[2] BURTON, R. E. – KEBLER, R. E.: The „half life" of some scientific and technical literatures = American Documentation, 71. köt. 1. sz. 1960. p. 18–22.
[3] LINE, M.B.: The „half life" of periodical literature: apparent and real obsolescence = Journal of Documentation, 26. köt. 1. sz. 1970. p. 46–52.
[4] BROOKES, B. C.: Obsolescence of special library periodicals: sampling errors and utility contours = Journal of the American Society for Information Science, 21. köt. 5. sz. 1970. p. 320–329.
[5] POPILOV, L. JA.: K voproszu o kriterijah kacsesztvennoj i kolicsesztvennoj ocenki informacionnüh potokov i trudoemkoszti obrabotki isztocsnikov informacii = Naucsno-Tehnicseszkaja Informacija, 2. sor. 3. sz. 1968. p. 3–6.
[6] DOBROV, G. M.: Nauka o nauke. Kiev, Naukova Dumka, 1966. 270 p.
[7] MIHAJLOV, A. I. – CSERNÜJ, A. I. – GILJAREVSZKIJ, R. SZ.: Osznovü informatiki. Moszkva, Nauka, 1968. 756 p.
[8] BRUKSZ, B. SZ.: Sztarenie naucsnoj literaturü. Problemü informatiki. Moszkva, VINITI, 1973. p. 74–102.
[9] VICKERY, B. C.: Editorial note = Journal of Documentation, 26. köt. 1. sz. 1970. p. 53–54.
[10] BROOKES, B. C.: The growth, utility, and obsolescence of scientific periodical literature = Journal of Documentation, 26. köt. 4. sz. 1970. p. 283–294.
[11] VJAL'JAOTSZ: Vozrasztnoj aszpekt ocenki fondov naucsnoj literaturü = Knygotyra, 11. köt. 4. sz. 1974.
[12] KOZACSKOV, L. SZ.: Szisztemü potokov naucsnoj informacii. Kiev, Naukova Dumka, 1973. 197 p.
[13] JUNISZISZT. Doklad ob iszszledovanii vozmozsnosztej szozdanija vszemirnoj szisztemü naucsnoj informacii. Moszkva, VINITI, 1972. 192 p.
[14] Perszpektivü razvitija fondov naucsnüh bibliotek. Szbornik naucsnüh trudov, 15. sor. Moszkva, Goszudarsztvennaja Biblioteka SZSZSZR im. V. I. Lenina, 1974. 82 p.
[15] IVANOV, R. N.: Metodü analiza v upravlenii processzami naucsno-tehnicseszkoj informacii. Moszkva, Sztatisztika, 1973. 112 p.
[16] KASAFUTDINOVA, E. Sz.: O racional'noj organizacii dokumental'nüh fondov = Naucsno-Tehnicseszkaja Informacija, 2. sor. 3. sz. 1975. p. 7–11.

/MOTÜLEV, V. M.: Ob opredelenii vremeni sztarenija dokumentov = Naucsno-Tehnicseszkaja Informacija, 2. sor. 12. sz. 1976. p. 3–7./

(Dezső László)

☆☆☆

A dokumentumok korának figyelembevétele a keresési folyamat során

Az információkeresés fő célja a felhasználó témájának megfelelő dokumentumok kikeresése. Ennek két fő oldala: a dokumentumokhoz valamely módszerrel tárgyszavak hozzárendelése (indexelés), és a felhasználó keresőképének, profiljának megszerkesztése. További, látszólag ezektől független szempontként felmerülhet az irodalom avulása, ami azt jelenti, hogy a felhasználók megkötéseket tesznek a dokumentumok korára vonatkozólag.

Az itt ismertetett módszer lehetőséget nyújt a dokumentumok tartalmi információinak és korának együttes kezelésére, aminek alapján hatékonyabb visszakeresési algoritmusok alkothatók meg. Ennek alapja olyan matematikai modell, amelyben *két változó* szerepel: egyrészt *a dokumentum és a keresőkép index-kifejezéseinek nyelvészeti azonossága,* másrészt *a dokumentum kora a kérdés feltevésének időpontjában.*

A nyelvészeti azonosság és a dokumentum korának közös kezelése

Az információkereső gyakorlatban általában nem szorítkoznak kizárólag az index-kifejezések egyeztetésével kapott találatokra. Más kritériumokat is szoktak megszabni a relevanciára, így *élnek a dokumentum típusára, nyelvére, publikálásának időpontjára stb. vonatkozó megszorításokkal* is, amelyek mind az output volumenét csökkentik.

Másfajta keresőstratégia is alkalmazható bizonyos keresőrendszereknél. Ennek lényege *az adatbázis összes dokumentumának rangsorolása* egy, valamennyi dokumentumhoz hozzárendelt szám szerint, amelyet a kérdező által tett lehetséges megkötések szerint két vagy több változó alapján határoznak meg. Ez lehetővé teszi a kikeresendő dokumentumok halmazának definiálását egyetlen küszöbérték szerint, szemben az előbbi módszerrel, ahol ehhez többkomponensű vektor szükséges.

A kétféle általános módszerrel kapcsolatban felvethető kérdések:

a két módszer egyenértékű-e, vagy pedig – adott lehívási vagy pontossági értékek mellett – az adatbázis különböző értékeléséhez vezet-e?

létezik-e egyértelmű kritérium annak eldöntésére, hogy meghatározott igények kielégítésére melyik módszert válasszuk?

hogyan lehetne olyan elméletet felállítani, amely gyakorlatban segíthet a két módszer közötti választásban?

A fenti kérdésekre adandó válaszokat a bevezetőben említett két változó összevetésével lehet megfogalmazni: *az index- és a keresőkifejezések nyelvészeti azonossága és a dokumentum kora jelenti a két változót.* E két változó két súlyozó tényezőt rendel minden dokumentumhoz, a tényezők kombinációjával egyetlen súlyérték képezhető.

Az irodalom avulása

Az irodalom avultságának *kettős értelme lehet:*

a) a dokumentumban foglalt információk értéke csökken a dokumentum korának növekedésével;

b) a dokumentum felhasználója úgy viselkedik, mintha a) igaz volna.

A következőkben feltételezzük, hogy objektív értelemben a)-nak nincs jelentősége, az avultság objektív meghatározásának alapjául a b) kiegészített változatát tekintjük:

b') A dokumentumok felhasználói az általuk relevánsnak minősített dokumentumok halmazának korára szubjektív módon egy valószínűségi eloszlást definiálnak. *Avultságról akkor beszélhetünk,* ha ennek az eloszlásnak az átlaga jelentősen eltér azon dokumentumhalmaz kormegoszlásának átlagától, amelyből a releváns dokumentumok ki lettek választva.

Az avultságnak *előjele* is van, attól függően, hogy az átlagok milyen irányban térnek el. Az avultság ilyen meghatározásához szükséges az összes dokumentum halmazának definiálása a kérdés feltevésének időpontjában.

Az információkeresésben az avultság tehát dinamikusan meghatározandó adat, formálisan valószínűségi eloszlásokkal kifejezve, amely felhasználók és kérdések szerint változik.

Az ismertetett modell egyik alapfeltevése, hogy *a releváns dokumentumok átlagos korának becslésére a felhasználó illetékes.* Ezt a feltevést különböző megfontolások igazolják, többek között az, hogy az információs szolgáltatások maguk is élnek a dokumentum kora, mint a relevancia egyik mutatója figyelembevételének gyakorlatával az SDI műveletek esetében.

Bonyodalmat jelent persze az, ha a felhasználó helytelenül becsüli meg a releváns dokumentumok átlagos korát.

A matematikai elmélet

A további tárgyalásban feltételezzük, hogy egy gyűjtemény minden dokumentumához két számot rendelünk a kikeresési folyamat során. Az egyik *a hagyományos vagy nyelvészeti súlyszám* ($\underline{x}$), amely a dokumentumhoz hozzárendelt szavak vagy szavak kódjainak halmaza és a felhasználó keresőkérdésében szereplő szavak halmaza egyeztetésére jellemző tényező. A másik szám *a dokumentum korára* jellemző ($\underline{t}$). A következőkben az $\underline{i}$ alsó index a relevanciát mutatja: $\underline{i} = 1$ a *releváns* dokumentumokat, $\underline{i} = 2$ a *nem-releváns* dokumentumokat jellemzi.

Felírhatjuk a két változó sűrűségfüggvényét, ezek legyenek rendre $f_i(\underline{x})$ és $g_i(\underline{t})$. Kimutatható, hogy $f_i(\underline{x})$ normál valószínűségi eloszlással, $g_i(\underline{t})$ negatív exponenciális eloszlással közelíthető. Az együttes kétváltozós valószínűségi sűrűségfüggvény így írható fel:

$$h_i(\underline{x}, \underline{t}) = f_i(\underline{x}) \cdot g_i(\underline{t}) \qquad (i = 1, 2).$$

E kétváltozós függvény értékei egy felületen helyezkednek el. Minden dokumentumnak az alapsíkban egy $(\underline{x},\underline{t})$ pont felel meg, ahol $\underline{x}$ a *nyelvészeti súlyszámot,* $\underline{t}$ a *dokumentum korát* jelzi. A felület minden pontjának egy $h(\underline{x}, \underline{t})$ érték felel meg, amely az illető dokumentum *relevanciájának valószínűségi sűrűségére jellemző.*

A dokumentum kikeresésére elméletileg kétféle eljárás áll rendelkezésre:

1. *A részhalmazok módszere.* Keressünk ki minden olyan dokumentumot, amelyhez rendelt $\underline{x}$ érték meghalad egy $\underline{x}_c$ küszöböt. Ebből a részhalmazból válogassuk ki azokat a dokumentumokat, amelyek $\underline{t}$ értéke kisebb egy $\underline{t}_c$ küszöbnél.

A két lépés sorrendje meg is fordítható, mivel a két változó független.

2. *A kétváltozós súlyozási módszerek.* Rendeljünk hozzá minden dokumentumhoz egy kétváltozós súlyfüggvényt: $z = z(\underline{x},\underline{t})$, vagyis egyetlen $\underline{z}$ számot, amely $\underline{x}$-nek és $\underline{t}$-nek függvénye. Szorítkozzunk ama dokumentumok kikeresésére, amelyekre nézve $z > z_c$ teljesül. A $\underline{z}$ megválasztható lineáris alakban: $z = k_1 x + k_2 t$, ahol k_1 és k_2 állandó.

Célunk a két módszerrel elérhető kikeresési hatékonyság összehasonlítása azzal a megszorítással, hogy a 2. módszerben a $\underline{z}$ függvény lineáris.

Hosszas matematikai levezetésekkel meghatározhatók a keresés különféle paramétereit (lehívási arány, pontosság stb.) leíró képletek, amelyekben az $\underline{x}_c$ és $\underline{t}_c$ küszöbértékek szerepelnek, mindkét módszerre külön-külön.

A kétváltozós súlyozási módszerre a levezetések eredményeképpen adódik a lehívás-pontosság összefüggés, amelyet grafikusan is ábrázolnak, tetszőleges $\underline{k}_1$ és $\underline{k}_2$ együtthatókkal. Felmerül a kérdés, hogy melyek e két együttható optimális értékei. Legyen $\underline{k}_2/\underline{k}_1 = \underline{k}$, akkor a $z = x + k_t$ súlyképletben szereplő együttható optimális értéke a lineáris diszkrimináns analízissel határozható meg. Az eljárás eredményeképpen adódó képletekben a $\underline{k}$ együttható értéke $\underline{x}$ és $\underline{t}$ várható értékének átlagával és varianciájával, valamint 2 és 0 közötti zárt intervallumban levő értékeket felvevő konstansokkal lesz kifejezve. A lineáris diszkrimináns analízis előnye, hogy a $\underline{k}$ optimuma igen egyszerűen meghatározható: akkor lesz $\underline{k}$ optimális, ha a képletben szereplő konstans értéke éppen 1.

Következtetések

A két módszer részletes gyakorlati elemzése alapján összehasonlításuk eredményeit a következőkben foglalhatjuk össze:

a) mindkét módszer *javíthatja az információkeresés hatékonyságát,* de ronthat is, ha nem következetesen alkalmazzuk;

b) *a pontosság-lehívás görbe* alakját adott kérdés esetén, két szélső helyzet között, egyrészt a részhalmazok módszere, másrészt a kétváltozós súlyozási módszer határozza meg;

c) a részhalmazok módszere a lehívási értékeket, míg a kétváltozós súlyozási módszer a pontossági értékek

k-tól függő intervallumát *nem teszi hozzáférhetővé* a felhasználók számára.

d) lineáris diszkrimináns analízissel a kétváltozós súlyozási módszerrel meghatározott k csaknem *optimális*, azonban a különböző paraméterek becsléséből adódó hibák csökkentése csak számítógépes szimulációs eljárással valósítható meg.

A közölt eljárással a dokumentumok kikeresése csak a nyelvészeti súlyszám és a kor alapján történik. Természetesen más változókat is figyelembe lehet venni, így pl. *a dokumentum nyelvét, a dokumentum típusát* stb. Számos változó ezek közül azonban úgy tekinthető, mint valamilyen nyelvészeti súlyszám. Mivel a keresésben felhasználható adatok kiválasztásának alapelve a változók függetlensége, a sokváltozós súlyozási módszer különösen olyan gyűjteményekből való visszakeresésre alkalmas, amelyeket *a fazettás osztályozási elv* szerint indexeltek.

A visszakeresés hatékonyságának javítása, amely egy összekapcsolt nyelvészeti/avultsági közelítésből adódik, függ az adatbázis hozzáférhető részének időbeli lefedésétől, amire egy A mennyiség jellemző (az adatbázisba még felvett dokumentum maximális kora). Pozitív előjelű avultság esetén bármelyik módszert is alkalmazzák a keresőrendszer javítására, a hatékonyság A függvénye lesz. Az A érték rögzítése a működő információkereső rendszerekben mindenképpen arra utal, hogy már eleve figyelembe veszik a dokumentum korát, mint a relevancia egyik mutatóját. Ezt azonban általában nem tudományos módszerességgel szokták meghatározni, hanem inkább ad hoc. Az A értékének optimális meghatározása, vagyis egy adatbázis optimális particionálása a dokumentumok kora szerint (vagy más változók szerint is) a vezetői döntések körébe tartozik, amelyhez az itt ismertetett módszer segítséget nyújthat.

/HEINE, M. H.: Incorporation of the age of a document into the retrieval process = Information Processing and Management, 13. köt. 1. sz. 1977. p. 35–47./

(Roboz Péter)

INFORMÁCIÓS HÁLÓZATOK, TÁJÉKOZTATÁSI SZOLGÁLTATÁSOK

A nyugat-európai tudományos, műszaki, gazdasági és társadalomtudományi információs hálózat, az EURONET

Az információs technika fejlődésének egyes szakaszai az utóbbi 15 évben a következőkkel jellemezhetők:

tezauruszok kialakulása;
on-line párbeszédes rendszerek kialakulása;
információs hálózatok kialakulása.

Pontosan ezt a sémát követte az információs tevékenység fejlődése az *Európai Közösségben* is.

1961-től 1965-ig tartott az EURATOM tezauruszának felépítése, amely az INIS *tezaurusszá* nőtte ki magát. 1967-ben kezdődött a gépi információkeresés először off-line, majd *on-line* üzemmódban. 1971-ben határozták el a tudományos és műszaki információs és dokumentációs tevékenység tagállamok közötti *koordinálását*. Az erre vonatkozó határozat szerint:

a) koordinált információs rendszert kell kifejleszteni a nyugat-európai információs hálózat létesítésének előfeltételeként;
b) meg kell teremteni a hálózat egységes üzeméhez szükséges módszereket és szabványokat;
c) gondoskodni kell az információs szakemberek és a felhasználók oktatásáról;
d) fejleszteni kell az információs technikát.

A határozat leszögezi, hogy az információellátást a legkorszerűbb módszerekkel kell biztosítani, a legkedvezőbb feltételek mellett, vagyis lehetőleg gyorsan és kis költség-ráfordítással.

A határozat végrehajtásáért az *Európai Közösségek Bizottsága* felelős. Létrehozták a megfelelő politika kialakításáért és gyakorlati megvalósításáért felelős különbizottságot is, amelynek elnevezése: *Tudományos és Műszaki Információs és Dokumentációs Bizottság (Committee for Information and Documentation in Science and Technology, CIDST)*.

Ez a bizottság az 1975–1977. évi időszakra hároméves *intézkedési tervet* dolgozott ki. Az ennek megfelelő három fő tevékenységi terület:

információs rendszerek létesítése különféle szakterületeken (pl. mezőgazdaság);
az információkezelő hálózat fizikai megvalósítása;
információs technikai eszközök és módszerek kifejlesztése.

1975-ben szerződés létesült a 9 tagállam postaigazgatóságainak konzorciumával, amely kialakítja és üzemelteti az adatátviteli hálózatot.

Ezt követően a tagállamok megvizsgálták, hogy *mely adatbázisok vesznek részt a hálózatban*, és milyen anyagi támogatásra lesz szükség. Megállapodtak abban, hogy a rendelkezésre álló pénzügyi eszközök figyelembevételével 1977 végére 95 adatbázis és adatbank bekapcsolódása látszik megoldhatónak.

1. Alapvető tervtanulmány

Az intézkedési terv alapját képező tanulmány a tudományos és műszaki információk várható felhasználásának volumenével, eloszlásával és alakulásával foglalko-