A szabad szöveges keresés és a szabályozott szótárat alkalmazó keresés viszonya

Amikor megjelentek az első olyan, számítógépes információszolgáltatás céljára szolgáló állományok, amelyeknek a rekordjai *szövegeket* tartalmaztak, az információkereséshez a rekordokat el kellett látni a szövegek tartalmát kifejező segédadatokkal. Vagy valamilyen osztályozó rendszer jelzeteit, vagy a természetes nyelv szavait, kifejezéseit lehetett használni erre a tartalmi feltárásra (azaz osztályozásra, indexelésre). A természetes nyelvi szavakat, illetve kifejezéseket egy-egy meghatározott készletből, meghatározott szabályok szerint kellett kiválasztani. Az e célra szolgáló szókészletek, az ún. *szabályozott* (kontrollált) szótárak összeállítása és alkalmazása egyaránt megkövetelte az adott szakterület átfogó ismeretét, a terület fogalmi rendszerének átlátását, az osztályozó terminusok (kifejezések) következetes alkalmazását.

Amikor megjelentek a nagy kapacitású és invertált fájlokat kezelő számítógépek, lehetővé vált a rekordok visszakeresése a bennük lévő szöveg *minden egyes szava* alapján is. A 60-as évek első felében kialakult, és évekig tartotta hadállásait az a vélemény, miszerint *a legjobb indexelő nyelv maga a természetes nyelv* – minimális szabályozással, vagy éppenséggel minden megkötés nélkül.

Időközben a nagy referáló szolgáltatásokat sorra gépesítették, és a 70-es évek folyamán mind természetesebbé vált, hogy online elérhető adatbázisokat készítsenek. A már korábban is létező tezauruszokat és más szabályozott szótárakat nem volt nehéz az online kereső rendszerekben alkalmazni, sőt újakat is készítettek, és ezeket a segédleteket a szabad szöveges kereséssel *együtt* lehetett használni. A tezaurusz mint eredményes keresési eszköz rendszeres kutatás tárgyává vált, alapelvei kikristályosodtak. *A keresések eredményességét a korábbitól eltérő megközelítéssel elemezték.* A 70-es évek végétől kezdve a mindennapi kereséseket tanulmányozva, "élesben" lehetett vizsgálni a különböző keresési technikákat. E vizsgálatok eredményeiből következtetéseket lehet levonni arról, hogy milyen előnyei és hátrányai vannak a szabályozott szótáras keresésnek és a szabad szavas keresésnek általában, illetve különböző környezeti feltételek mellett.

Ellentmondásos vizsgálati eredmények

A vizsgálatok megtartották a 60-as években alkalmazott paraméterekből a keresés eredményességének két mutatóját: a pontosságot és a teljességet.

$$\text{Pontosság} = \frac{\text{releváns találatok száma}}{\text{összes találat száma}},$$

$$\text{Teljesség} = \frac{\text{találatok száma}}{\text{az állományban lévő összes releváns tétel száma}}$$

– A ref.

Ellentétben azonban a szabad szöveges keresés egyértelmű előnyeit mutató korábbi eredményekkel, az elemzett (a 70-es évek végétől a 80-as évek közepéig tartó) időszakban változatos, sőt egymásnak ellentmondó eredményekre jutottak a kutatók.

- *Henzler* [1] úgy találta, hogy szabályozott szótárral általában nő a teljesség, és a pontosság is nagyobb; de nem akkor, ha valamilyen új dologról, új témáról van szó, ezért a szabad szöveges és a szabályozott szótáras keresés kombinálását ajánlotta ideális megoldásként.
- *Hersey és társai* [2] tapasztalatai szerint az osztályozó jelzetek (az eredetiben "subject codes", vagyis "témakódok", ami osztályozó jelzeteket és kódolt deszkriptorokat egyaránt jelenthet – A szerk.) szerinti keresés a teljesség szempontjából 30–40%-kal, a pontosság szempontjából 15–20%-kal jobb eredményeket hozott, mint a szabad szöveges keresés.
- *Weinberg* [3] rámutatott, hogy sok szignifikáns szó/kifejezés csak a teljes szövegekben fordul elő, a címekben és a tartalmi kivonatokban nem.
- *Duckitt* [4] arra hozott példát, hogy egy fontos alapfogalom ("dosage" – dózis, adag, adagolás) sem a címben, sem a teljes szövegben nem fordult elő, így a szabad szöveges keresés csődöt mondott volna. Az összeállítás szerzőjének hasonló példája: a kávészakmában a közhasználatú "instant coffee" (neszkávé) megnevezés helyett a "soluble coffee" (oldódó kávé) kifejezés használatos, ezért az első kifejezés előfordulása címben, kivonatban és teljes szövegben egyaránt valószínűtlen.
- *Markey és társai* [5] megállapították, hogy az ERIC adatbázist használó tapasztalt keresők 68%-a használta a szabályozott szótárt. A szabályozott szótár nagyobb pontosságot, a szabad szöveges keresés nagyobb teljességet nyújtott, ami megfelelt a várakozásnak (vö. *Johnston* [6] korábbi felméréseivel). *Fenichel* [7] további ERIC-felmérései során az ERIC-használók három csoportját különböztette meg: a kezdő, a közepesen tapasztalt és a nagy gyakorlattal rendelkező keresőket. Az ERIC-ben járatlan felhasználók inkább a szabad szöveges keresést alkalmazták, a tapasztalt ERIC-használók inkább a tezauruszt. A teljesség a tapasztalt ERIC-felhasználóknál jelentősen nagyobb (58%) volt, a közepesen tapasztalt csoport érte el a következő legnagyobb teljességet (47%), és a legkisebb keresési költséggel is a közepesen tapasztalt csoport dolgozott. A pontosság minden csoportnál megközelítőleg azonos volt. Következtetés: az egyes módszereket csak a felhasználók igényeinek pontos tisztázása után, az igények ismeretében lehet reálisan értékelni.

▶ *Tatalias* [8] egy kutatóintézetnek az online adatbázisokkal ismerős, de közvetlen kereséseket nem végző munkatársai körében vizsgálódott: 91%-uk mindenképpen tezauruszt kívánt, mielőtt belevágna a keresésbe.

A *teljes dokumentumszövegeket tartalmazó adatbázisok* közül csak keveset vizsgáltak, ám igen szignifikáns pontossági és teljességi értékek jelentkeztek.

▶ *Tenopir* [9] vizsgálatainak eredménye szerint a teljes szövegekben végzett szabad szövegszavas keresések átlagosan 73,9% teljességet és 18% pontosságot nyújtottak. Ugyanezek az arányok a referátumok szövege esetében 19,3% és 35,6%, szabályozott szótáras keresésnél pedig 28% és 34% voltak (a vizsgált adatbázisban egy tétel legfeljebb 5 deszkriptort kaphatott).

A pontosság meredeken – 78,57%-ra – ugrott, amikor a szabad szöveges és a szabályozott szótáras keresést kombinálták.

▶ *Bryant* és *Terapane* [10] hasonló eredményekre jutottak: szabadalmi címek és kivonatok szabad szöveges keresésével a kérdéseknek csak 9,7%-ára kaptak választ, a teljes szövegek alapján a kérdések 87%-ára. A találatok nagyfokú relevanciájának biztosításával tervezett kísérletekben a szabad szöveges és osztályozó rendszeres keresés kombinációja ugyanazon szabadalmi adatbázisban az egyik vizsgált szakterületen 73,3% teljességet, a másikon 62,75%-ot hozott.

Az online információszolgáltatók a fentiekből általában arra a felismerésre jutottak, hogy együtt kell kínálni a szabad szöveges és szabályozott szótáras keresési lehetőségeket.

Ismérvek a kétféle keresés elemzéséhez

Szemantika. Az azonos jelentésű (szinonim), ellentétes jelentésű (autonim), rokonértelmű (kváziszinonim) szavak és az azonos alakú, de eltérő jelentésű (homonim) szavak mind bonyolítják a szabad szöveges keresést. A szóeleji és szóvégi csonkolásokkal a szótöveket azonosítani lehet, de a szinonimák problémája aligha kezelhető valamiféle szótár nélkül.

Szövegkörnyezet. A homonimaprobléma is speciális eljárásokat igényel. Az "adminisztráció" ügyintézést is jelent és "kormány" értelemben is használatos ("a Reagan-adminisztráció"), a "páholy" szó egyes szövegkörnyezetben építészeti értelemben szerepel, más kontextusban a szabadkőműves páholyokra vonatkozhat. A fogalmak terjedelmére vonatkozó megjegyzéseket (értelmezéseket) tartalmazó tezaurusz kiküszöbölheti a többértelműséget.

A nyelvek fogalmi szerkezete. A legtöbb modern nyelvben igen sok a generikus szó, egyes nyelvekben azonban nagyon kevés. Ha valaki a déligyümölcsök New York-i piacáról tájékozódik és csak szabad szavas keresést használ, nem találhatja meg a specifikus, de releváns piaci tanulmányokat a "pawpaw" és "guava" gyümölcsökről – hacsak előzetesen nem fektet kemény munkát a trópusi, ill. déligyümölcsök világának feltárásába, jóllehet a keresés még így is nehézkes és időigényes lesz. A tezaurusz a GYÜMÖLCS vagy DÉLIGYÜMÖLCS generikus kifejezésekkel segíthet, a nem hierarchikus nyelvekben pedig utalások (pointerek) nyújtanak megoldást.

Szakterület. Az egyes területek terminológiájának kötöttsége nagyon különböző. A kémiában ugyanarra a vegyszerre akár kilenc vagy még több szinonim kifejezés is használatos, a jogban egyedi kifejezések a jellemzőek. A gyorsan fejlődő területek terminológiája különösen képlékeny lehet.

Az emberi tényező. Valamely rendszer teljesítményének értékelése szempontjából igen jelentős, hogy az ember mit hogyan *érzékel*. Több pszichológiai tényező is jelentőséggel bír. Így:

▶ *Felelősség.* A szabad szöveges rendszerekben a kereső vállán van a felelősség (a szó pozitív értelmében is, hiszen szabadon kiaknázhatja a szövegben rejlő lehetőségeket), míg ha a szabályozott szótáras keresés eredménye nem kielégítő, az adatbázis előállítóját lehet okolni [11].

▶ *Az adatbázis áttekinthetősége.* A felhasználó számára általában észrevehetetlen a gyönge keresési teljesítmény, mert nem tudhatja, mi minden van az adatbázisban, és így egy keresés eredményének megítélésére nemigen van módja, különösen, ha elég sok a releváns találat [12].

▶ *Az elégedettségi tényező.* Sokféle tényezőn múlik; nemcsak a keresési eredmény, hanem a gyorsaság is befolyásolja, sőt a felhasználó típusától és a környezettől is függ.

▶ *A kezelés könnyűsége.* Nem szabad alábecsülni azt, hogy minimális erőfeszítéssel lehessen eredményt kapni az adatbázisból – bár minél jobban ismeri valaki a rendszert és a szakterületet, ennek a tényezőnek a szerepe annál kisebb. Az állandó és az alkalmi felhasználó valószínűleg másképp értékeli ugyanazt a rendszert.

A két keresési technikához kapcsolódó tényezők

A két technika előnyeit és hátrányait többen számba vették [11, 13, 14, 15], de ezeknek a tényezőknek – amelyeket az *1. ábra* szemléltet – a hatása egyáltalán nem olyan egyértelmű, mint amilyennek első pillantásra tűnhet.

Költségek. Szabad szöveges keresési technika használata esetén az adatbázis előállítója elvileg nélkülözni tudja a magas fizetésű osztályozókat-indexelőket – gyakorlatilag valószínűleg csak abban a speciális esetben, ha az adatbázis teljes szövegeket tartalmaz. A szöveges adatbázisokban többnyire nem teljes szövegű eredeti dokumentumok, hanem ezeket helyettesítő kivonatok vannak. Ezekben az esetekben valószínű, hogy a deszkriptorok kiválasztásához és leírásához szükséges idő alig számít(ana) ahhoz képest, amit a bibliográfiai leírás megszerkesztése, a referátum megírása stb. igényel. Ami költséges, az a tezaurusz létrehozása a semmiből.

A SZABAD SZÖVEGES KERESÉS ELŐNYEI

- olcsó
- egyszerű
- a dokumentum teljes információtartalma kereshető
- a keresés szempontjából minden szó azonos értékű
- emberi indexelési hibák fordulhatnak elő
- az új kifejezések késedelem nélkül beépülnek

A SZABAD SZÖVEGES KERESÉS HÁTRÁNYAI

- nagyobb teher nyugszik a felhasználón
- elveszíthetjük a szövegben implicit módon benne rejlő, de explicit módon nem kifejezett információt
- hiányzik a specifikus-generikus fogalmak közötti kapcsolat
- ismerni kell a szakterület szókészletét

A SZABÁLYOZOTT SZÓTÁRAS KERESÉS ELŐNYEI

- sok szemantikai problémát megold
- lehetővé teszi a generikus kapcsolatok azonosítását
- leképezi az adott szakterület fogalmi rendszerét

A SZABÁLYOZOTT SZÓTÁRAS KERESÉS HÁTRÁNYAI

- drága
- esetleg nem fedi le megfelelően a területet
- nem zárhatók ki az emberi (indexelési) hibák
- a szókészlet elavulhat
- nehéz szisztematikusan beépíteni a szavak/kifejezések között fennálló összes releváns kapcsolatot

1. ábra

A keresés egyszerűsége. Jóllehet a szabad szöveges kereséssel rögtön hozzá lehet férni az adatbázishoz, igencsak gondot kell fordítani a szótövek helyes megállapítására, a szinonimákra, a szövegkörnyezetre. Persze jó, hogy nem kell a tezauruszt vagy az osztályozó rendszert tanulmányozni, főleg, ha az ismeretlen. Ha viszont már ismerős, akkor a keresést valószínűleg nem bonyolítja, hanem egyszerűsíti. Sokat lehetne még tenni azért, hogy könnyebb legyen a tezauruszok kezelése; például "ablakot" lehet nyitni a képernyőn a tezauruszba való betekintéshez. A szabályozott szótárak szolgáltatóinak ügyelniük kell arra, hogy

- késedelem nélkül beépítsék az új szavakat/kifejezéseket;
- a teljes szakterület fogalomállományát leképezzék a szótár fogalmi rendszerében, és a rendszert könnyen áttekinthető hierarchiában jelenítsék meg a felhasználók számára, ellenkező esetben a szótár súlyosan félrevezető lehet, és a rossz keresési segédlet diszkreditálhatja az adatbázist;
- igen nehéz minden releváns relációt beépíteni; *Willetts* [16] tíz közismert tezauruszt elemezve jutott arra a következtetésre, hogy a relációk specifikálása és használata gyakran következetlen.

Egyelőre tisztázatlan a felsorolt tényezők egymáshoz viszonyított fontossága, hogy ti. melyik mennyire befolyásolja a teljesítményt.

A keresési módszer megválasztása

A legtöbb általánosan használható adatbázisban mind a szabad szöveges, mind a szabályozott szótáras módszerrel lehet keresni. A Dialog adatbázisainak több mint 60%-a kínál tezauruszt vagy egyéb szabályozott szótárat. Ritkán használnak osztályozórendszert a tényadatokat tartalmazó (faktografikus) adatbankok. Nagyon ritka a kizárólagos szabad szöveges keresési lehetőség, erre még leginkább újságok és hírszolgálatok szövegeit tartalmazó adatbázisok adnak példát. Az adatbázisok létrehozásakor a keresési módszer megválasztását illetően az alábbi szempontokat ajánlatos átgondolni.

A szakterület jellege. A szakterület nagyságát, korlátait és terminológiájának pontosságát vizsgáljuk.

A módszer költségei. Már szó volt róla, hogy a dokumentumok indexelése nem drága akkor, ha az eredeti dokumentumok valamiféle kivonatát visszük az adatbázisba, az új tezaurusz viszont sokba kerül. Azok a javaslatok, melyek szerint a tezauruszok csak keresési segédletek legyenek, és indexelésre nem kell őket használni [14], csak teljes szövegeket tartalmazó adatbázisok esetén lehetnek – esetleg – gazdaságosak. Ezen adatbázisoknál fel kell hívni a figyelmet arra, hogy az alacsony pontosságú keresés a teljes szövegben nyilvánvalóan drága, a hosszú keresési időt és az eredményként kapott igen terjedelmes anyagot figyelembe véve.

A felhasználói típusok. A felhasználók viselkedése és preferenciáik különbözőek, ezért ajánlott az igényeket előre elemezni.

A teljesítménytényezők prioritása. A nagy pontosság és teljesség adott esetben második helyre szorulhat a keresés gyorsasága és egyszerűsége mögött. Átfogóan elemezni kell nemcsak a felhasználó preferenciáit, hanem az információs szolgáltatás mélyebb belső céljait is.

Kívánatos további kutatások

Ahhoz, hogy a szabad szöveges és a szabályozott szótáras keresés értékeit és problémáit illetően szilárd következtetéseket vonhassunk le, további kutatások szükségesek, például az olyan témákban, mint

- a keresők viselkedése,
- keresési változatok szakterületek szerint,
- keresési változatok a keresők tapasztaltsága szerint,
- meggondolások a rendelkezésre álló keresési technika alapján,
- a keresési teljesítmény elemzése,
- az eredmények elemzése (míg a teljesítmény kizárólag a rendszer működési paramétereit jelenti, eredményeken azok az előnyök értendők, amelyeket a keresőket alkalmazó intézmények vagy az

önálló vállalkozói, vagy független kutatói minőségükben eljáró keresők szereznek a rendszer alkalmazásából).

Szaké rtő rendszerek

A valós környezetben végzendő kutatások szükségességét növeli az a körülmény, hogy tezauruszok és más hasonló keresési segédletek valószínűleg hamarosan láthatatlanná válnak az olyan szakértő rendszereken belül, amelyek lehetővé teszik a keresőkérdések természetes nyelven történő megfogalmazását. E szempontból fontos az ilyen rendszerek készítőinek megállapítása **[71:** a tezaurusz rendkívül (ontos eszköz annak biztosítására, hogy a rendszer sok különböző módon megfogalmazódott azonos kérdés nyomán ugyanazt a tartalmat keresse.

A tezaurusznak a szakértő rendszerekben való alkalmazása külön kérdés, egyesek integrálni akarják az ismeretbázisba, mások inkább megőriznék önálló segédeszköznek.

a a a

A keresés hatékonyságát annyi külső változó befolyásolja, hogy valószínűleg nincsenek általános érvényű optimális megoldások. Még az sem állítható, hogy mindig a két módszer együttes alkalmazását kell ajánlani. Leginkább arra van szükség, hogy a módszerek működését egyre jobban megértsük, mégpedig pontosan leirt környezetekben vizsgálva őket, amelyekben a releváns változót pontosan azonosítani lehet.

Irodalom

|i] HENZLER. R. G : Freeor controlled vocabularies = International Classification, 5. köt. 1978. p. 2 1 - 26.

[21 HERSEY. D. F. et al.: Free text word retrieval and scientist indexing and retrieval. = Journal ol Documenlatron. *27.* köt. 1971 p 167- 183.

[31 WEIN8ERG. H. B.: Multiple sets ol humán indexing for civil engineering documenls: Comparison of structure and occurence rates in full text. = Science and Technology Líbraries. 2. köt. 3. sz. 1982. p. 13- 33.

[4] DUCKITT, P.: The value 0(controlled indexing systems in online full texl databases. = Proceedings ot the Sth International Online Information Meeting. Oxford. Learned Information, 1981.p. 447- 453.

[5] MARKEY, K. et al.: An analysis of controlled vocabulary and free text search statements in online searches. = Online Review, 4. köt. **1**980 **p.** 225- 236.

[61 JOHNSTON. S. M.: Effect **pl** thesaurus indexing in retrieval írom machine-readable databases. = Ouarterly Bulletin ol the International Association of Agricultural Librarians and Documentalists, 27. köt. 1982. **p.** 90- 96.

[7) FENICHEL, C. H.: An examination of the relationship between searching behavior and searcher background. = Online Review. 4. köt. 4. sz. 1980. **p.** 341- 347.

[8] TATALIAS, J.: Attiludes and expectations ol polenlial and user online searchers. = Proceedings of the National Online Meeting. New York. 1985. **p.** 457- 462.

[9] TENOPIR. C: Full text database retrieval performance. = Online Review, 9. köt. 2. *sz.* 1985. **p.** 149- 164.

[10) BRYANT, J. H.-TERAPANE, J. F.: Online searching in the US Patent and Trademark Office. = World Patent Information. 7. köt. 1 - 2. sz. 1985. **p.** 133- 138.

[11] ROTHMAN. J : Is indexing obsolete? = Feinberg. H. (ed.) Indexing Specialized Formats and Subjects. Metushen. New Jersey and London. Scarecrow Press. 1983. **p.** 22- 23.

[12] DUBOIS, C. P. R.: The use of thesauri in online retrieval. = Journal of Information Science. 8 köt. 2. sz. 1984. **p.** 63- 66,

[13] PEREZ. E.: Text enhancement. controlled vocabulary versus tree text. = Special Líbraries, 73. köt. 3. sz 1982. **p.** 183- 192.

[14] GASTALDY. S. B : Les thesaurus de recherche: des outils pour l'interrogation en vocabulaire libre. = Árgus, 13. köt. 2. sz. 1984. p. 51- 56.

[15] FEINBERG, H.: The thesaurus in indexing and searching - I. 11, **p.** 260- 281.

[I 6] WILLETTS, M.: Investigation of the relation between terms in thesauri. = Journal ol Documentation, 31. köt 3.**sz.** 1975. **p.** 158- 184

[17] BERNSTEIN. L M - WILLIAMSON, R. E.: Testing of a natural language retrieval system for a full lest knowtedge base. = Journal of the American Socíely for 1nformalion Science. 35. köt. 4. sz. 1984. **p.** 235- 247.

/DUBOIS. C. P. R.; Free text vs. controlled vocabulary; a reassessment. = Online Review, 11. köt. 4. sz. 1987. p. 243- 253./

(Szöllősy Éva)

A hasonlósági keresés úttörője: a duplumrekordok kizárása a Dialógnál

Az online információkeresés mostanáig mindig pontos egyezéses keresési Jelentett, vagyis a keresökifejezés és a kikeresett dokumentumban előforduló kifejezés - a csonkolás vagy maszkolás megengedte szabadsággal - pontosan meg kellett egyezzen egymással Az ennél általánosabb hasonlósági keresésre, vagyis arra, amikor a kikeresett dokumentumban a keresökifejezéshez hasonló, de atlól azért többé kevésbé eltérő kifejezés előfordulását várjuk el, már vollak kísérletek. Az itl ismertetetthez hasonló dunlumkizárással például főleg mikioszámilógépes utófeldolgozás kerelében próbálkoztak. Az itt ismerteteti eljárás azonban az elsö. amely a hasonlósági keresést nyilvános keresőrendszerben valósítja meg A ref