

okok lehetnek összetett események, elemi okok Boole-algebrai kombinációi,

- ◆ rangsorolja a hipotéziseket a valószínűség mértéke alapján és attól függően, hogy milyen szabályok szerint állították fel őket, illetve, hogy az egyes szabályok alkalmazása során mennyi volt a lépések és az adatok száma,
- ◆ megállapítja azokat az eseteket, amikor az eseménytulajdonság párokra vonatkozóan két "konkurrens" hipotézis azonos valószínűséggel rendelkezik (adott esemény rendelkezik, ill. nem rendelkezik A tulajdonsággal),
- ◆ kiszámítja a kapott eredményt, és ennek alapján osztályozza az objektumokat.

A rendszerrel végzett eddigi kísérletek egytől-egyetig kedvezőek. Az adatbázisban szereplenie kell mind az események (objektumok), mind az eseményeket jellemző tulajdonságok halmazának. Az ese-

mények és tulajdonságok között létezniük kell empirikus összefüggéseknek. Emellett az adatbázisnak rejtve tartalmaznia kell bizonyos információt a tulajdonság- és eseményhalmaz közti korreláció meglétének, ill. hiányának okairól. Mindezeket a halmazokon meg kell adni azt a részlegesen (nem mindenütt) meghatározott kapcsolatot, amely az adatbázisban levő információ hiányos voltát tükrözi.

A kifejlesztett rendszer minden gyengén formalizált diszciplínában jól alkalmazható, vagyis olyan tudományterületen, ahol hiányzik az elméleti állítások, tételek szisztematikus axiomatizálása.

/GILÁREVSKIJ, R.S.: O značenii razrabotki ékspertnyh informacionnyh sistem v informatike = NTI, 2.Ser. 1984. 11.sz. p. 1-4./

(Környei Márta)

Az online és a hagyományos irodalomkeresés összevetése: a keresési stratégiát és a keresés határfokát vizsgáló kísérlet érdekes eredményei

Tíz évvel ezelőtt, amikor az online információkeresés még újdonság volt, az informatikai szakirodalom kedvencei közé tartozott egy téma gépi és kézi keresése eredményeinek összehasonlítása. A gépi módszerek azóta sokat fejlődtek, kiépültek a nagy, nemzetközi elérésű adatbázisok, így újra érdemes egy ilyen összehasonlítást elvégezni. Az összehasonlítás valós anyagon történt: Hollandia és Izrael közös vízhasznosítási programja keretében megjelent egy cikk az *Agricultural Water Management* c. folyóirat 1982. évi 5. számában, szerzői: Th. M. Boers és J. Ben-Asher, címe: *A review of rainwater harvesting* (Az esővízkitermelés áttekintése). A cikk 105 bibliográfiai hivatkozást tartalmaz, ezeket a szerzők hagyományos "kézi" kereséssel találták.

Egy későbbi időpontban gépi keresést végeztek ugyanerre a témára, megállapítandó, hogy melyek azok az irodalmi források, amelyeket a kézi módszerrel nem találtak meg. A gépi keresést úgy végezték el, hogy az egyes felhasznált adatbázisok tartalma és a keresés költségei összevethetőek, értékelhetőek legyenek.

A vizsgált adatbázisok

Hogy a lehető legteljesebb eredményre jussanak, minden olyan adatbázist bevontak a keresésbe,

amelyről gyanítható volt, hogy releváns anyagot tartalmaz, nevezetesen:

- ◆ WATER RESOURCES ABSTRACTS (WRA)
- ◆ AGRICOLA
- ◆ COMMONWEALTH AGRICULTURAL BUREAUX ABSTRACTS (CAB)
- ◆ AGRIS
- ◆ PASCAL (francia)
- ◆ AQUALINE
- ◆ BIOSIS
- ◆ COMPENDEX
- ◆ GEOARCHIVE
- ◆ GEOREF
- ◆ LIFE SCIENCE COLLECTION
- ◆ NTIS
- ◆ COMPREHENSIVE DISSERTATION INDEX

A felsorolt 13 adatbázisban szabad keresőszavakból álló, eltérő keresőprofilokkal haladtak végig, a keresőszavak és kombinációik megválasztása az adatbázis tárgykörétől, a változtatás az első keresés eredményétől függött. A kísérlet logikáját a következőképpen alakították ki:

1. *A keresőszavak I. halmaza* a négy első egymással logikai VAGY, ezek a két utolsóval ÉS kapcsolatban:

WATER	(víz)
RAINWATER	(esővíz)
PRECIPITATION	(csapadék)

RUNOFF/RUN OFF (lefolyás, két írásmódja
használatos)
COLLECT... (gyűjtés)
HARVEST... (kitermelés)

2. A keresőszavak 2. halmaza, itt is az első négy VAGY-ban, ezek az utolsó négygel ES-ben:

WATER
RAINWATER
PRECIPITATION
RUNOFF/RUN OFF
STORING... (tárolás)
STORAG...
HARVEST
COLLECT

3. halmaz, keresőszavak egymással VAGY kapcsolatban:

DRINKING (ivóvíz)
WATER
WELL WATER (kútvíz)
SOIL WATER (talajnedvesség)
UNDERGROUND (hévíz)
WATER
GROUND WATER (talajvíz)

4. halmaz (VAGY-ban):

IRRIGAT... (öntözés)
SPRINKL... (esőztető öntözés)

5. halmaz (VAGY-ban):

DESERT STRIP (parlagos vetésforgó)
FARMING
RUNOFF BASED (danaidás vízhozam-
mérő)
PITCHER
WATER-PONDING (belvizi elöntés)
EVAPORATION (párolgáscsökkentés)
SUPPRESSION

A kezdő keresőprofil szerkezete (stratégia).

(I. VAGY (2. ÉS 4.)) ÉS NEM 3. vagy 5.

Ezt a profilt futtatták végig az AGRICOLA és a CAB teljes állományán. A keresés eredményeként a 2. halmaz keresőszavait csökkentették, így a módosított, II. jelű halmaz (első szó ES-ben a második kettővel)

WATER
STORAG...
COLLECT...

Az új keresőprofil pedig, amelyet a NTIS, BIOSIS, COMPENDEX és WRA adatbázisra alkalmaztak:

I. VAGY (II. ÉS 4.) VAGY 5.

A WRA-ban olyan sok a módosított profilnak megfelelő választ találtak, hogy — utólagos megítélés szerint — jobb lett volna az eredeti keresőprofil alkalmazása.

További egyszerűsítések következtek, az 1. csoport a

WATER
RUNOFF
HARVEST

keresőszavakra redukálódott (első ES-ben a két utóbbival), jele I lett.

Az AQUALINE, LIFE SCIENCE COLLECTION, GEOARCHIVE és GEOREF adatbázisokon a keresés stratégiája tovább egyszerűsödött:

I vagy (II és 4).

Az 1. táblázat foglalja össze a kísérlet céljára különböző keresési stratégiákkal elért eredményeket. A táblázat első oszlopa az első, a részletes stratégiával elért találatok számát mutatja, a második oszlopban csak az 1. halmazzal végzett keresés találat száma áll, a harmadik oszlop a két egyszerűsített keresőprofil (azért kettő, mert az adatbázisok két csoportjához különbözőket kellett megállapítani), a negyedik oszlop pedig az 5. halmaz keresési eredményeit mutatja.

1. táblázat

Különböző keresési stratégiákkal
kapott találati számok

Adatbázis	A rész- letes stratégia	1. halmaz	1. VAGY 5. halmaz (2. és 4) 1. VAGY (II és 4)	5. halmaz
AGRICOLA	215	185	217	5
CAB	180	161	187	8
NTIS	17	17	3	
BIOSIS	40	37	40	
COMPENDEX	14	17	17	3
WRA	258	254	258	9

Érdekes összehasonlításra ad lehetőséget a táblázat első két oszlopa: a gondosan kidolgozott, részletes keresőprofil és az 1. halmaz alkalmazása profilként olyan találati számokat eredményez, melyek között a legnagyobb különbség az AGRICOLA esetében van. Ha az 1. halmaz keresőszavainak egyszerűsített keresését végezték volna csak el, az eredmény így is a kombinált keresés eredményének 86%-a lenne.

A használt adatbázisok összehasonlítása

Az online és a manuális irodalomkeresés összehasonlítását elvégezték valamennyi adatbázisra, az eredményt a 2. táblázat foglalja össze.

T = az online kereséssel kapott összes tételek száma

R = az online kereséssel kapott releváns tételek száma

U = a más adatbázisban nem tárolt (unikális) tételek száma

B = a manuális módszerrel is megtalált tételek száma

A = a szerző neve szerint manuálisan talált tételek száma

WH = azon tételek száma, melyeknek a címében szerepelt a

WATER HEARTVESTING (vízkitermelés) szakkifejezés.

2. táblázat

A kézi és gépi irodalomkeresés eredményei
(találatszámok)

Adatbázis	F	R	U	B	A	WH
WATER RESOURCES ABSTRACTS	258	117	56	43	30	44
AGRICOLA (1970-present)	215	100	28	36	22	83
CAB	180	122	72	29	35	57
AGRIS	71	35	8	10	7	29
BIOSIS	40	17	3	6	5	13
PASCAL	28	10	5	2	2	6
GEOREF	19	3	1	-	2	2
NTIS	17	11	5	-	4	6
COMPENDEX DISSERTATION INDEX	14	10	1	3	1	6
AQUALINE	8	6	3	1	-	2
GEOARCHIVE	7	4	2	-	1	2
LIFE SC. COLL	3	3	-	1	2	3
Összesen:	863	441	184			

A táblázat sorai a találatok csökkenő sorrendjében vannak rendezve: az első négy adatbázis szolgáltatja a találatok 83%-át. Érdekes eredmény azonban az, hogy míg a négy vezető adatbázis esetében a találatoknak — durván — a fele volt releváns, a szám szerinti utolsó adatbázisokban a találatok 80–100%-a.

Megjelenési idő, átfedés

A kísérlet kiterjedt még arra is, hogy a találatok hogyan oszlanak meg publikálásuk éve szerint; 1979 előtt vagy után jelent-e meg a talált forrás. A 2. táblázat első négy adatbázisának tételei közül meglepő módon 33, 3, 10 és 25% volt csak az 1979 után publikáltak aránya. Ez a nagy kétsédelem bizonyos fokig megbocsátható a WRA és a CAB esetében, ezekben kivonattal látják el a tételeket, ami természetesen késést jelent.

Az egyes adatbázisok természetesen tartalmaznak azonos tételeket, az átfedést (a WRA 117 releváns tételét véve alapul) a 3. táblázat mutatja.

3. táblázat

A WRA-val átfedésben előforduló, azonos releváns találatok száma

Adatbázis	WRA	AGRI- COLA	CAB	AGRIS	Egyebek
WRA	-	43	24	12	26
AGRICOLA	43	-	38	25	25
CAB	24	38	-	18	26
AGRIS	12	25	18	-	15
Egyebek	26	25	26	15	-

A listavezető adatbázisokban van néhány olyan tétel, amely egyik más adatbázisban sem szerepel (unikális találatok), a lista végén állók tételei azonban mind megtalálhatók az előbbieken. Elenyészően kevés a csak egy, de kis találatszámú adatbázisban megtalált szakirodalmi tételek száma.

Meglepőnek mondható még az az eredmény is, amit a 2. táblázat utolsó oszlopából lehet kikövetkeztetni: nevezetesen, ha csak a WATER ÉS HARVEST... keresőszóval végezték volna a keresést, a legtöbb releváns tételt megkapták volna az AGRICOLA és az AGRIS esetében, viszont a WRA és a CAB találatainál a releváns tételek felét sem éri el ez szám.

A keresés költségeit részletesen kiszámították, így a teljes, egy találatra eső, egy releváns találatra eső, és az unikális tételekre eső költségeket. A lista élén álló adatbázisok keresési költsége az alacsonyabb.

A kézi keresés eredményei

A kiindulásként használt cikk 105 hivatkozásából 42 olyan, amelyet a gépi keresés nem talált meg. 39 tétel címe tartalmazta a WATER HARVESTING szakkifejezést, és 9 (tehát 8,6%) cikk jelent meg 1979 után. A 105 hivatkozásból a WRA 43-at, az AGRICOLA 36-ot tartalmazott.

A kézi és gépi keresés együttesen 316 releváns tételt szolgáltatott.

A csak cím szerint kereső online módszer ennek 13%-át nem találta volna meg, a szerző szerint kereső pedig 10%-át.

Az online talált címek 43%-a tartalmazta a két kiemelt keresőszót, a kézi módszerrel találtaknak pedig 39%-a: ez a különbség elhanyagolható.

Végső következtetés: a kereső stratégiát a lehető legegyszerűbben érdemes kialakítani, nem kell minden lehetséges logikai kapcsolatot végigfuttatni a fontos keresőszavakkal. Költség szempontjából: nem kell minden szóbaejhető adatbázist lekérdezni, érdemesebb a téma "nagy" adatbázisaira szorítkozni a kereséssel.

/NABER, G.: Online versus manual literature retrieval: a test case shows interesting results in retrieval effectiveness and search strategy = Database, 8. köt. 1. sz. 1985. p. 20–24./

(Domokos Miklósné)