Két online információkeresési módszer, a szabadszöveges és a deszkriptoros visszakeresés összehasonlítása

A számítógépes információkeresés állandóan visszatérő kérdése: a szabadszöveges vagy a szabályozott kulcsszavas keresés az előnyösebb, a hatékonyabb? Ha egy adatbázisban mindkét típusú keresésre mód van, választanunk kell, mégpedig a két módszer előnyeit és hátrányait mérlegelve. A mérlegelésben figyelembe veendők a következők:

- szabadszöveges kereséssel automatikusan válogatunk ki *minden* olyan információs tételt, amelyet a szóban forgó deszkriptorokkal indexeltek, továbbá azokat is, amelyek címében, referátumában stb. előfordulnak ezek a keresőszók;
- szabályozott kulcsszavak (deszkriptorok) alapján végzett információkereséssel a *leginkább releváns* tételeket válogathatjuk ki, mert az indexelő feltehetően a valóban releváns tételekhez rendeli hozzá a megfelelő deszkriptorokat.

Ezeket a kérdéseket vizsgáljuk meg részletesebben az amerikai *Környezetvédelmi Hivatal (Environmental Protection Agency, EPA)* kutatóintézeti könyvtárának egy konkrét információkeresési esete kapcsán.

Az EPA könyvtárának gyakorlata

A környezetvédelmi információs szolgáltatások különféle diszciplinákra terjednek ki. Ezért a keresést is különféle, nevezetesen orvostudományi, kémiai, biológiai, toxikológiai, műszaki és más adatbázisokban kell elvégezni. Az EPA könyvtárában számítógépes online információkereséseket részben a környezetvédelmi kutatások támogatására, részben a levegő-, talaj- és vízszennyezési előírások, szabványok kidolgozására és ezek betartásának ellenőrzésére végeznek. Az online információszolgáltatások valamennyi költségét a könyvtár fedezi.

A könyvtár két profilszerkesztő munkatársa évente átlag 2100, vagyis havonta átlag 180 retrospektív keresést hajt végre online módszerrel. Egy-egy keresőkérdés megválaszolására átlagosan 3,2 adatbázist vesznek igénybe. A profilszerkesztők a keresőstratégiát rendszerint megbeszélik a kérdést feltevő felhasználóval, aki többnyire maga is részt szokott venni a keresésben.

A módszer

Az esettanulmány céljára kiválasztott keresőkérdés a területfeltöltések kilúgozására vonatkozik (l. később), mert ezt a kérdést olyannak ítélték, amely sokféle tipikus döntésre kínál alkalmat. A keresést – hosszas megfontolás után – végül is két adatbázisban végezték el, méghozzá a felhasználó jelenlétében. A felhasznált adatbázisok a COMPENDEX és az ENVIROLINE voltak.

A kísérlet körülményeit a következőkben határozták meg:

1. A vizsgálat célja az online keresés során mérlegelendő döntések értékelése. A döntéseket a következő tényezőkre kell meghozni;
   a) az adatbázis tematikai tartalma;
   b) az információkeresés fogalmainak, paramétereinek meghatározása;
   c) a keresőstratégiák kiválasztásához alkalmas keresőszavak kiválasztása;
   d) a keresőszavak és kombinációik megváltoztatása a keresés közben.
2. Előzetesen elkészítenek három keresőprofilt: egy-egy „rövid keresőprofilt" mindkét adatbázishoz, amely csak deszkriptorokat tartalmaz, továbbá egy „hosszú keresőprofilt", amely mindkét adatbázisban használható és tetszés szerinti keresőszavakat tartalmaz (a szabadszöveges kereséshez).
3. A kísérletben a Lockheed DIALOG rendszer használata szerepel, mégpedig az új SuperSELECT utasításrendszer alkalmazásával és a szavak csonkolási lehetőségének maximális érvényesítésével.

A keresőstratégiák elve

A kitűzött módszer kitűnő lehetőséget kínált az alapkérdés, vagyis a kétféle információkeresési megközelítés összehasonlítására, mivel általa

a hosszú, átfogó keresőstratégia is a rendelkezésre állt, amely a szabadszöveges keresésre nyújtott lehetőséget mindkét adatbázisban;

a rövid, csak deszkriptorokra és osztályozási jelzetekre (tehát szabályozott keresőszavakra) alapozott stratégiák – mindkét adatbázisban külön-külön – ugyancsak a rendelkezésre álltak.

A kísérlettől azt várták, hogy az első módszerrel az adatbázisokban levő *valamennyi* releváns dokumentumot ki fogják keresni, és ezt a stratégiát – akár a későbbiekben is – bármely adatbázisban változtatás nélkül fel tudják majd használni. Mivel az ilyen megoldáshoz nem kell igénybe venni az egyes adatbázisok használatához készült nyomtatott segédleteket (tezauruszok, osztályozási rendszerek stb.), sokan – az időmegtakarításra hivatkozva – ezt a látszólag egyszerűbb módszert kedvelik. A kísérletnek tehát be kellett volna bizonyítania, hogy a szabadszöveges keresés a magasabbrendű online módszer.

A keresés előkészítése

A kiválasztott keresőkérdés konkrétan így hangzott: *Hulladékkal, szeméttel való területfeltöltés kilúgozásának hatása a talajvíz minőségére.*

A kérdést feltevő személy a területfeltöltések egészségügyi következményeivel foglalkozó kutatási program felelőseként dolgozik az EPA-ban. Célja a témára vonatkozó új kutatási, fejlesztési eredmények megismerése, az áttekintő szakirodalom információs anyagának összegyűjtése, valamint publikálásra alkalmas bibliográfia összeállítása volt, tehát átfogó irodalomkutatásra törekedett.

A keresés előkészítéséhez feltétlenül figyelembe kellett venni, hogy a kérdést feltevő felhasználó nem egy meghatározott szűk terület abszolút releváns dokumentumait kívánja, hanem inkább teljességet, vagyis egy viszonylag bő témával foglalkozó valamennyi releváns dokumentum kikeresését igényli. A felhasználó a keresés előkészítésében a vonatkozó keresőszavak, szakkifejezések megadásával, a keresés online végrehajtásában pedig a válaszok megtekintésével és a relevancia elbírálásával vett részt.

A keresőkérdést négy fogalomcsoportra lehetett bontani:

1. Területfeltöltés,
2. Szemét, hulladék,
3. Kilúgozott anyagok vagy talajvíz,
4. Mérgező anyagok.

A „hatás"-t és rokonértelmű kifejezéseit mint keresőszókat eleve elvetették, mert a sok előfordulás egyfelől meglassította, tehát megdrágította volna a keresést, másfelől a stratégiát is túlságosan behatárolta volna.

A stratégia elemzése során később úgy döntöttek, hogy a negyedik fogalomcsoportot is kihagyják, mert valamennyi várható mérgező anyag felsorolása nagyon hosszúvá tenné a keresőprofilt, tehát ez is gazdaságtalanná tenné a keresést. A gyakorlatban úgyis az a helyzet, hogy a szemétből, hulladékból kilúgozott *bármely* anyag eleve szennyezi az ivóvizet, tehát ebben a vonatkozásban feltehetőleg mérgező hatású. A jó keresőstratégiához nemcsak azt kell tudni, hogy milyen fogalmakat vegyünk fel a profilba, hanem azt is, hogy *mit hagyjunk ki belőle.*

Mindezt a stratégiák és a keresőprofilok összeállítása és a keresések elvégzése követte, először a COMPENDEX, majd az ENVIROLINE adatbázisban.

A COMPENDEX adatbázisban végzett keresés

Először a deszkriptorok és osztályozási kódok alapján szerkesztett, ún, „rövid profillal" végezték el az online információkeresést.

Ahhoz, hogy a jó rövid profilt ügyesen meg lehessen szerkeszteni, alaposan ismerni kell a COMPENDEX keresőrendszerét, amely ötféle kiadvány ismeretét és optimális használatát tételezi fel. Ezekhez jól használhatók a DIALOG rendszer ilyen céllal rendekezésre bocsátott kézikönyvei.

Az eredmény „tiszta" találatokból állt, vagyis az információs „zaj" minimális volt. A keresés végrehajtásához alig néhány percre volt szükség, tehát ez a fajta keresés igen olcsónak bizonyult.

A rövid profillal végzett keresés eredménye a COMPENDEX adatbázis 1970–1979-es évfolyamából származó 152 találat volt. A kereséshez mindössze 4,5 percre volt szükség, aminek költsége 4,86 $. A 152 találat off-line kinyomtatási költsége 15,20 $-ba került, tehát az online szolgáltatás összköltsége 20,06 $-t tett ki.

Mint mondottuk, a kérdést feltevő felhasználó maximális teljességre törekedett. Ezért megpróbálkoztak a hosszú keresőprofillal is, amelyben olyan szabad keresőszavakat (a tezaurusztól független szakkifejezéseket) használtak, amelyek az információs tételek bármelyik részében (cím, referátum, kulcsszavak) előfordulhattak.

A keresőstratégiát a már közölt három fogalomcsoport szerint építették fel, az egyes csoportokat az ÉS logikai operátorral kombinálták. A keresőstratégia elve ez volt:

[*Terület-feltöltés* szakkifejezései] ÉS [*Szemét, hulladék* szakkifejezései] ÉS [*Talajvíz, kilúgozás* szakkifejezései]

E három csoport mindegyike a rokon- és kapcsolódó értelmű szavakat, szakkifejezéseket és változataikat egymással VAGY kapcsolatban tartalmazta. Ahol lehetett, a szavakat csonkoltan alkalmazták, ami megsokszorozta a találatok valószínűségét. Összesen 26 (egyszerű vagy összetett) csonkolt keresőszóból állt a profil.

A keresés eredménye 300 információs tétel volt, közülük 148 a rövid profillal is találatként adódott. A 162 új találat közül igen sok bizonyult zajnak, ami szabadszöveges keresés esetében természetes. A zaj egy része a tengeri, óceáni hulladékra vonatkozó információ volt. Az erre vonatkozó keresőszavak kizárásával (NEM operátor használatával) 151 találat maradt meg a 162 közül.

A hosszú profillal végzett keresés ideje 21 perc, költsége 23 $ volt (szemben az előző nem egészen 5 $ költséggel); a találatok száma viszont az előzőnek kétszerese lett. A két profil *nem közös* találatainak összege: 152 + 151 találat, ezek összes nyomtatási költsége 30,30 $. A szabadszöveges keresés + az összes találat nyomtatási költsége 53 $-t tett ki.

Mi következik ebből? Látható, hogy a szabadszöveges keresés *nem hozza ki az összes releváns dokumentumot,* annak ellenére, hogy jóval hosszabb keresési időt vett igénybe, mint a deszkriptoros keresés. Az utóbbi, a

szabályozott keresőszavakkal végzett keresés sokkal olcsóbb. Ez sem hozta ki ugyan az összes releváns dokumentumot, de a leginkább relevánsakat mindenképpen. *A maximális teljesség igénye esetén mind a kétféle keresést el kell végezni!*

Az ENVIROLINE adatbázisban végzett keresés

Ugyanezt a témát keresték az *Environmental Abstracts* online hozzáférésű változatában, az ENVIROLINE adatbázisban is. Ennek használati díja a DIALOG rendszerben magasabb, mint a COMPENDEX adatbázisé: az előbbié óránként 90 $, az utóbbié 65 $. Sajnos, a két adatbázis eltérő indexelési és osztályozási rendszere miatt a COMPENDEX adatbázisra kidolgozott rövid profil nem volt használható az ENVIROLINE-ban, a hosszú profil viszont változtatás nélkül bizonyult alkalmasnak.

A deszkriptorokon alapuló rövid profillal 183 találat adódott, amelyek mindegyike relevánsnak tűnt. A keresés összköltsége 45,60 dollárt tett ki (9 $ keresési díj + 36,60 $ nyomtatási díj).

A megőrzött hosszú profillal 323 találatot kaptak, a felhasználó véleménye szerint ezek 2/3-a volt releváns. Költségek: a 9 perc kapcsolási költsége 13,50 $, a 323 találat nyomtatási költsége 64,60 $, összesen 78,10 $. Feltételezték, hogy ezzel a módszerrel valamennyi releváns dokumentumot kikeresték.

Amikor a deszkriptoros keresés 183 találata közül NEM operátorral kizárták azokat, amelyeket a szabadszöveges keresés kihozott, az derült ki, hogy az előbbi kereséssel 28 elsőrendűen releváns dokumentumot nem találtak meg. Ez tehát a rövid stratégia vesztesége, amit az olcsóbb keresés ellenértékeként fizetni kell.

A következtetés hasonló, mint a COMPENDEX adatbázis esetében: a releváns anyag maximális kihozatalához *a kétféle stratégiát kombinálni kell.* A kétféle keresés az ENVIROLINE adatbázisban összesen 351 tételt eredményezett. Ezek keresési költsége (22,50 $) és nyomtatási díja (70,20 $) összesen 92,70 $-ra rúgott.

A felhasználó véleménye

A felhasználó a kereséssel és eredményével meg volt elégedve. Ebben közrejátszott az a tény, hogy együttműködött a profilszerkesztővel, és néhány fontos döntést szaktudásával elősegített.

A két adatbázisból kikeresett információkat mennyiségben és minőségben egyaránt megfelelőnek ítélte. Megállapította, hogy a két adatbázis kiegészíti egymást, a COMPENDEX a műszaki szakirodalom javát képviseli, az ENVIROLINE pedig főleg az értékes vállalati irodalmat és kutatási jelentéseket tartalmazza.

Következtetések

A két adatbázisban végzett kísérleti online információkeresés előkészítése és végrehajtása közben a következő döntéseket kellett hozni:

A keresés mértéke

Meg kellett határozni előzetesen azt, hogy a keresés nagy teljességre törekvő vagy korlátozott mértékű legyen-e. Ezt főleg az információkeresés végső célja döntötte el.

Fogalomcsoportok

Azonosítani kellett a keresés fogalomköreit, paramétereit. Dönteni kellett ezt követően abban, hogy mely fogalomcsoportokat fognak használni az optimális információkeresés igényét fenntartva, mégpedig túl nagy kompromisszumok nélkül. Általában 2–4 fogalomkört szoktak definiálni, és mint ennél az esetnél is bebizonyosodott, sokszor egy-egy csoport elhagyása is igen fontos és eredményes döntés lehet.

Adatbázisok

A keresésre használandó adatbázisokat annak alapján kell kijelölni, hogy mely tématerületekre terjednek ki, valamint milyen dokumentumokat foglalnak magukban és hogyan indexelik őket. A kérdést feltevő szakember témájának szakterülete (kémia, műszaki tudományok, orvostudomány stb.) határozza meg az elsődleges forrást, a többi ennek kiegészítéseként jöhet számításba. Az adatbázis(ok) kiválasztásának egy fontos tényezője a számítógépi kapcsolat és a nyomtatás díja is.

Keresőstratégia

Ennek meghatározásához el kell dönteni, hogy szabályozott indexkifejezésekre (deszkriptorokra, osztályozási kódokra) alapuló vagy szabadszöveges keresést fogunk-e alkalmazni a releváns dokumentumok kikeresésére. Az előző típusú keresés eredménye rendszerint korlátozott számú, de a leginkább releváns dokumentumok kiválasztása, ezek némi veszteségével számolva. Az ilyen típusú információkeresés a legolcsóbb. Az ismertetett kísérlet bebizonyította, hogy a drágább szabadszöveges keresés sem hozott ki minden releváns dokumentumot. Maximális kihozatalt (minden releváns dokumentumot, de tekintélyes zajjal) csak a két módszer kombinálásával lehet elérni.

Keresőprofil-módosítás

Az online módszer a legrugalmasabb információkeresési eljárás, mert az első lépésben kikeresett információs tételek megtekintése alapján tetszés szerint megváltoztat-

hatók a keresőszavak (egyesek elhagyhatók, újak bevihetők) és ezek logikai kombinációi. Az erre vonatkozó döntéseket célszerűen a felhasználó hozza meg, és ezzel a lehetőséggel maximálisan kell élni.

A kinyomtatási formátum

Ugyancsak az információt felhasználó személy döntheti el azt, hogy számára melyik adott nyomtatási formátum (az információk mely adatainak kiíratása) a legkedvezőbb.

A gyakorlott profilszerkesztők tudják, hogy mennyire nehéz annak eldöntése, hogy melyik módszer az optimális; keresés közben sokszor automatikusan kényszerülnek döntésekre. A keresési hatékonyság növelésének egyik legfontosabb kritériuma a *jó kommunikációs lehetőség a profilszerkesztő és a felhasználó között.*

A profilszerkesztőnek a feladott témára vonatkozó szakmai ismeretei meghatározó fontosságúak a keresés eredményességében, ui. csak megbízható szakmai háttér esetében képes a megfelelő keresőszavak kiválasztására és az optimális stratégia kidolgozására. Ismernie kell továbbá a használt adatbázisok tartalmát, indexelési, osztályozási rendszerét, deszkriptorainak használati módját. Ehhez ugyancsak gyakran kell forgatnia az adatbázisokhoz tartozó nyomtatott segédleteket. Az online szolgáltatások kézikönyvei is jó szolgálatot tehetnek a keresősoftware lehetőségeinek, korlátainak, valamint az egyes adatbázisok jellemzőinek megismerésében.

Röviden: a legeredményesebb az online keresés akkor, ha a profilszerkesztő abban a helyzetben van, hogy a felhasználók által feltett különféle keresőkérdéseket olyan stratégiákká, keresőprofilokká tudja alakítani, amelyek a várt információkat a legkisebb idő- és pénzráfordítás mellett, optimális kihozatallal szolgáltatják.

/CALKINS, M. L.: Free text or controlled vocabulary? A case history step-by-step analysis... Plus other aspects of search strategy = DATABASE, 1980. június, p. 53–67./

(Roboz Péter)

Online információkeresés több adatbázisból

Az online információkereső szolgáltatások egyre növekvő mértékű igénybevételével együtt emelkedik az egyazon keresési művelethez igénybe vett adatbázisok száma. Ez a tendencia nemcsak a keresést végző személyek munkáját, hanem a szolgáltatás üzemeltetőjét is jelentősen érinti. Az új irányzat az adatbázisok termelőire is hatással van, de az előbbieknél kisebb mértékben.

A keresést végző személy

Egy retrospektív információkeresés megoldásához az online korszak előtt nemigen volt idő vagy mód több adatbázisból való információkeresésre. Régebben rendszerint a problémára vonatkozó elsődleges adatbázist hasznosították, a többit nem. Az online szolgáltató rendszerek ma már módot nyújtanak több adatbázis használatára is az adott rendszeren belül, s ezért a kereső személy nemcsak a „primer" adatbázisban férhet hozzá releváns információkhoz. Az idő- és a költségtényezők sem zárják ki a keresés teljességének növelését a másodlagosnak tekintett adatbázisok hasznosításával.

Minél tapasztaltabb valaki az online keresés technikájában, annál inkább több adatbázist is kipróbál egy-egy keresésnél. A technika elsajátításának legnehezebb fázisa az, amikor az első két-három adatbázist „kell megtanulni", utána már egyre könnyebb lesz a dolog. Az online információkeresés módszerének megismerése hármas feladat:

az online keresés elméletének és terminológiájának elsajátítása;

a számítógéppel való kommunikálás, a terminálkezelés és a keresőnyelv megismerése;

az adatbázisok használatának begyakorlása.

Az adatbázis termelője

Amint a fent vázolt folyamatban a rendszer felhasználója egyre nagyobb jártasságra tesz szert, egyre inkább felvetődik benne az egységesítés igénye. Könnyen belátható előny ugyanis, ha minél több adatbázis-termelő használja pl. az egységes folyóiratcím-kódokat (CODEN) vagy a szerzői nevek egységes írásmódját. Sajnos, az ilyenfajta szabványosítás igen lassú folyamat, a keresők bármennyire is szeretnének egy „ülésben" több adatbázist hasznosítani.

Ennek nem az az oka, hogy az adatbázis-termelők nincsenek tisztában a problémával, de azt szeretnék, ha az adatbázisok szabványosítását, egységes jelrendszerre hozatalát nem ők, hanem a szolgáltató rendszerek üzemeltetői oldanák meg. Úgy vélik ugyanis, hogy évek óta kialakított indexelési rendszereik megváltoztatása – az egységesítés kedvéért – csaknem lehetetlen feladat elé állítaná őket. A keresőprogram eszközeivel sokkal olcsóbban lehetne megvalósítani bizonyos fokú egységesítést. Ilyen lehetőség pl. a csonkolás alkalmazása, amely bizonyos mértékig kiegyenlíti az indexkifejezések vagy a név-írásmódok eltéréseit.

Az online szolgáltatás üzemeltetője

A több adatbázisra alapozott online retrospektív keresés (amit multiadatbázis-keresésnek vagy keresztt-