokazatelja - društvene, ekonomske i tehničke efikasnosti. Društvena efikasnost informacijskog sistema ocjenjuje se njegovim utjecajem na tehnički, socijalni i kulturni progres okoline u kojoj sistem djeluje. Ocjenjivanje može biti vrlo kompleksno jer treba uzimati u obzir niz najrazličitijih parametara. Ekonomska efikasnost podrazumijeva odnos troškova izgradnje, održavanja, obrade informacija i koristi od rezultata odredjene informacijske djelatnosti. U praktičnoj izvedbi zahtijeva se da trošak izgradnje i obrade bude manji od koristi. Tehnička efikasnost ocjenjuje relativno samostalan i od okoline isključeni sistem. Zasniva se na ocjenjivanju i mjerenju unutarnjih svojstava informacijskog sistema, koja se općenito zasnivaju na relevantnosti informacije izdate korisniku.

Cilj ovog rada je razmatranje slijedećih pitanja: kriteriji za ocjenjivanje efikasnosti ovisni o relevantnosti informacija, pojavni oblici relevantnosti u pretraživanju informacija te logička i lingvistička sredstva identifikacije dokumenata i zahtjeva.

U teoriji pretraživanja informacija jedan od ključnih problema

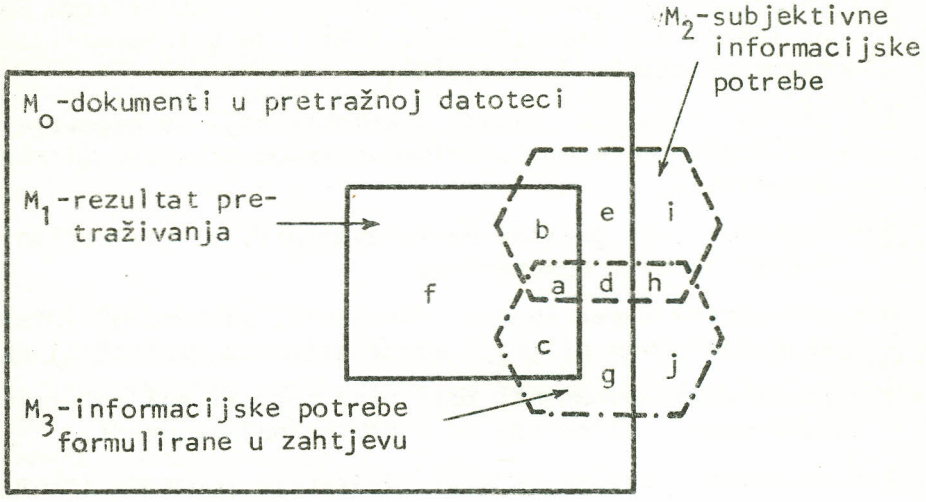
je odredjivanje relevantnosti informacija (dokumenata).*) Većina kriterija za ocjenjivanje efikasnosti sistema za pretraživanje informacija (kao što su: točnost pretraživanja, potpunost pretraživanja, pretražni šum, koeficijent pertinentnosti i dr.) zasniva se na pojmu relevantnosti izdate ili neizdate informacije. Poteškoće se pojavljuju zbog toga što pojam relevantnosti sadrži aspekt individualno-pragmatičke ocjene korisnika o vrijednosti nadjene informacije, a što u znatnom stupnju zavisi od psiholoških faktora i ne podliježe iscrpnoj formalizaciji (1,5).

Informacijske potrebe ne uspijevaju se uvijek točno,jednoznačno i iscrpno formulirati u obliku informacijskog zahtjeva. Pod jednim te istim informacijskim zahtjevom različiti korisnici, u principu, mogu podrazumijevati različite informacijske potrebe. Elementi subjektivizma, koji neminovno prate postupak jezičnog formuliranja različitih semantičkih kategorija, postali su razlog mnogobrojnih znanstvenih sporova oko pojma relevantnosti (pertinentnosti), neophodnog za ocjenjivanje stupnja sličnosti nadjene informacije i informacijskog zahtjeva (informacijskih potreba).

2. KRITERIJI OCJENJIVANJA EFIKASNOSTI OVISNI O RELEVANTNOSTI INFORMACIJA

Kod pretraživanja pojavljuje se nekoliko vrsta informacija(dokumenata). Kriterij klasifikacije je značenje informacije za primaoca (korisnika) i relevantnost rezultata pretraživanja (2,202). Te vrste informacije prikazane su na sl.1, uz dodatna objašnjenja.

*) *U informacijsko-dokumentacijskoj djelatnosti pojmovi informacija i dokument su sinonimi jer su informacije obično materijalizirane u obliku nekog primarnog ili sekundarnog dokumenta.*



Sl.1. Vrste informacija (dokumenata) kod pretraživanja u odnosu na njihovu pertinentnost i relevantnost

M_o - potencijalno relevantni dokumenti u pretražnoj datoteci sistema za pretraživanje informacija,

M_1 - dokumenti izdati iz pretražne datoteke pri nekom pretraživanju,

M_2 - pertinentni dokumenti, tj. dokumenti koji stvarno odgovaraju postojećim informacijskim potrebama korisnika,

M_3 - relevantni dokumenti, tj. dokumenti koji odgovaraju informacijskim potrebama formuliranim u informacijskom zahtjevu,

a - izdati pertinentni dokumenti,

b - slučajno pogodjeni i izdati pertinentni dokumenti (greškom u pretraživanju), premda nisu obuhvaćeni informacijskim zahtjevom,

c - relevantni dokumenti koji odgovaraju informacijskom zahtjevu ali koji korisniku nisu interesantni jer su mu već poznati ili se iz subjektivnih razloga na njih ne osvrće itd,

d - informacijski gubitak na pertinentnim dokumentima koji se ne izdaju kod pretraživanja premda su obuhvaćeni informacijskim zahtjevom,

e - informacijski gubitak na pertinentnim dokumentima koji se ne izdaju kod pretraživanja i koji su u informacijskom zahtjevu nepotpuno formulirani,

f - pretražni balast, izdati dokumenti koji ne odgovaraju formuliranim potrebama u informacijskom zahtjevu (irelevantni dokumenti),

g - informacijski gubitak na relevantnim, ali za korisnika ne i pertinentnim dokumentima,

h - relevantni i pertinentni dokumenti, obuhvaćeni informacijskim zahtjevom ali koji se ne nalaze u pretražnoj datoteci,

i - pertinentni dokumenti koji nisu formulirani informacijskim zahtjevom i ne nalaze se u pretražnoj datoteci,

j - relevantni dokumenti koji odgovaraju informacijskom zahtjevu ali nisu interesantni za korisnika i ne nalaze se u pretražnoj datoteci

Najznačajniji kriteriji ocjenjivanja efikasnosti sistema za pretraživanje informacija ovisni o relevantnosti informacija *) jesu: potpunost pretraživanja, točnost pretraživanja, pretražni šum i koeficijent pertinentnosti. U nastavku će se ukratko prikazati odnosi koje prate gore navedeni kriteriji.

Potpunost pretraživanja

Potpunost pretraživanja (P) odredjuje odnos nadjenih i izdatih relevantnih dokumenata prema ukupnom broju relevantnih koji se nalaze u pretražnoj datoteci, a koji odgovaraju nekom informacijskom zahtjevu. Odredjuje se po formuli:

$$P = \frac{100(a+b+c)}{(d+g)} \quad (u\ \%)$$

Točnost pretraživanja

Točnost pretraživanja odredjuje odnos nadjenih i izdatih relevantnih dokumenata prema ukupnom broju izdatih dokumenata na odredjeni informacijski zahtjev. Odredjuje se po formuli:

$$T = \frac{100(a+b+c)}{a+b+c+f} \quad (u\ \%)$$

*) *Osim ovih kriterija postoje i drugi koji prate druge parametre, kao što su: vrijeme pretraživanja, troškovi i sl.*

Pretražni šum

Pretražni šum definira se kao odnos nerelevantnih dokumenata, izdatih od sistema za pretraživanje na informacijski zahtjev, prema ukupnom broju izdatih dokumenata. Pretražni šum(S) određuje se po formuli:

$$S = \frac{100 \cdot f}{a+b+c+f} \text{ (u \%)}$$

Koeficijent pertinentnosti

Koeficijent pertinentnosti (K) uzima u obzir korisnikovu definiciju pertinencije udjela svih do tada korisniku nepoznatih,a informacijski značajnih dokumenata, dobivenih kao rezultat pretraživanja. Izračunava se po formuli:

$$K = \frac{100 \ (a+b)}{a+b+c+f} \text{ (u \%)}$$

3. VRSTE RELEVANTNOSTI U PRETRAŽIVANJU INFORMACIJA

Postoji suglasnost niza autora (1,6) da se višeaspektni pojam "relevantnosti" ne može jednoznačno definirati, već da je svrsishodno diferencirano razmatranje koje omogućava precizno razgraničenje pojmova relevantnosti i pertinentnosti. Kod detaljnijeg razmatranja mogu se razlikovati tri vrste relevantnosti (2,200):

- sintaktička ili strukturna relevantnost,
- semantička relevantnost i
- pragmatička relevantnost.

Sintaktička, strukturna ili formalna relevantnost postignuta je tada ako postoji, prema pretražnoj logici zahtijevano, podudarnost izmedju pretražnog tipa izdatog dokumenta*) i pretražnog uputstva.**) Dakle, postiže se podudaranjem jednih ili drugih

*) *Pretražni tip dokumenta je formalizirani sadržaj informacija, datih u dokumentu na prirodnom jeziku. Formalizacija se vrši u cilju pretraživanja informacija i postiže se korištenjem jezika za indeksiranje.*

**) *Pretražno uputstvo je formalizirani oblik informacijskog zahtjeva, prikazan na jednom od jezika za indeksiranje.*

komponenata formaliziranih tekstova zahtjeva i dokumenata. Postizavanjem sintaktičke relevantnosti još nije osigurana i semantička relevantnost (sadržajna ili stvarna relevantnost) izdatog dokumenta. Za postizavanje semantičke relevantnosti nužan je daljnji dodatni atribut da s a d r ž a j izdatog dokumenta odgovara informacijskim potrebama korisnika izraženim u informacijskom zahtjevu (suglasnost sadržaja dokumenta sadržaju zahtjeva).

Problemi osiguravanja semantičke relevantnosti jesu objekt istraživanja teorije o jezicima za indeksiranje, tj. o informacijskim jezicima sistema za pretraživanje informacija. Sintaktička i semantička definicija relevantnosti usko su vezane s gnoseološkim osnovama uzajamnog odnosa jezika i mišljenja.

Pri razmatranju pojma relevantnosti važno mjesto daje se njezinoj trećoj komponenti - pragmatičkoj ili subjektivnoj relevantnosti ili, kako se najčešće naziva u stručnoj literaturi, pertinentnosti. Pragmatička relevantnost postignuta je tada ako korisnik s t v a r n o koristi informacije sadržane u izdatim dokumentima, ako se njegova informiranost povećava.

Pertinentnost izdatog dokumenta ne ovisi samo o njegovom stvarnom sadržaju. Uz objektivne činjenice, koje pertinentnost utvrđuje, utječe takodjer i čitav niz s u b j e k t i v n i h koje postoje u osobi korisnika. Izvanredno značajnu ulogu imaju ovdje činjenice koje postoje u osobama pojedinih korisnika, kao što su: pristup k informiranju i informacijama, njegova individualna organizacija rada, znanje stranih jezika, raspoloživo vrijeme za obradu informacija, pripadnost odredjenoj znanstvenoj školi, stav prema odredjenim autorima, institucijama i publikacijama te sl.

Ukazujući na subjektivni karakter pertinentnosti J.Bar-Hillel smatra (1,9) da je razrada metoda, koje bi dale prihvatljive mjere za ocjenjivanje razlike izmedju pojedinih tema, "nedostižna" ili barem "djavolski teška". Jedan od utemeljitelja teorije pretraživanja informacija, američki znanstvenik Mortimer Taube, smatra da je nemoguće izgradjivati bilo kakav matematički model za ocjenjivanje efikasnosti funkcioniranja sistema za pretraživanje informacija koji se zasniva na subjektivnom pojmu pertinentnosti. Medjutim, čitav niz drugih autora smatra ako smo svijesni prisustva subjektivne komponente u pojmu relevantnosti,da se pri vještom korištenju iz nje može izvući odredjena korist.

L. Doyle (3) smatra da je relevantnost intelektualna "kvaka" kojom mi, premda i ne strogo, možemo razmišljati o problemu pretraživanja informacija, a bez nje o tome uopće ne bismo mogli razmišljati. Kao dokaz tome služe mnogobrojna teoretsko-eksperimentalna istraživanja specijalista u svijetu koji široko primjenjuju statističke metode obrade podataka pri analizi psiholoških aspekata pojma relevantnosti-pertinentnosti. Predmet analize uglavnom su indeksatori (izdvajaju se tipološke grupe indeksatora, sa svojstvenim im karakteristikama načina mišljenja, stilom rada i sl.) i korisnici informacija (tipologija korisnika).

Kod projektiranja automatiziranih sistema za pretraživanje informacija potrebno je imitirati one čovjekove intelektualne mogućnosti koje omogućavaju ostvarivanje prijelaza od sintaktičke k semantičkoj i pragmatičkoj relevantnosti. Postojeći automatizirani sistemi u svijetu oslanjaju se u postupku pretraživanja na sintaktičku relevantnost, generiranu elektroničkim računalom, tj. na podudarnost jednih ili drugih komponenti zahtjeva i dokumenata. Kod toga elektroničko računalo operira logičkim i lingvističkim sredstvima identifikacije koja se zajedno nazivaju kriterijima semantičke sličnosti dokumenata (informacija) i zahtjeva, tj. različitim sredstvima (logičkim i lingvističkim) nastoji se postići semantička relevantnost. Postojeći sistemi još uvijek ne pokušavaju postignuti pragmatičku relevantnost u opsluživanju korisnika.

U nastavku će se ukratko razmatrati logička i lingvistička sredstva identifikacije dokumenata i zahtjeva za postizavanje semantičke relevantnosti.

4. LOGIČKA SREDSTVA IDENTIFIKACIJE DOKUMENATA I ZAHTJEVA

Pri razmatranju sintaktičke i semantičke relevantnosti isključujemo iz razmatranja pragmatički aspekt pojma relevantnosti.

Prvi je pokušao uvesti kvantitativno mjerenje informacija, uz isključivanje psiholoških faktora, R. Hartley 1928. god. 1933. god. V. V. Koteljnikov je dokazao fundamentalni teorem o prikazivanju informacija konačnim brojem diskretnih saopćenja sadržanih u neprekinutim funkcijama s ograničenim spektrom učestalosti (1,11).

1949. god. C. Shannon uveo je univerzalnu mjeru formalno-kvantitativne ocjene informacija koja, medjutim, nije mogla pos-

lužiti kao mjera za iscrpnu ocjenu stupnja semantičke sličnosti dokumenta i zahtjeva. To se objašnjava poteškoćama povezanim s formalnim definiranjem i točnim mjerenjem semantičkog značenja informacije. Semantička teorija ili teorija informacija u širem smislu još je malo razradjena. Niz specijalista pretpostavlja pokušaje njegove izgradnje na bazi teorije informacija u užem smislu, tj. teoriji informacija u interpretaciji C. Shannona.

Bez obzira na spomenute teškoće principijelno-teoretskog karaktera praktički zahtjevi izgradnje automatiziranih sistema za pretraživanje informacija u različitim područjima ljudske djelatnosti dovode do stvaranja čitavog niza kriterija semantičke sličnosti pojedinih tekstova. Iako se smatra da je apsolutno točna ocjena semantičke sličnosti nemoguća, ti kriteriji se, medjutim, pokazuju sasvim prihvatljivim s točke gledišta njihove praktične primjene.

Jedan od jednostavnijih kriterija ocjene semantičke sličnosti definiran je kao postotak deskriptora pretražnog tipa dokumenta koji se podudara s deskriptorima pretražnog uputstva. Naknadno su specijalisti pojačali taj kriterij s dva, tri, ili višestupnjevanim rangiranjem deskriptora po njihovoj važnosti.

Daljnji razvitak metoda rangiranja deskriptora po njihovoj važnosti dovodi do stvaranja metode "težinskih" koeficijenata. Začeci metode pojavljuju se u stručnoj literaturi 1962. godine kada se daju detaljno opisani principi organizacije i rezultati eksploatacije sistema za pretraživanje informacija u kojima je ocjenjivanje semantičke sličnosti realizirano po metodi "težinskih" koeficijenata. Bit metode sastoji se u korisničkom rangiranju deskriptora pretražnog uputstva s "težinskim" koeficijentima i priznavanju relevantnosti dokumenta u toku pretraživanja ako je suma njegovih "težinskih" koeficijenata deskriptora veća od neke unaprijed odredjene veličine. Danas metoda "težinskih" koeficijenata ima široku primjenu i postoji čitav niz njezinih različitih modifikacija.

Usporedna analiza razmatranih kriterija ocjene semantičke sličnosti omogućuje zaključak, bez obzira na različite principe njihove izgradnje, da postojeći sistemi za pretraživanje informacija u pretežnoj većini sadrže u pretražnom tipu ključne riječi dokumenata. Pojavljuje se tendencija uključivanja u pre-

tražni tip dokumenta i drugih atributa, kao: prezime autora, bibliografski citati, različiti klasifikacijski indeksi i sl, što dovodi do mogućnosti višeaspektnog pretraživanja po različitim atributima.

Treba naglasiti da je posljednjih godina znatno porastao interes specijalista za metodu "težinskih" koeficijenata, a što je uvjetovano traženjem mogućnosti strojnog generiranja pretražnih uputstava u sistemima za pretraživanje informacija.

5. LINGVISTIČKA SREDSTVA IDENTIFIKACIJE DOKUMENATA I ZAHTJEVA

Sintaktička i semantička relevantnost u značajnoj mjeri ovisi o problemu racionalnog konstruiranja jezika za indeksiranje. Pod jezikom za indeksiranje ovdje se podrazumijeva jezik sistema za pretraživanje informacija. To je umjetni jezik, specijalno konstruiran za formuliranje osnovnog sadržaja dokumenta (informacije) i zahtjeva s ciljem njihovog usporedjivanja u postupku pretraživanja.

Idealni jezik trebao bi posjedovati slijedeće karakteristike (4,26):

- treba raspolagati s leksičko-gramatičkim sredstvima neophodnim za točno izražavanje centralne teme nekog dokumenta i informacijskog zahtjeva,
- ne smije biti dvosmislen: svaki zapis na njemu smije dopustiti jedno, i samo jedno, tumačenje,
- treba biti pogodan za algoritamsko usporedjivanje pretražnog tipa dokumenta s pretražnim uputstvom, izraženim u tom jeziku,
- treba biti što sličniji prirodnom jeziku (da ih indeksatori i korisnici što lakše shvate i nauče).

Jezici za indeksiranje različitih struktura i namjena predmet su izučavanja niza specijalista u svijetu. Teorija jezika za indeksiranje dopunjuje se stalno novim elementima koji omogućavaju postizavanje veće semantičke relevantnosti izdate informacije u praktičkoj izvedbi sistema za pretraživanje informacija.

U suvremenoj praksi pretraživanja informacija najrasprostranjeniji su jezici za indeksiranje deskriptorskog tipa. Za razliku od tradicionalnih jezika (hijerarhijske klasifikacije, jezici predmetnih odrednica, facetne klasifikacije) deskriptorski jezici imaju razvijene paradigmatičke i sintagmatičke relacije izmedju leksičkih jedinica svojih rječnika čija razvijenost i omogućava postizavanje veće ili manje semantičke relevantnosti. Suvremena istraživanja imaju akcent na pronalaženju i prikazivanju relacija medju leksičkim jedinicama jezika za indeksiranje, što je vrlo složeni zadatak, povezan s modeliranjem intelektualne čovjekove djelatnosti. Taj problem ni približno nije do kraja razriješen te se od znanosti na ovom području tek očekuju novi rezultati.

L I T E R A T U R A

1. *Avetisjan,D.O., Problemi informacionnogo poiska, Finansji i statistika, Moskva, 1981.*

2. *Engelbert,H., Informationsrecherchesysteme in der Wissenschaft, Akademie-Verlag, Berlin, 1978.*

3. *Doyle,L.B., Is relevance an adequate criterion in information system evaluation? ASC, Washington,1963.*

4. *Černij,A.I., Vvedenie v teoriju informacionnogo poiska, Nauka, Moskva, 1974.*

Primljeno: 1982-06-30

Žerjav F. Information Relevance and Criteria for the Efficiency Assessment of Information Retrieval Systems

S U M M A R Y

The concept of relevance plays a key role in the efficiency assessment of information retrieval systems. The paper analyses the problem of relevance in detail,defines its forms, discusses methods and attempts at objectivization,and points to future trends in the development of research in this field.

Criteria for the efficiency assessment of information retrieval systems in which relevance is of significance are presented.