Az indexelés minősége a könyvtár- és információtudományi adatbázisokban

Az indexelés célja, hogy lehetővé tegye a dokumentumok tartalmuk szerinti visszakeresését. A visszakeresés hatékonysága az indexelés minőségétől függ. A szakirodalom különféle módszereket ír le ennek megállapítására. A használói megelégedettségen és a relevancia megítélésén alapuló módszerek szubjektív elemeket tartalmaznak; másoknál, amelyek az indexelő vagy az indexelési rendszer teljesítményét azzal mérik, mennyire képes a tárgyszavakat úgy megválasztani, hogy a dokumentumról a lehető legtöbb információt nyújtsa, nehéz a gyakorlati értékeket kiszámítani.

White és *Griffith** eljárása kiküszöböli a fenti hátrányokat. Módszerükkel sikerül kimutatni, hogy az adatbázisok ellenőrzött szótárai mennyire képesek (1) összekapcsolni az összetartozó dokumentumokat, (2) nagy vonalakban megkülönböztetni ezen dokumentumok halmazait a teljes fájlban, (3) pontosan megkülönböztetni az egyes dokumentumokat. A vizsgált adatbázisok mindegyikében szereplő dokumentumok tesztelési csoportjainak segítségével lehet összehasonlítani *az indexelés minőségét* a fenti kritériumok alapján. Az első kritérium (a *következetesség* mérve) megállapítja, hogy az adatbázis milyen indexkifejezésekkel írja le az egyes csoportok dokumentumainak közös tartalmát, és sikerül-e így összekapcsolni az összetartozó dokumentumokat. A második kritérium (az *összefogás-megkülönböztetés* mérve) szerint az indexkifejezések akkor különböztetik meg nagy vonásokban az összetartozó dokumentumok csoportjait, ha az illető csoport dokumentumainak legalább felére vonatkoznak, és ritkán használják őket az adatbázisban. Végül az egyes dokumentumok pontos megkülönböztetésének mértéke (a *részletezés* mérve) a leírásukhoz felhasznált indexkifejezések átlagos számával állapítható meg (minél kimerítőbb az indexelés, annál könnyebben hozzáférhető a dokumentum).

Ez a tanulmány a könyvtár- és információtudományi szakirodalom indexelésének minőségét vizsgálja a *Library and Information Science Abstracts (LISA)* 1969– 1987-es (83 450 tétel), a *Library Literature (LL)* 1984– 1987-es (31 000 tétel) és az *Information Science Abstracts (ISA)* 1966– 1987-es (119 400 tétel) évfolyamai alapján az *1. táblázatban* megadott témakörökben.

A módszer

A tesztelési dokumentumcsoportokat többféleképpen lehet kialakítani. Itt a tartalmilag összefüggő dokumentumokat szakértők és a szerzők válogatták ki. Hogy ne vádolhassák őket elfogultsággal, nem az egyik vagy másik adatbázisból indultak ki, hanem a felsőoktatási kötelező és ajánlott olvasmányok listáit használták fel. Eredetileg 12 csoportot, egyenként legalább öt dokumentummal kívántak felállítani, de minthogy a kiválasztott dokumentumok közül nem mindegyik volt meg mindhárom adatbázisban, meg kellett elégedni 9 csoporttal, s bennük 3– 7 dokumentummal. Ez azonban nem befolyásolta a végeredményt.

1. táblázat
Az indexelés kiértékelésének témakörei

Könyvtár- és információtudomány			
Könyvtártudomány		Információtudomány	
Katalogizálás	Tájékoztató szolgálat	Bibliometria	Indexelés
Mikroszámítógépes szoftverek katalogizálása	Tájékoztatási szolgáltatások kiértékelése	Idézetelemzés	Automatizált indexelés
Sorozati kiadványok katalogizálása	Bibliográfia-használati oktatás	Kölcsönzések elemzése	Az osztályozás használata az online keresésben
	Online tájékoztatási szolgáltatások		

Miután minden tételt azonosítottak az adatbázisokban, lehívták a hozzájuk tartozó valamennyi indexkifejezést. Ezeket betűrendbe rakták, hogy mindegyik csak egyszer szerepeljen. Majd megállapították, hogy az egyes adatbázisokban hányszor rendelték hozzá e kifejezéseket az összes feldolgozott dokumentumhoz. Ez a vizsgálat – tekintet nélkül a három adatbázis eltérő gyakorlatára – indexkifejezésnek tekintett minden tárgyszót, amelyekkel az indexelők a dokumentumok tartalmát leírták. (Így a tezaurusz-relációkat – pl. szinonimákat – is önálló indexkifejezéseknek vették.)

A dokumentumok tartalmi hasonlóságára utaló kifejezések kiszűrése érdekében megnézték, melyek azok, amelyek két vagy több dokumentumot fogtak össze, és megállapították e kifejezések előfordulásának gyakoriságát a csoporton belül. (Azokat a kifejezéseket, amelyek csak egy-egy dokumentumnál fordultak elő, figyelmen kívül hagyták.)

A következő lépés az volt, hogy megvizsgálják, mennyire képesek megkülönböztetni e kifejezések az egyes csoportokat. Ha az adatbázisokban túl sok dokumentumhoz lettek hozzárendelve, túlzott, ha túl

* WHITE, H. D.– GRIFFITH, B. C.: Quality of indexing in online data bases. = Information Processing & Management, 23. köt. 1987. p. 211– 224.

kevéshez, csekély a megkülönböztetés mérve. A megkülönböztetés mérvének kifejezésére két eljárást is alkalmaztak a szerzők: a White és Griffith által kidolgozottat (A típusú) és a sajátjukat (B típusú).

A megkülönböztetési mutatót az A eljárás szerint a következő képlettel lehet kiszámítani:

$$\text{Megkülönböztetési mutató}_{\text{"A" kifejezés}} = 1/\log_{10} \text{Előfordulások száma az adatbázisban}_{\text{"A" kifejezés}}$$

A mutató értéke 0 és 1 között változik; 0,25 a megkülönböztetési küszöb, s a magasabb érték jobb megkülönböztetést jelent. Ez az eljárás azonban nem számol azzal, hogy az adatbázisok nagysága eltorzítja a mutatót, továbbá esetenként 1-nél nagyobb értékek is adódhatnak a mutatóra.

A B eljárás szerint a következőképpen számítható ki a megkülönböztetési mutató:

$$\text{Megkülönböztetési mutató}_{\text{"A" kifejezés}} = \text{Előfordulások száma az adatbázisban}_{\text{"A" kifejezés}} / \text{Az adatbázis nagysága}$$

A mutató értéke itt is 0 és 1 között változik, de az alacsonyabb érték jelent jobb megkülönböztetést.

Mindkét mutatót kiszámították minden indexkifejezésre, amely két vagy több dokumentumot fogott össze egy-egy csoportban (lásd *2. táblázat).* Az A mutató értéke 0,23 és 0,81 között változik az egyes adatbázisokban, s a mindháromra vonatkozó átlagos értéke 0,35 (egyenként: LL – 0,57; LISA – 0,32; ISA – 0,34). Az LL-é a legjobb, de nem szabad elfelejteni, hogy ez egy kicsi adatbázis. A B mutató értéke 0,0001 és 0,339 között változik az egyes adatbázisokban; a mindháromra vonatkozó átlagos értéke 0,039 (LL – 0,003; LISA – 0,06; ISA – 0,02). A 0,05-nek vett küszöbértékhez képest az LL túl aprólékosan, a LISA jól különbözteti meg az egyes témaköröket.

2. táblázat

Két vagy több dokumentumot összefogó kifejezések a "kölcsönzések elemzése" csoportban

LL	LISA	ISA
Kölcsönzések elemzése [4] (0,61; 0,001)	Matematikai modellek [2] (0,41; 0,003)	Könyvtári és információs szolgáltatások – Kölcsönzések nyilvántartása [3] (0,50; 0,001)
	Technikai eljárások és szolgáltatások [3] (0,26; 0,1)	Kölcsönzés [4] (0,31; 0,01)
	Szolgáltatások [2] (0,23; 0,34) Olvasói szolgáltatások [2] (0,26; 0,08) Kölcsönzés [2] (0,31; 0,2)	

[ ] = Az összefogás gyakorisága.
(A, B) = Az A és a B típusú megkülönböztetési mutató értéke.

Az indexelés minőségének megállapítására az összefogás-megkülönböztetés együttes mércéjével vizsgálták meg az indexkifejezéseket, méghozzá kétféleképpen: összefogó-megkülönböztető kifejezések azok, amelyek a csoportba tartozó dokumentumoknak legalább a felére érvényesek és (1) az A típusú megkülönböztetési mutató értéke 0,25 fölött, a B típusúé 0,05 alatt van; (2) az A típusú megkülönböztetési mutató értéke 0,25 és 0,75, a B típusúé pedig 0,001 és 0,05 közé esik.

Végül a kielégítő indexelés megállapítására kiszámították a részletezés mérvét is az egyes dokumentumokra vonatkozóan.

Az eredmények

A *3. táblázat* foglalja össze a különféle mércék szerint elvégzett kiértékelés eredményeit. A 9 oszlop egy-egy tesztelési szempontot jelent, a következők szerint:

1. *oszlop:* azoknak a kifejezéseknek a száma, amelyek egy-egy csoport minden dokumentumát összefogják.
2. *oszlop:* azoknak a kifejezéseknek a száma, amelyek egy-egy csoport dokumentumainak felét (vagy ennél többet) összefogják.
3. *oszlop:* azoknak a kifejezéseknek a száma, amelyek összefogják egy-egy csoport dokumentumainak felét (vagy többet), és A típusú megkülönböztetési mutatójuk 0,25-nél magasabb.
4. *oszlop:* azoknak a kifejezéseknek a száma, amelyek összefogják egy-egy csoport dokumentumainak felét (vagy többet), és B típusú megkülönböztetési mutatójuk 0,05-nél alacsonyabb.
5. *oszlop:* azoknak a kifejezéseknek a száma, amelyek megfelelnek az összefogás kritériumainak, és az A típusú megkülönböztetési mutatójuk 0,25 és 0,75 közé esik.
6. *oszlop:* azoknak a kifejezéseknek a száma, amelyek megfelelnek az összefogás kritériumainak, és a B típusú megkülönböztetési mutatójuk 0,001 és 0,05 közé esik.
7. *oszlop:* az egyes csoportokban szerepő dokumentumok száma.
8. *oszlop:* az egyes csoportokban szereplő dokumentumok indexelésére használt kifejezések száma.
9. *oszlop:* az egyes csoportokban szereplő dokumentumok indexelésére használt kifejezések dokumentumonkénti átlagos száma.

A kiértékelésben a LISA került az élre, mivel az 1., a 2., a 3. és az 5. oszlop kritériumai szerint a legjobb értéket mutatja fel, s dokumentumonként átlagosan a

3. táblázat
A könyvtár- és információtudományi adatbázisok indexelésének összehasonlítása az egyes dokumentumcsoportokban

Csoport	Adatbázis	1	2	3	4	5	6	7	8	9
Katalogizálás										
Mikroszámítógépes	LL	1	2	2	2	1	1	4	7	1,75
szoftverek	LISA	7	12	12	8	12	7	4	38	9,5
katalogizálása	ISA	3	5	4	4	4	4	4	23	5,75
Sorozati	LL	1	2	2	2	2	2	6	21	3,5
kiadványok	LISA	5	5	5	0	5	0	6	37	6,17
katalogizálása	ISA	2	4	4	4	4	4	6	25	4,17
Tájékoztatás										
Tájékoztatási	LL	0	0	0	0	0	0	4	7	1,75
szolgáltatások	LISA	0	4	3	2	3	2	4	43	10,75
kiértékelése	ISA	0	4	4	3	4	2	4	21	5,25
Bibliográfia-	LL	0	1	1	1	1	1	6	18	3
használati	LISA	0	4	4	1	4	1	6	58	9,67
oktatás	ISA	0	1	1	1	1	1	6	34	5,67
Online	LL	0	1	1	1	1	1	6	17	2,83
tájékoztatási	LISA	0	3	2	1	2	1	6	56	9,33
szolgáltatások	ISA	0	2	2	2	2	1	6	37	6,17
Bibliometria										
Idézetelemzés	LL	1	1	1	1	1	1	7	19	2,71
	LISA	3	5	5	4	5	4	7	69	9,86
	ISA	0	1	1	1	1	1	7	45	6,43
Kölcsönzések	LL	1	1	1	1	1	1	4	10	2,5
elemzése	LISA	0	5	4	2	4	2	4	32	8
	ISA	1	2	2	2	2	1	4	20	5
Indexelés										
Automatizált	LL	1	1	1	1	1	1	3	9	3
indexelés	LISA	4	5	5	1	5	1	3	20	6,67
	ISA	0	3	3	3	3	3	3	14	4,67
Az osztályozás használata	LL	0	2	2	2	1	1	3	7	2,33
az online	LISA	4	9	9	4	9	4	3	27	9
visszakeresésben	ISA	1	6	6	6	6	6	3	20	6,67

legtöbb kifejezést (8,8) használja a tartalom leírására. Az ISA a 4. és a 6. oszlopban foglalt kritériumok szerint kissé megelőzi a LISA-t (lásd az adatbázisok nagyságából adódó eltéréseket az A és a B típusú mutatóban). Az LL minden szempont szerint gyenge eredményt ért el, kivéve a 3. és 5. oszlop tesztjében, de ez főként kicsinységének következménye.

A LISA indexelése volt a legjobb mind a könyvtár-, mind az információtudomány területén. Egyik adatbázisnál sem lehetett azonban megállapítani, hogy indexelése jobb lett volna egyik vagy másik nagy területen.

A szerzők nem tettek kísérletet a kifejezések minőségi összehasonlítására valamely indexkifejezés preferálása érdekében. Ezt az olvasó maga is megteheti a függelékként csatolt táblázatok áttanulmányozásával. E táblázatok részletezik az egyes csoportok indexkifejezéseit, a kifejezések mutatóinak értékeit, s felsorolják a csoportba tartozó dokumentumokat.

/CHU, C. M.– AJIFERUKE, I.: Quality of indexing in library and information science databases. = Online Review, 13. köt. 1. sz. 1989. p. 11– 35./

(Papp István)