*nak szabályai.* Más és más szabályok érvényesek az alaptudományokban, a komplex tudományokban, az alkalmazott tudományokban és az integrált tudományokban. Az alaptudományok területén például

a matematikai tezauruszba matematikai,
a fizikaiba fizikai + matematikai,
a vegyészetibe kémiai + fizikai + matematikai,
a biológiaiba biológiai + kémiai + fizikai + matematikai,
a társadalomtudományiba társadalomtudományi + biológiai + kémiai + fizikai + matematikai

terminológia is bekerül.

A szerző a javasolt egyetemes vezérfonal alapján *sikerrel fejlesztett ki több kompatibilis tezauruszt* (társadalmi kommunikáció, papíripar, hajózás, atomenergetika). Az egyetemes vezérfonal alkalmasságát egy, az USA-ban készült és fazettás építkezésű oktatásügyi tazeurusz összehasonlító elemzésével is eredményesen bizonyította be.

*/SZOKOLOV, A. V.: Ob odnom vozmozsnom podhode k obeszpecseniju szovmesztimoszti IPT. Universzal'naja fabula informacionno-poiszkovüh tezauruszov. = Naucsno-Tehnicseszkaja Informacija, 2. sor. 1. sz. 1977. p. 19–24./*

(Futala Tibor)

## A TÁJÉKOZTATÁS GÉPESÍTÉSE

### Automatikus profilszerkesztést segítő módszer

A számítógépes információkeresés munkaigényes megelőző feladata a *keresőprofilok szerkesztése.* Célszerű ezért olyan eljárást keresni, amellyel a ráfordított idő és a manuális munka csökkenthető, illetve automatizálható legyen.

A végső cél olyan módszer kifejlesztése, amelyben a felhasználó egyetlen feladata a kikeresett dokumentumok relevanciájának az elbírálása, míg a kiválasztott dokumentumokat a számítógép már automatikusan elemzi, értékeli.

#### *Módszertan*

A releváns dokumentumokat az irreleváns dokumentumoktól a *felhasznált terminológia alapján lehet megkülönböztetni.* Bizonyos szakkifejezések, szókombinációk vagy szótöredékek a releváns dokumentumokban gyakrabban fordulnak elő, mint az adatbázis egészében, és ezek képezik a potenciális keresőkifejezések halmazát. Tehát egy szó vagy kifejezés hasznosságát úgy definiálhatjuk, mint annak előfordulási gyakoriságát a releváns anyagban, viszonyítva a teljes adatbázisban való előfordulási gyakoriságához. Erre jellemző szám a kifejezés *specifikussága,* amelynek definíciója az előzőek szerint:

$$S = \frac{R}{N}, \quad 0 \leqslant S \leqslant 1,$$

ahol $R$ a kifejezést tartalmazó *releváns* dokumentumok száma,
$N$ a kifejezést tartalmazó *összes* dokumentum száma.

*Hasznos kifejezések* azok, amelyek nagy gyakorisággal jelennek meg a releváns kis gyakorisággal az irreleváns dokumentumokban. Hasznos kifejezésekre $S$ értéke 1-hez közeli szám. Viszont pontosan 1 lehet az értéke akkor is, ha az egész adatbázisban csak egyetlen egy releváns dokumentum van és csak ebben, egyszer jelenik meg a kifejezés, hiszen ebben az esetben $R = N = 1$. A specifikusság rendjében készült szójegyzékben éppen ezek a csak egyszer előforduló kifejezések helyezkednének el legfelül, torzítva az eredményeket. *Az ellentmondás feloldására a kifejezés specifikusságát másképpen célszerűbb definiálni:*

$$S = \frac{R^2}{N}$$

Ha $R > 1$, és az irreleváns anyagban való megjelenés gyakorisága $(X = N - R)$ kicsi, $S > 1$ adódik. Tehát minél hasznosabb egy kifejezés a keresés szempontjából, azaz minél több releváns dokumentumban fordul elő, annál nagyobb $S$ érték tartozik hozzá, és a specifikussági rangsorban annál magasabban helyezkedik el.

A módszer kipróbálására mindegyik minta-kérdésre három CAC mágnesszalag összesített adatbázisából keresték ki a releváns dokumentumokat. Kiszámították az ezekben szereplő kifejezések $S$ értékét, rangsorolták azokat csökkenő értékük szerint, majd az így adódó rangsor alapján a fenti adatbázisból ismét kikeresték a releváns dokumentumokat és ezeket hasonló módon értékelték ki. Ezt az egyre kevésbé specifikus kifejezésekkel való keresésen alapuló közelítést mindaddig megismételték, amíg több új releváns dokumentum nem jelentkezett. Majd minden kifejezésre *kiszámították a keresési pontosságot (precision) és a relatív lehívási értéket (recall)* (a kifejezés alapján kikeresett releváns tétel/összes kikeresett releváns tétel).

Kezdetben e módszerrel csak *egykifejezéses* specifikussági listákat készítettek. Számos fogalom azonban nem adható meg egykifejezéses listákkal. Ezek a fogalmak a kifejezések logikai ÉS kapcsolatával határozhatók meg, így keletkeznek a *kifejezés-párok.* Megtörténhet, hogy az alacsony specifikusságú kifejezések esetleg magasabb specifikusságú kifejezés-párokat adnak. Ennek megfelelően előállították a kifejezés-párok specifikussági listáját.

Hasonló módon elkészítették a *szótöredékek* specifikussági listáját is. A szótöredékek alkalmazása minimálja a profilban szükségszerűen megadandó kifejezések számát, kiterjeszti a kifejezés változataira történő keresés lehetőségét a minta adatbázison túl is.

A specifikussági listák jelentős tulajdonsága, hogy mutatják a *kumulativ profiljellemzőket.* A specifikussági küszöbérték feletti összes kifejezést felvéve a profilba, a küszöbérték kijelölésétől függően, változó működési jellemzőkkel meghatározható profilok készíthetők. A felhasználó dolga, hogy a profilok így kialakítható spektrumában meghatározza az optimálisat, a kívánt minőségi jellemzők és az ehhez szükséges költségek mérlegelésével.

A specifikussági listákat megfelelő programmal *keresőprofilokká alakították,* és ezeket a profilokat a CAC adatbázisból vett másik három füzet összesített adatbázisával vetették össze. A keresés eredményét összehasonlították az ugyanebből az adatbázisból manuálisan készített profilokkal végzett keresés eredményeivel.

### *Az automatikusan szerkesztett profilok fajtái*

Négyféle profilt szerkesztettek a fenti módszerrel: *egykifejezéses, egykifejezéses/kifejezéspáros, egykifejezéses/szótöredékes és csak kifejezéspáros profilt.* A rendelkezésre álló 142 kérdés-mintából 68-ra készítették el a két előbbi típusú profilt és ezek közül 10 kérdésre a két utóbbi típusú profilt is.

*A legegyszerűbb az egykifejezéses profilok készítése,* ami a logikai VAGY operátorral kapcsolt szavak listáját jelenti. A legnagyobb specifikusságú dokumentumok a leginkább relevánsak.

Az egykifejezéses/kifejezéspáros profilokat a kétféle specifikussági lista egyesítésével alakították ki. A kifejezéspárokat a kifejezések logikai ÉS kapcsolatával állították elő. A kifejezéspárokat és kifejezéseket külön-külön kezeli a program, egymással VAGY kapcsolatban. A tisztán kifejezéspárokkal készült profilok találatait összehasonlították a megfelelő egykifejezéses és vegyes profilokkal kapott eredményekkel.

### *Specifikussági lista előállítása*

*142 profil specifikussági listáját állították elő és ezeket részletesen elemezték.* A profilszerkesztésbe bevont hipotetikus átlagos felhasználónak mintegy 100 találatot kell kiértékelnie az egykifejezéses specifikussági lista készítéséhez, és további mintegy 50 tételt a kifejezéspárok listájának készítéséhez.

Az automatikus profilszerkesztés költség-hatékonysága függ a kifejezések előfordulási forrásainak kijelölésétől (deszkriptorok vagy a cím szavai), a profilok fent részletezett típusaitól, a profiltól megkívánt jellemzőktől (pontosság, lehívás), a gép adataitól stb.

### *Információkeresés automatikus profilokkal*

Az összes profil valamennyi változata alapján végzett keresési kísérlet output-jait használták fel a profilok csoportjaihoz tartozó össztalálatszám, összes releváns találatok száma, minden profil különböző változatai által szolgáltatott releváns találatok száma, a pontosság és a lehívás értékének kiszámítására. Ezenkívül meghatározták az egyes profilcsoportokra jellemző összértékeket a találatok, a releváns találatok és a lehívás bázisának összegezésével, valamint a pontosság és a lehívás adatainak kiszámításával. Az eredményeket az *1. táblázat* összesíti.

A keresési idők összevetéséből úgy tűnik, hogy a legrövidebb idő alatt, tehát *legolcsóbban futtatható automatikus profil főleg kifejezéseket és alig néhány kifejezéspárt és szótöredéket tartalmazhat.*

Valamennyi profilhalmaz értékeléséből az adódik, hogy *a manuálisan szerkesztett profilok jobbak, mint az automatikus profilok,* az egykifejezéses és az egykifejezéses/kifejezéspáros profilok különbsége pedig statisztikailag nem szignifikáns. A keresési időket tekintve, a manuális profilok eltérése másképp értékelhető. Az egykifejezéses profilok keresési ideje sok esetben egy nagyságrenddel kisebb a többinél. Az egykifejezéses/kifejezéspáros profilok keresése drágább, mint a manuális profiloké, az eredmények viszont csaknem azonosak; tehát a kifejezéspárok alkalmazása viszonylag kevésbé hatékony keresési eszköz. A kifejezéspárok inkább a keresés pontosságát javítják, mint a lehívási arányt.

A kifejezéspárok hatásának jobb megismerését szolgálja az a kísérlet, amelyet csak kifejezéspárokat tartalmazó 10 profillal végeztek. A táblázat ezek összehasonlítását mutatja az egykifejezéses profilok hatásával is. A kifejezéspárok sokkal kevesebb adatot szolgáltattak, mint az egykifejezéses profil, noha az átlagos pontosság a két esetben azonos nagyságrendű. A kifejezéspárok lehívási aránya viszont sokkal kisebb (18,7%), mint a kifejezéses profil alkalmazásával nyert megfelelő érték (46,8%).

Mindent összevetve tehát legjobb az egyszerű egykifejezéses lista, mely viszonylag olcsó és mégis eredményes. A szótöredékek értékes lehívási eszköznek tekinthetők, viszont futtatásuk meglehetősen drága, és a pontosság csökken.

A profilok teljes halmaza két csoportra osztható úgy is, hogy külön vizsgálják a csak egy fogalmat kifejező profilokat *(egyfogalmú profilok)* és az egymással kapcsolatban álló, de eltérő fogalmakat tartalmazó profilokat *(többfogalmú profilok).* Azt gondolhatnók, hogy az egykifejezéses profilok megfelelőek az egyfogalmú kérdések teljes maghatározására, és a többfogalmú kérdésekhez kifejezéspáros profilok szerkesztése szükséges. Az eredmény meglepően nem pontosan így alakult. Egyfogalmú kérdésekre a fenti állítás beigazolódott; a manuális

*1. táblázat*

**Automatikusan és manuálisan szerkesztett profilok keresési eredményeinek összefoglalása**

| | Egykifejezés | Egykifejezés/ kifejezéspár | Manuális | Egyfogalmú – egykifejezéses | Egyfogalmú – egykifejezéses/ kifejezéspáros | Egyfogalmú – manuális | Többfogalmú – egykifejezéses | Többfogalmú – egykifejezéses/ kifejezéspáros | Többfogalmú – manuális | Egykifejezéses/ szótöredék | Kifejezéspár | Egykifejezéses | Egykifejezés/ kifejezéspár | Manuális |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Átlagos találatszám | 81,8 | 72,8 | 90,7 | 95,6 | 79,8 | 93,0 | 80,0 | 71,9 | 90,4 | 61,8 | 15,9 | 35,1 | 30,5 | 42,2 |
| Átlagos pontosság, % | 24,1 | 24,1 | 36,3 | 39,2 | 40,3 | 61,5 | 22,0 | 21,9 | 32,9 | 12,2 | 23,8 | 22,7 | 19,4 | 32,5 |
| Átlagos lehívás, % | 57,8 | 56,4 | 76,5 | 79,3 | 75,7 | 85,1 | 56,3 | 54,9 | 76,3 | 53,3 | 30,5 | 37,7 | 49,5 | 63,6 |
| Átlagos keresési idő profilonként, sec | 5,4 | 229,4 | 151,1 | 6,1 | 157,3 | 218,4 | 5,3 | 239,0 | 142,1 | 7,1 | 338,4 | 5,0 | 123,4 | 58,3 |
| Összes találatszám | 5565 | 4952 | 6165 | 707 | 638 | 744 | 4858 | 4314 | 5421 | 618 | 159 | 351 | 305 | 422 |
| Összesített pontosság, % | 25,6 | 27,3 | 31,0 | 36,5 | 39,7 | 39,9 | 24,0 | 25,5 | 29,8 | 14,4 | 16,4 | 18,5 | 19,3 | 24,6 |
| Összesített lehívás, % | 61,3 | 58,1 | 82,1 | 76,3 | 74,9 | 87,9 | 58,7 | 55,3 | 81,2 | 64,0 | 18,7 | 46,8 | 42,2 | 74,8 |

profilok ilyenkor kissé jobbak, de az automatikus profilok keresési ideje rövidebb. Azonban többfogalmú kérdések esetén az egykifejezéses profilok és az egykifejezéses/kifejezéspáros profilok eredménye alig tér el egymástól; a manuális profilok itt is kissé jobb eredményre vezetnek, de az automatikus profilok gyorsabban futtathatók.

Tehát ez esetben is beigazolódott a kifejezéspárok viszonylagosan kisebb hatékonysága; bizonyos profilok szerkesztéséhez azonban feltétlenül szükséges ezek alkalmazása.

Sok automatikus profil megtalált olyan új releváns dokumentumokat is, amelyeket a manuális profilok nem hoztak ki. Ennek két oka van: a specifikussági lista készítésében használt ismétlések során olyan új keresőszavak is előkerültek, amelyek a manuális profilokból hiányoztak; a manuális profilok komplex logikája miatt maradhattak ki egyes hivatkozások.

Az összes releváns találat (2327) közül 415 dokumentumot csak az automatikus profilok alapján találtak meg, tehát a hagyományos módszerrel a potenciális találatok mintegy 1/5-e nem volt elérhető. Ugyanakkor viszont az automatikus profilokkal 462 olyan találat nem jelent meg, amelyek a manuális profilokkal kikereshetők voltak.

### *Következtetések*

Az automatikusan szerkesztett profilok alkalmazásával nyert eredmény összemérhető a manuális profilokéval. *A leghatékonyabbak azok, amelyek főleg egykifejezéses fogalmakból állnak és szótöredékeket is tartalmaznak.* A kifejezéspárokat tartalmazó profilok általában (de nem mindig!) kevésbé eredményesek, mint az egykifejezéses profilok: az utóbbiak költséghatékonysági mutatója is jobb.

Az ismertetett eljárás *on-line* alkalmazást feltételez, annak ellenére, hogy a kíséreltek során *batch* feldolgozást végeztek. Az automatikus profilszerkesztési eljárás ugyanis úgy kezdődik, hogy a felhasználó néhány szót (vagy akár egy szót) visz be terminálon keresztül a rendszerbe. Az így kapott találatokat relevancia szempontjából értékeli, ennek alapján a rendszer automatikusan kiszámítja minden kifejezés specifikussági értékét, majd újabb találatlista jelenik meg. Ezt addig ismétlik, amíg egy elfogadható egykifejezéses specifikussági lista nem jön létre (a határértéket a felhasználó állapítja meg iterációs lépésenként). A listából – kívánságra – kifejezéspár- és szótöredék-listák is készülnek. A specifikussági listákból az automatikus profilt a lehívás, a pontosság, a specifikusság értéke és a kumulativ output számok alapján a felhasználó utasításai szerint készíti el. Ez még tovább finomítható új specifikussági érték megállapításával (valamely másik helyzet a pontosság-lehívás görbén), továbbá új logikai operátor (pl. NEM) bevezetésével.

*/ROBSON, A. – LONGMAN, J. S.: Automatic aids to profile construction. = Journal of the American Society for Information Science 27. köt. 4. sz. 1976. p. 213–223./*

(Roboz Péter)

## Könyvtári számítógép-terminál kiválasztása

A kereskedelemben igen sokféle terminál kapható. Leggyakrabban aszerint osztályozzák őket, hogy képesek-e az eredetiről papírmásolatot (hard copy) szolgáltatni, vagy pedig vizuális megjelenítők-e. Az alkalmazott kódrendszer szerint is két csoportra oszthatók: vagy az *ASCII (American Standard Code for Information Interchange = Amerikai Szabványkód Információcsere Céljára)*, vagy az *EBCDIC (Extended Binary Coded Decimal Interchange Code = Kiterjesztett Kettőskódolású Tizedes Csere Kód)* kóddal kompatibilisek. A lassabbak tíz, a leggyorsabbak 150 karaktert képesek átvinni másodpercenként.

A könyvtári terminál kiválasztásakor az alábbi tényezőket célszerű figyelembe venni.

### *Az output*

*Vizuális megjelenítés elég-e vagy szükség van papírmásolatra is?* Az utóbbi esetben figyelembe kell venni, hogy egyes terminálok tetszőleges minőségű papírra másolnak, míg sok típusnál speciális – pl. hőre érzékeny – papír szükséges, ami nagy mennyiségben igen drága lehet. Aránylag kevés papírmásolat esetén a katódsugaras (CRT) terminálhoz kapcsolható olcsóbb másoló alkalmazása ajánlatos;

a megfelelő *karakterkészlet, betűtípusok* kiválasztása;

*olvashatóság, képernyőméret.* Kis képernyő nem képes egy soktételes bibliográfia egyidejű megjelenítésére. Ezen ún. *„görgető egység"* beépítésével lehet segíteni, ami lehetővé teszi, hogy a képernyő alján megjelenő új információ a felül levőt úgy iktassa ki, hogy azt kívánságra vissza lehessen hozni; a pozicionáló segítségével a képernyő bármely pozíciójában javítás végezhető stb.

### *Kompatibilitás*

Adott célú felhasználásra, pl. on-line bibliográfiai keresésre, csak bizonyos típusú terminálok felelnek meg. Az adatbázis-szolgáltató ezért mindig közli a *felhasznál-*