

Videotex a mezőgazdaságban

Jelenleg hat mezőgazdasági szolgáltatás van a Prestel rendszeren és kb. kétszer ennyi olyan szolgáltatás, amely hasznos információt tartalmaz a mezőgazdaság számára (pl. pénzügyek, banki információk) is. A Prestel 1984 áprilisa óta üzemelteti a *Farm-hálózat* (Farm-link) rendszerét. 700 előfizetője van. Tervezik, hogy 1985 őszére egész Angliában és Walesben hozzáférhető lesz. Amellett, hogy előfizetőinek híreket és információkat közvetít a piaci árakról, az időjárásról, a termékekről stb., felkínálja az *ADÓ* (PAYE) programjait, árurendelő és üzenetközvetítő szolgáltatást is.

A videotex fejlesztésében jelentős szerepet játszott az ICI Mezőgazdasági Részlege. Az itt dolgozó szakemberek oldották meg a tv-készülékek telefonvonalas csatlakozását az ICI központi számítógépéhez.

Az *AGVISER* szolgáltatás több szinten épült ki. Az első szint csak egy alapterminált és egy modemet igényel. Segítségével a következő információs forrásokhoz lehet csatlakozni: időjárás-előrejelzés, piacok, tenyészállatok kínálata, termelőterülettel

való gazdálkodás, üzenetközvetítés. A második szinthez már mikroszámítógép kell. Ez a szint széles választékát kínálja azoknak a szoftvereknek, amelyeket le lehet hívni a központi gépről és használni lehet a gazdálkodási adatok kezelésére, fizetési jegyzékek elkészítésére vagy szövegszerkesztésre.

Az ICI Növényvédelmi Részlege (ICI Plant Protection Division) ugyancsak újdonságnak számító fejlesztést hajtott végre a videotex és a számítógép együttes alkalmazásával. Nem egészen egy év alatt hozta tető alá a *COUNSELLOR* rendszert, amely döntés-előkészítést végez a gabonafélék betegségmegelőzésében. Egyelőre belső rendszerként működik, mivel közreadásra még nem teljesen "érett". Ez a könnyen kezelhető faktografikus tanácsadó rendszer képviseli a végső formát a nem információs szakemberek számára készült mezőgazdasági információs rendszerek között.

/HOEY, P. O'N.: Use of information technology in UK agricultural information work: a review. = *Journal of Information Science*, 11. köt. 1. sz. 1985. p. 9–17./

(Szabó Sándor)

A teljesség és a pontosság vizsgálata teljes dokumentumszövegek visszakeresésében

A teljes szövegű információkeresés azt jelenti, hogy az online hozzáférhető adatbázis teljes dokumentumszövegeket tárol és tesz hozzáférhetővé párbeszédes üzemmódban. Az adatbázisok többsége ma még a bibliográfiai adatok mellett csak rövidebb-hosszabb kivonatokat, referátumokat tartalmaz, s ezek mutatnak rá a teljes primer dokumentumra, amelyet — ha szükség van rá — külön lépésben kell beszerezni. A teljes szövegű adatbázisok azonkívül, hogy a primer dokumentumokat teljes egészükben szolgáltatják, lehetővé teszik a teljes szövegben (vagy annak kiválasztott részeiben) való keresést is.

A teljes szövegben végzett információkeresés hatásfokának mértékét — a többi információkereső rendszerhez hasonlóan — két mennyiség, a teljesség (T) és a pontosság (P) mutatja. Az előbbi azt jelenti, hogy egy adatbázisban meglevő releváns dokumentumok hány százalékát sikerült visszakeresnünk, vagyis

$$T = \frac{\text{a visszakeresett releváns dokumentumok}}{\text{az adatbázisban levő összes releváns dokumentum}} \cdot 100\%$$

A másik mutató, a pontosság azt jelenti, hogy az online keresés során visszakeresett összes dokumentum hány százaléka releváns, vagyis

$$P = \frac{\text{releváns dokumentumok}}{\text{visszakeresett összes dokumentum}} \cdot 100\%$$

A teljes szövegű adatbázisokból végzett online keresések T és P értékének meghatározásával nem sokan próbálkoztak eddig. A következőkben egy ilyen vizsgálatot ismertetünk.

A vizsgálat körülményei

A kísérletben szereplő jogi adatbázis 40 ezer dokumentumot tartalmaz, amelyek összesen 100 ezer online kereshető és megjeleníthető szövegoldal tesznek ki. Az adatbázis — nevével ellentétben — nem tartalmazza a dokumentumok teljes szövegét, hanem csak azok válogatott részeit, amelyek a szerkesztők szerint az eredeti dokumentumok legfontosabb mondanivalóját tartalmazzák, tömörítvények-

nek foghatók fel. A tömörítés mértékére jellemző, hogy a 40 ezer dokumentum teljes terjedelmében összesen 350 ezer oldalt tesz ki.

A kísérletre az IBM jól ismert információkereső szoftverjét, a STAIRS-t, pontosabban annak egy módosított változatát használták. A STAIRS rendszer lehetővé teszi a bevitt keresőszavak összekapcsolását Boole-féle logikai operátorokkal (AND, OR, NOT), megszabhatjuk továbbá a dokumentumszövegben egymáshoz viszonyított helyzetüket is, pl. szomszédos szavak, egy mondatbeli szavak. Meghatározott adatmezőben előforduló szavakra is rá lehet irányítani a keresést (pl. csak a szerzőmezőben, csak a dátumban előforduló szavakra, számokra). A STAIRS lehetőséget nyújt a visszakeresett dokumentumok rendezésére valamely szempont szerint, pl. a dátum növekvő vagy csökkenő rendje szerint, vagy a szerzőnév ábécérendje szerint. 200-nál kevesebb visszakeresett dokumentum esetén kiválasztott keresőszavak előfordulási gyakorisága szerint is rangsorba állíthatók a dokumentumok, tehát egyetlen utasítással megnézhetjük, hogy mely dokumentumok tartalmazzák 10-szer, 9-szer, 8-szor stb. pl. azt a szót, hogy "adózás". Végül a módosított STAIRS képes egy tezauruszt is online kezelni, ami lehetővé teszi a deskriptorok közötti szemantikai összefüggések megismerését és alkalmazását a kötöttszavas keresés szélesítésére, ill. finomítására.

A kísérlet végrehajtása

A teljes szövegű jogi adatbázis online keresése teljességének és pontosságának kísérleti vizsgálatához jogászokat kértek fel, akik a visszakeresett dokumentumokat (pontosabban: a teljes dokumentumok tömörítvényeit) peres ügyükhöz használták fel. Ők akkor hasznosíthatják az online keresés eredményét ebből az adatbázisból, ha a visszakeresés teljessége 75%-os a releváns dokumentumoknál, de legalább 10%-os az alapvetően fontos dokumentumoknál. A dokumentum minősítésére négy kategóriát használtak:

- ◆ alapvető,
- ◆ megfelelő,
- ◆ határeset,
- ◆ nem releváns dokumentum.

A kísérletben az adatbázist ugyanúgy kellett használni, mint egy valószínű peres ügyben. A keresés szabadszavas volt, vagyis nem deskriptorokon, hanem a jogi gyakorlatban elterjedt terminológián alapult. A kísérletben két ügyvéd és két, a STAIRS-ben jártas jogász gyakornok vett részt. Összesen 51 kérdést tettek fel, amelyeket a gyakornokok alakítottak át online keresőkérdésekké, majd elvégezték

a keresést. A keresés eredményét, a találatokat a ki-nyomtatott teljes dokumentumok formájában átadták az ügyvédeknek, akik a fenti négy kategória egyikével minősítették ezeket aszerint, hogy mennyiben felelnek meg eredeti kérdésüknek, ill. mennyiben támogatják az aktuális peres ügyet. Ezután a két ügyvéd — ha szükségesnek látta — a keresés további finomítását kérte a gyakornokoktól, ami további releváns dokumentumok visszakeresését eredményezte. Ezt addig folytatták, amíg az ügyvédek ki nem jelentették, hogy a kikeresett dokumentumokkal elégedettek, kérdéseiket megválaszolni tekintik. Fontos mozzanat, hogy az ügyvédek és a gyakornokok egymással mindaddig folytathatták a "párbeszédet", amíg szükségesnek tartották, miközben az utóbbiak a keresési stratégiát egyre tovább pontosították, az ügyvédek pedig újabb és újabb találathalmazokat kaptak kiértékelésre.

Ezután kiszámították a P-t, úgy, hogy a relevánsnak minősített (a négy kategória közül a három elsőbe eső) dokumentumok teljes számát elosztották az összes visszakeresett dokumentum számával. Az első keresés és a további finomítások során kapott találatokat összesítették, és a többször is visszakeresett azonos dokumentumokat csak egyszer vették figyelembe a számításnál.

A T kiszámítása nehezebb, mert ehhez ismerni kellene a vissza nem keresett releváns dokumentumokat. Ezt úgy oldották meg, hogy az adatbázisnak azokból a rekordjaiból (dokumentaiból), amelyek nem voltak találatok a keresés egyik lépésében sem, a kísérlet szervezői különböző halmazokat képeztek. A halmazokat úgy állították össze, hogy mindegyikükben legyen relevánsnak ítélt dokumentum is. Ezekből a dokumentumhalmazokból véletlen mintákat vettek, amelyeket az ügyvédeknek átadtak értékelésre. Az ügyvédek nem tudták, hogy most egy ilyen, nem találati rekordokból álló halmazt kaptak, vagy a keresési lépések valamelyikében előállított találati rekordok halmazát. A nem találati minták alapján becsült, a teljes adatbázisra extrapolált releváns dokumentumok száma jelenti a T számítási képletében a nevezőt. Mivel a teljes adatbázisra való extrapoláció csak maximált lehet, a T is egy maximumértéknek adódik minden információs kérésre.

A kísérlet eredménye

Az 51 kérdés közül 40-re számították ki a T-t és a P-t, a többi 11 a mintavételes módszer ellenőrzésére szolgált. Az eredményeket az 1. táblázat mutatja.

A táblázatból látható, hogy a P 19,6% és 100% között változik, az átlag 79%, a hozzá tartozó szórás

1. táblázat
40 keresőkérdésre végzett online információkeresés
teljessége és pontossága

Keresőkérdés sorszáma	Teljesség	Pontosság
1	*	*
2	45,5%	92,6%
3	*	*
4	*	*
5	*	*
6	8,9	60,0
7	20,6	64,7
8	43,9	88,8
9	13,3	48,9
10	10,4	96,8
11	12,8	100,0
12	9,6	84,2
13	15,1	85,0
14	78,7	99,0
15	*	*
16	*	*
17	*	*
18	13,0	38,0
19	15,8	42,1
20	19,4	68,9
21	41,0	33,8
22	22,2	94,8
23	2,8	100,0
24	*	*
25	13,0	94,0
26	7,2%	95,0%
27	50,0	42,6
28	50,0	19,6
29	*	*
30	7,0	100,0
31	*	*
32	12,5	100,0
33	18,2	79,5
34	14,1	45,1
35	*	*
36	4,2	33,3
37	15,9	81,8
38	24,7	68,3
39	18,5	83,3
40	4,1	100,0
41	18,3	96,9
42	45,4	91,0
43	18,9	100,0
44	10,6	100,0
45	20,3	94,0
46	11,0	85,7
47	13,4	100,0
48	13,7	87,5
49	17,4	87,8
50	13,5	75,7
51	4,7	100,0

Átlagos teljesség = 20% — (Szórás = 15,9)
Átlagos pontosság = 79% — (Szórás = 23,2)

23,2. Vagyis 40 keresőkérdés átlagában 100 dokumentum közül 79 volt releváns (alapvető vagy meg-

felelő, vagy határeset). A T maximuma 78,7%, minimuma 2,8%, az átlag 20% (15,9-es szórással). Vagyis 40 kérdés átlagában az adatbázisban meglevő minden 100 releváns dokumentum közül csak 20-at kaptak meg az ügyvédek (miközben úgy gondolták az alapfeltevés szerint, hogy a releváns dokumentumok 75%-a vissza lett keresve).

Egyéb számításokat is végeztek az online információkeresés hatásfokának becslésére. Így a kísérlet eredményét meghatározták külön a két ügyvédre:

	Átlagos T	Átlagos P
1. ügyvéd	22,7%	76,0%
2. ügyvéd	18,0%	81,4%

Megnézték azt is, hogy változik-e a T, ha nem a közvetítők (a jogász gyakornokok), hanem maguk a végfelhasználók (az ügyvédek) végzik a keresést a terminálnál. Ehhez taláalomra kiválasztottak öt kérdést, és a keresést az ügyvédekkel végeztették el, akik a képernyőre kiírt találatok alapján interaktív módon módosították a keresési stratégiát. A kísérlet korábbi szakaszából már ismert volt a kérésre kapható minimális számú releváns válasz az adatbázisban, ennek és az ügyvédek által visszakeresett találatok alapján ismét kiszámították a T-t, és összehasonlították a gyakornokok keresésére meghatározott T értékkel. Az öt kérdés közül háromra az ügyvédek keresése adott nagyobb T-t, kettőre fordítva. Tehát öt kérdés átlagában nincs lényeges eltérés a végfelhasználó és a közvetítő által végzett információkeresés T értékében.

A T és P értéke eltérő, ha a relevanciaértékelés különböző szintjeit tekintjük. A három kategória: alapvető (A), megfelelő (M) és határeset (H) szerint minősített dokumentumok esetében a teljesség és pontosság átlagértékei 40 dokumentumra a következőknek adódtak:

A relevancia-minősítés esetei	T	P
A+M+H	20,0%	79,0%
A+M	25,3%	56,6%
A	48,2%	18,2%

Látható, hogy a T növekedésével a P csökken, megfelelve a teljesség és a pontosság jól ismert inverz összefüggésének az információtudományban, ami a visszakeresés gyakorlatában szükséges állandó kompromisszumok forrása: minél nagyobb a teljesség, annál kisebb a pontosság és fordítva.

Értékelés

Mint a kísérlet eredménye tanúsítja, a tipikus online keresés — teljes szövegű adatbázis szabad-szavas keresése esetében — az adatbázisban meglevő minden öt releváns dokumentum közül csak

egy dokumentumot eredményez (a teljesség átlagban 20%). Ez a tény az online rendszereket használók számára meglepő lehet, mivel a felhasználó csak a visszakeresett dokumentumok halmazát látja, nem az összeset, továbbá a visszakeresett halmazban is a legtöbbet relevánsnak látja. Automatikusan a pontosságot ítélik meg – mint a visszakeresés jóságának mutatóját –, a teljességről hajlamosak elfeledkezni.

Sokan úgy érvelhetnek, hogy a vizsgálat eredménye nem általánosítható, legfeljebb ebben a vizsgált adatbázisban lehet igaz. Valószínűleg azonban nem így van. A teljes szövegű adatbázisokban való online keresés egyszerűsége abban rejlik, hogy – az általános hiedelem szerint – a dokumentumokban levő közismert szavak és szakkifejezések alapján igen nagy biztonsággal kereshetők vissza a releváns dokumentumok. Ennek a feltételezésnek az a hibája, hogy a közhasználatban elterjedt szakkifejezések nagy része, amelyről a kereső úgy véli, hogy a releváns dokumentumokban feltétlenül benne van, számos nem releváns dokumentumban is szerepel. A tökéletes keresési stratégiának csak azokra a szakkifejezésekre kellene korlátoznia a keresést, amelyek a releváns dokumentumokban megvannak, de a nem relevánsokban nincsenek meg. Ez azonban lehetetlenül nehéz stratégia lenne, a természetes nyelv nagy hajlékonysága miatt.

A vizsgálat egy másik fontos tanulsága, hogyan kell felépíteni egy adatbázist a jó minőségű visszakeresés célját szem előtt tartva. Az világossá vált, hogy az igen egyszerű szabadszavas visszakeresés nem vezet jó eredményre a teljes szövegű adatbázisok keresésében, elsősorban a teljesség vonatkozásában. Ha indexelést is alkalmaznak az adatbázis-

ban, akkor erre nem megfelelőek egyedül a dokumentum szövegéből vett kulcsszavak. Mivel a releváns dokumentumok 80%-át (átlagban) nem sikerült visszakeresni, ezek nyilván nem tartalmazták a keresésre eredetileg kidolgozott kulcsszavakat, szókapcsolatokat, pedig a vissza nem keresett 80% is arról a témáról szólt, amire a kereső kíváncsi volt. Az automatikus indexelés kizárólag a dokumentumokban meglevő szavakon alapul, ezért nem biztos, hogy így eredményes és jó minőségű keresés jöhet létre. Minthogy az egyszerű szabadszöveges, szabadszavas keresés a vártnál jóval gyengébb eredményre vezetett, azt a következtetést vonhatjuk le, hogy a szövegből vett szavakon alapuló automatikus indexelés is a vártnál rosszabb visszakeresési mutatókat eredményezne, mind a T, mind a P vonatkozásában.

Végül: egy nagy adatbázis indexelési rendszeréről, ill. hatékonyságát vélt keresési stratégiájáról addig nem mondhatjuk meg, hogy mennyire jó, amíg nem mértük meg egy megfelelően megtervezett és kiértékelt kísérlettel a keresés hatékonyságát. Az információkeresés egyik sajnálatos jellemzője, hogy bármely keresési eljárás vagy indexelési rendszer jóságát csak kis adatbázisok esetében tudjuk többé-kevésbé pontosan megítélni, nem gondosan szabályozott körülmények között. A szigorú kísérleti vizsgálatok költségei viszont igen nagyok, esetünkben csaknem félmillió dollárt tettek ki.

/BLAIR, D. C.: Full text retrieval: Evaluation and implications. = International Classification, 13. köt. 1. sz. 1986. p. 18–23./

(Roboz Péter)