deszkriptorokra alapozott keresés. Ennek oka, hogy az előbbi esetekben egyaránt keresünk a cím, a referátum és a deszkriptorok adatmezőjében előforduló tárgyszavak szerint, az utóbbi esetben viszont csak a deszkriptorok szerint. Ezt bizonyítja a *2. táblázat*, ahol a repülőgép által okozott zaj (AIRCRAFT NOISE) kereső-kifejezést különféleképpen keressük egy adatbázisban.

2. táblázat

Az AIRCRAFT NOISE szakkifejezés keresése különféle módszerekkel

A kérdés sorszáma	A találatok sorszáma	Kereső-kifejezés	Magyarázat
1	2250	AIRCRAFT NOISE	csak szabályozott deszkriptor
2	3662	AIRCRAFT(W)NOISE	a kifejezés bármely adatmezőben előfordulhat, de csak egymás mellett, ilyen sorrendben
3	4979	AIRCRAFT(F)NOISE	a tárgyszavak ugyancsak bármely adatmezőben előfordulhatnak, de nem kell egymás mellett állniok
4	5148	AIRCRAFT(C)NOISE	a tárgyszavaknak nem kell ugyanabban az adatmezőben sem lenniük
5	55 906	AIRCRAFT	logikai kombináció, a két tárgyszó bárhol előfordulhat a rekordban, az eredmény ugyanaz, mint a 4. számú kérdésre
6	30 074	NOISE	
7	5148	5 ÉS 6	

Látható, hogy mindegyik kontextuális logikai módszerrel feltett kérdés (2., 3. és 4. kérdés) tartalmazza az 1. kérdésben hivatkozott deszkriptort is. Valószínű, hogy a leginkább releváns dokumentumokat az 1. kérdés szolgáltatja, mert az indexelők ezt a deszkriptort nyilván azért rendelték hozzá a dokumentumhoz, hogy tükrözze annak fő mondanivalóját. Ilyenkor persze elveszítjük azokat a dokumentumokat, amelyek az AIRCRAFT NOISE szakkifejezést a címben vagy a referátumban tartalmazzák, de ilyen deszkriptoruk nincs. Ezeket (a deszkriptoros tételekkel együtt) a 2. kérdés „hozza ki". Ha viszont a 3. és 4. kérdés kontextuális közelítését vagy az 5–6–7. kérdések egyszerű szavas, kombinációs közelítését alkalmazzuk, hamis válaszokat is kaphatunk, mert az AIRCRAFT és a NOISE nem feltétlenül egymáshoz rendelt szavak lehetnek.

A keresési stratégiát bizonyos fokig befolyásolja az indexelés. Ha az információs tételeket egyáltalán indexelték, azaz intellektuális munkával valamilyen tárgyszavakkal látták el, akkor a keresőnek az indexelő fejével kell gondoskodnia: ki kell találnia, hogyan indexelhették az általa keresett dokumentumokat. Ha ez nem vezet eredményre, akkor a témát olyan kérdésekkel kell megközelítenie, amelyek segítenek neki a releváns keresőszavak kikutatásában.

A szabályozott és szabad tárgyszavas online keresési módszerek közül a keresőnek kell választania, az előnyök és hátrányok alapos ismeretében. Döntésében olyan tényezők is szerepet játszanak, mint az online keresésre fordítható, rendszerint korlátozott idő, a kívánt válaszok száma, a keresés pontossága vagy teljessége.

/RAITT, D. I.: Aspects of searching via on-line systems using controlled and uncontrolled vocabularies. = IATUL Proceedings, Online issue, 12. köt. 1980. p. 3–21./

(Roboz Péter)

Egy újabb stratégia az információkeresésben

Bevezetés

A működő információkereső rendszerek (Information Retrieval System, IRS) száma a százat is meghaladja; közös céljuk: ellátni az olvasókat igényeik szerinti dokumentumokkal. Ami megkülönbözteti ezeket az IRS-eket az adatkereső rendszerektől (Data Retrieval System, DRS), amilyenek pl. a banküguyviteli vagy légiforgalmi helyfoglaló rendszerek, az a *relevancia*. A DRS-ek esetében egy kérdésre kapott válasz relevanciája *objektív módon dönthető el*, míg az IRS-ekben a feltett kérdésre kapott válasz relevanciáját voltaképpen csak maga *a felhasználó tudja megítélni*. A DRS-eket – e szempontból – az IRS-ek speciális eseteként tekinthetjük. A relevancia az a tényező, amely meghatározza az IRS-ek hatékonyságát.

A működő IRS-ek, amelyek közül egyesek a nagy, online hozzáférésű adatbázisok keresését is lehetővé teszik, azonos alapelven működnek: a felhasználó kérdését deszkriptorok és Boole-operátorok segítségével keresőprofillá alakítják át, és összehasonlítják a fájlban tárolt dokumentumokhoz hozzárendelt deszkriptorokkal. A fájl minden dokumentuma egyenként átesik ezen a

műveleten, a dokumentumok deszkriptorai és a keresőprofil stratégiájának azonossága esetén „találatot" szolgáltat a rendszer. Ez azonban nem mindig nyújt a felhasználó számára is releváns információkat.

A Boole-módszerű keresési stratégia már régóta kritika tárgyát képezi. [1, 2] Megemlíthetjük ezzel összefüggésben a MARON és KUHNS által ajánlott „valószínűségi indexelési" rendszert [3], amelynek lényege, hogy a dokumentumokhoz rendelt indexelő kifejezések mindegyikéhez egy számot is hozzárendel. E szám kifejezi annak mértékét, hogy a kiválasztott indexkifejezés mennyire írja le a tartalmat. DOYLE [4] az egy fájlban elhelyezett dokumentumok egymással való összefüggését tartja a visszakeresésre alkalmas tényezőnek: asszociációs térképet javasol, és az olvasónak e térkép alapján kell megtalálnia a számára releváns információkat.

SALTON munkássága az információs rendszerek kialakításának legkülönbözőbb területeihez fűződik. Munkái közül főleg az automatikus osztályozás és a hierarchikus fájl-szervezés kapcsolódik szorosan e cikk mondanivalójához. [5]

Salton módszerei sok hasonlóságot mutatnak az alábbiakban ismertetendő módszerrel, a különbség a relevancia fogalmának szemléletmódjában van.

GOFFMANN *indirekt kereső módszere* (Indirect Method, IDM) [6] tíz évvel ezelőtt látott napvilágot, ez hatékonyabb a Boole-logikán alapuló módszereknél, de Maron és Kuhns módszerénél is. Az IDM egy-egy dokumentum relevanciáját az adott keresőkérdéssel operáló műveletben *már megtalált egyéb dokumentumokkal való összehasonlítás révén állapítja meg.* Több kísérlet bizonyítja az IDM hatékonyságát. [6, 7, 8, 9]

Azt, hogy az IDM-et a nagy információkereső rendszerek mégsem alkalmazzák, igen költséges voltának tulajdoníthatjuk. Az alábbiakban tárgyalt keresési stratégia ezen kíván javítani. Esetében az IDM-nek egy olyan változatáról lesz szó, amely az eredetinél rugalmasabb keresési lehetőségeket nyújt, a költségei mégsem nagyobbak, sőt a Boole-logikán alapuló keresési rendszerek költségeivel is összemérhetők. Az alábbi ismertetés csak a keresési stratégiával foglalkozik, bár ez nem független az IRS egyéb elemeitől.

A keresőkérdés

Mielőtt a keresési stratégiát elemeznénk, definiálnunk kell azon kérdések lehetséges típusait, amelyeket a felhasználó feltesz az információkereső rendszernek.

A kérdéseket két rendező elv alapján osztályozhatjuk:

a válasz típusa szerint, amit a kérdező vár, vagyis a kérdés célja szerint (*A* típus);

a kérdés megfogalmazásának módja szerint (*B* típus).

A típusú kérdések:

a) *standard kérdés:* a felhasználó azoknak a tételeknek egy korlátozott nagyságú és rendezett jegyzékét várja, amelyek felelnek a kérdésre. Sokszor 500 rendezetlen tételből kellene a valóban releváns, mondjuk 10 darabot kiválasztani;

b) *nem-standard kérdés:* a felhasználó nagy mennyiségű választ vár, mindent, ami kérdésére felel, tehát „teljes" bibliográfiát. Ez esetben nincs szükség finom keresési stratégiára vagy szűrésre.

B típusú kérdések:

a) *meghatározott válaszból kiinduló kérdés:* a felhasználó kérdésként egy igényeit kielégítő és a fájlban benne levő dokumentumot jelöl meg, amely reprezentánsa kérdésének és a keresés kiindulópontjaként használható;

b) *Boole-megfogalmazású kérdés:* a felhasználó úgy fogalmazza meg kérdését, hogy ismeri a szabályokat, amelyekkel az indexkifejezések Boole-operátorok segítségével kérdéssé állíthatók össze;

c) *keresgélő kérdés:* ez esetben a felhasználó a fenti módok egyikén sem tudja igényét pontosan megfogalmazni, nem nagyon tudja maga sem, mire van szüksége. Áttekinti az egész adattárat, megtalál benne néhány releváns dokumentumot, és ennek eredményeképpen már tud talán pontos Boole-típusú kérdést fogalmazni. A módszer sokkal hatékonyabb, ha a keresgélést a fájl egy kis részére tudja korlátozni.

A keresési stratégia – feltételek és követelmények

A keresési stratégia valójában egy szabályrendszer, amely lehetővé teszi, hogy a kérdéssel összevessük a fájlban levő információs tételeket, és kiválasszuk közülük a kérdésre megfelelő választ adó néhányat. A jó stratégia csak a releváns tételeket emeli ki, méghozzá valamennyit. Minthogy a relevancia fogalma eléggé meghatározhatatlan, ez a kívánság túl általános. Kíséreljük meg a „jó" válasz ismérveit és előfeltételeit pontosabban körülírni.

– Ajánlatos, hogy a válasz rendezett legyen, azaz a tételek olyan sorrendben kövessék egymást, ahogyan olvasni szeretnénk őket.

– A felhasználó hozzávetőlegesen határozza meg a várt válasz nagyságát.

– A stratégia vegye figyelembe a „menet közbeni" változtatást, vagyis egy-egy tétel felhasználói relevanciájának mértékét egy később felbukkanó másik minden további nélkül megváltoztathatja.

– Az interaktivitás a felhasználó és a rendszer között legyen megvalósítható, mert ez erősen javítja a rendszer hatékonyságát.

A relevancia időtől is függő, szubjektív fogalom. A kereséskor nem ismerjük jól sem a felhasználót, sem problémáját. A rendszer legfeljebb arra lehet képes, hogy olyan dokumentumokat nyújtson válaszként, amelyeknek jó esélyük van arra, hogy relevánsak legyenek a kérdésre. A végső ítéletet ebben azonban csak maga a felhasználó mondhatja ki. Így tulajdonképpen megtörténhet az is, hogy ugyanarra a kérdésre – ha különböző időpontokban tesszük fel – a rendszer ugyanazt a dokumentumegyüttest választja ki, de a felhasználó e dokumentumoknak más-más részhalmazát találja majd relevánsnak. Ugyanez előfordulhat azonos időpontban is, ha két felhasználó azonos kérdést tesz fel.

Az indirekt módszer

Goffman IDM-je általános kommunikációs modellt ad, [10] ebből származtatja keresési stratégiáját. Egy D adatbázis minden elempárja között a P_{ij} feltételes valószínűség jelentése: ha az i-edik tétel releváns egy adott kérdésre, akkor a j-edik is az.

Minden kérdéshez tartozik egy k_0 küszöbérték, amely elválasztja az „alig" relevánsat az irrelevánstól. Minden P_{ij} érték, amely kisebb vagy egyenlő e küszöbszámnál, nulla értékűnek tekintendő.

Az IDM szerint kétféle módon lehet a kérdéseket definiálni:

egy releváns dokumentummal – *B*a) típusú kérdés – és egy küszöbértékkel;

Boole-operátorokkal összekapcsolt indexkifejezések csoportjával – *B*b) típusú kérdés – és egy k_0 küszöbértékkel.

Ha a kérdés az első módon adott, akkor a visszakeresett dokumentumok sorozatát először a kiinduló dokumentummal való összevetés, majd sorban a következő kikeresett dokumentummal végzett összehasonlítás adja, és így tovább, amíg nem marad releváns tétel, vagyis mindaddig, amíg a $P_{ij} > k_0$ feltétel teljesül. A fájl dokumentumaiból csökkenő P_{ij} értékeink alapján rendezett sorban kapjuk a választ alkotó dokumentumokat.

A másik módon megfogalmazott kérdés esetén a stratégia kapcsolódó tételek különálló (diszjunkt) osztályaira (ekvivalencia-osztályokra) tagolja az adatbázist, k_0 szerint. Minden osztályból kiválaszt egy tételt, és összeveti a kérdéssel. Az így kapott legmegfelelőbb tétel elvezet egy ekvivalencia-osztályhoz, amelynek minden tagján végigfuttatják a keresőkérdést. A kérdéshez legjobban illeszkedő dokumentum lesz az előző típusú keresési stratégia kiinduló tagja.

CLEVELAND [7] ezt a stratégiát *Geometrical Model (GM)* néven terjesztette ki; ez a fentiekhez hasonló módon, de több lépcsőben – kevésbé szigorú feltétellel – válogat. A GM – az IDM-hez hasonlóan – nagyon jó hatásfokú keresést biztosít, de használata igen költséges.

A javasolt lánc-eljárás

Az IDM és a GM eljárás egy k_0 küszöbszám és egy releváns Q tétel meghatározása után mindig az előzőleg kiválasztott tételhez való hasonlóság alapján választja ki a következő releváns tételt. A relevancia pontos megítélése a felhasználó korábbi ismereteitől függ. Közelebb juthatunk azonban a felhasználó kívánságának teljesítéséhez, ha a relevánsnak tudott Q kiválasztása után a Q-hoz hasonló J-t választunk ki. Ezután már nemcsak a J-hez hasonló harmadik releváns dokumentumot választjuk ki, hanem mindkettőhöz, tehát a Q-hoz *és* a J-hez is hasonló harmadikat, majd mindhármukhoz hasonló negyedik stb. dokumentumot. A keresés úgy alakítható, hogy a legutóbb talált dokumentumhoz való hasonlóság nagyobb súllyal szerepeljen, mint a megelőzőkhöz való. Ily módon rendezhetjük is a kiválasztott dokumentumok sorát.

A javasolt keresési stratégia

Mindkét fentebb említett stratégia, az IDM és a GM is a küszöbérték megszabását egyaránt a felhasználótól várja. Rendszerint azonban a felhasználónak nincs tapasztalata ennek meghatározásában, legfeljebb a keresés lefolytatása után tudna erről nyilatkozni. Hosszú és költséges módszer lenne egy találomra felvett k_0 értékkel elindítani a keresést, majd az eredményből egy javított k_0-hoz jutni, és több iterációs lépés után megkapni a felhasználó számára kielégítő választ.

Az IDM és GM módszerek keresési stratégiájának módosítása az alábbi eljárás.

A *Keresőkérdés* c. fejezet jelöléseivel élve:

Ba) típusú kérdés

Adott egy Q kiinduló dokumentum. A javasolt stratégia kikeresi a Q-hoz leginkább hasonló két dokumentumot. Második lépcsőben mindkettőjük alapján két-két további tétel választódik ki és így tovább. Eredményül egy fa-szerkezet rajzolódik ki, ezt választérképnek hívják *(1. ábra)*.

A felhasználónak minden elágazásnál két dokumentum között kell választania, így releváns dokumentumok láncát kapja (pl. a szaggatott vonallal jelzett A út). Megtörténhet, hogy egy tétel többször is szerepel a választérképen, csak nem ugyanazon az útvonalon. A módszer segítségével nemcsak két tétel válogatható ki egy lépés során, hanem tetszőleges n számú is.

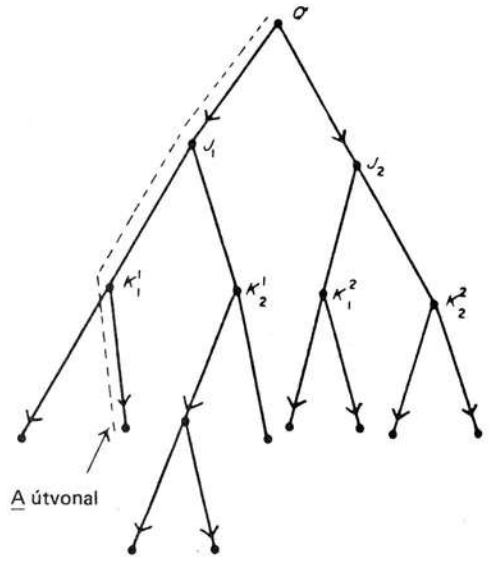
Bb) típusú kérdés

Nem valószínű, hogy sok Boole-típusú kérdés érkezik a rendszerhez, mivel az ilyen keresőkérdések megfogalmazása profilszerkesztőt kíván. Ilyen megfogalmazású

kérdést gyakran helyettesíthetünk egy megfelelő dokumentummal. Az IDM algoritmusa jó megoldást nyújt a Boole-típusú keresőkérdésekre is. Az adattárat célszerű 20–30 tételt tartlmazó független csoportokra bontani, és minden csoportból egy reprezentánts tételt kiemelni, amely lehetőleg az egész csoportot képviseli. A kérdést az IDM-nél leírt módon e reprezentánsokkal hasonlítja össze a keresési stratégia.

Bc) típusú kérdés

A *keresgélő kérdés* esetében az egész adattár átvizsgálása idő- és költségtényezők miatt lehetetlen. Meg kell elégedni ez esetben is a fájl csoportokra osztásával és e csoportok reprezentáns képviselője vizsgálatának stratégiájával.



1. ábra Választérkép

A javasolt stratégia elemzése

Az információkeresés e módszere iterációs lépéseken át közelít a megoldáshoz, ellentétben az olyan módszerrel, amely „teljesen releváns" válaszokat nyújt. A relevanciát végső soron a felhasználó ítéli meg, speciális érdeklődésének, előzetes ismereteinek függvényében. A felhasználó és a rendszer közötti interaktív kapcsolatra szükség van, de az értékelés történhet a felhasználó munkahelyén, a választérkép egy kinyomtatott példányának segítségével. Ez olcsóbbá teszi a rendszer működését.

A válaszul kapott tételek mennyisége igen gyorsan növekszik a keresési lépések számával. A standard kérdésekre 8, 10 vagy 12 lépésben javasoljuk a közelítést, az optimális terjedelmű válasz előállításához.

Ha a választérkép a sok lehetséges útvonallal elkészül, és a felhasználó kiválaszt egy utat, máris tudhatjuk, hogy érdeklődése milyen típusú dokumentumokat tüntet ki. Például, ha a sokszor idézett dokumentumokat választja a kevésbé idézettekkel szemben, az útvonal folytatását már ezt figyelembe véve jelölhetjük ki.

A megvalósítás gazdaságossága

A javasolt módszer költségeit a Boole-logikán alapuló módszer költségeivel lehet összevetni.

- A mátrix nagysága

  A P_{ij} mátrix nagysága a fájl nagyságától függ; S darab dokumentum esetén $S \times S$ nagyságú. A módszer ismeretében világos, hogy ennek csak a felére van szükség.

- A mátrix szükségessége

  A mátrixot a keresés során nem kell online hozzáférésűvé tenni. Elegendő az egyik tételtől a másikhoz vezető mutató-értékeket egyszer kiszámítani, és így a mátrix maga off-line tárolható. A *B*a), *A*a) típusú kérdések esetében egyáltalán nem szükséges a mátrix a visszakereséshez.

- A mátrix naprakészen tartása

  Úgy tűnik, hogy a mátrixot újra ki kell számítani a fájl minden új tétellel való kiegészítésekor. Ezt kerüli meg az a módszer, amely az új tételeket – átmenetileg, a Boole-típusú kérdéseknél leírtak szerint – egy reprezentánssal helyettesíti. A teljes mátrix átdolgozását nem tételenként, hanem nagyobb időközökben végzik.

A költségek összetétele

A teljes költség két összetevőből áll: az állandó, nem a kérdések számától, hanem az adatbázis nagyságától függő költségekből és a változó, a kérdések számától függő kiadásokból. Mindkét költségtípusnak több eleme van.

Állandó költségek:

a) beszerzés
b) feldolgozás (indexelés, lyukasztás stb.)
c) az adatbázis kiegészítése új tételekkel
d) a tételek tárolása
e) a kereséshez szükséges adatok tárolása
f) az adatbázison belüli mutatószámok kiszámítása.

Változó költségek:

g) a kérdés megszerkesztése
h) a keresés művelete.

Összehasonlítások:

Látható, hogy az a), b), c) és d) alatti tevékenységek költségei az itt javasolt és az IDM módszer esetében nagyjából azonosak. A g) alatti költségek a hagyományos rendszerekben nagyobbak, a felhasználó és a gép közötti közvetítő személy szükségessége miatt. A h) költségek lényegesen magasabbak a hagyományos információkereső rendszerekben és a fájl növekedésével tovább nőnek.

Az e) alatti költségek részletezése:

A javasolt keresőrendszerben minden tételhez tárolni kell a más tételekhez való hasonlósági kapcsolat mutatószámait. A pontos költségszámítás helyett megbecsüljük a tárolási kapacitás növelésének szükségességét a d)-hez képest.

A hagyományos keresőrendszernek minden dokumentum esetén a következő adatokat kell online módon tárolnia:

- szerzők neve,
- a dokumentum címe,
- folyóirat címe,
- kötet, füzet, oldal,
- 8–15 deszkriptor,
- a deszkriptorok és dokumentumszámok invertált fájlai.

Sokszor még referátum is társul a fentiekhez. Az első öt csoport tételenként mintegy 2000 bit kapacitást igényel. A javasolt rendszerben, 1 024 000 tételes fájl esetén, minden mutató tárolásához 20 bit kell. Ez a szám átlagosan 12-re csökken, tekintettel a hasonló dokumentumok egy csoporthoz tartozására. Így a tárolókapacitás mintegy 6–18%-kal növelendő, ami nem jelent túl nagy költségtöbbletet, különösen, ha a nagyon gyors keresést tekintjük. A javasolt információkereső rendszert gazdaságosnak tekinthetjük, mert bár állandó költségei magasabbak, változó költségei alacsonyabbak a hagyományosnál. Az állandónak nevezett költségek természetesen csak egy bizonyos kérdésmennyiségig tekinthetők fixnek; ha a rendszer fizikai kapacitása eléri határát, és új tárolókat kell igénybe venni, ugrásszerűen nőnek a költségek.

Befejezés

A leírt keresési stratégiát az IDM rugalmas kiterjesztésének tekinthetjük; az IDM ennek a módszernek speciális esete. Legegyszerűbb változata helyettesítheti a hagyományos Boole-módszert alkalmazó rendszereket, hatásfoka jóval nagyobb, költségei pedig nagyságrendileg összemérhetők.

Irodalomjegyzék

1. VERHOEFF, F. – GOFFMANN, E. – BELZER, J.: Inefficiency of the use of Boolean functions for information retrieval = Communications of the ACM, 4. köt. 1961. p. 557–559.
2. VAN RIJSBERGEN, C. J.: Information retrieval. Butterworths, London, 1975.
3. MARON, M. E. – KUHNS, J. L.: On relevance, probabilistic indexing and information retrieval = Journal of the ACM, 7. köt. 3. sz. 1960.
4. DOYLE, L. B.: Semantic road maps for literature searches = Journal of the ACM, 8. köt. 1961. p. 574–578.
5. SALTON, A.: Automated information organization and retrieval. New York, McGraw-Hill, 1968.
6. GOFFMANN, W.: An indirect method of information retrieval = Information Storage Retrieval, 4. köt. 4. sz. 1968.
7. CLEVELAND, D. B.: An n-dimensional retrieval model = Journal of the American Society for Information Science, 27. köt. 5/5. sz. 1976.
8. DERINGER, D. K.: An information retrieval system for a computer center. Doktori disszertáció, Cleveland, CWRU, 1972.
9. CROFT, W. B. – VAN RIJSBERGEN, C. J.: An evaluation of Goffmann's indirect retrieval method = Information and Processing Management, 12. köt. 1976. p. 327–331.
10. GOFFMANN, W. – NEWILL, V. A.: Communication and epidemic processes = Proceedings of the Royal Society A, 298. köt. 1967. p. 316–334.

/MANSUR, O.: An associative search strategy for information retrieval. = Information Processing and Management, 16. köt. 3. sz. 1980. p. 129–137.

(Domokos Miklósné)