TMT 52. évf. 2005. 7–8. sz.

Csík Tibor – Varga Katalin

A tudás és az információfeldolgozás

Az információ korunk legismertebb árucikke, terjesztése, rendszerezése gazdaságilag egyre fontosabb tevékenység. A technológiai fejlesztéseknek köszönhetően (olcsó digitális információkezelés és kommunikáció) az információhoz jutás mind kevésbé jelent problémát. Egyre inkább növekszik viszont azon ismeretek jelentősége, amelyek az információ megszerzésének mikéntjére, illetve a források, szolgáltatások megválasztására, igénybevételére vonatkoznak. A tanulmány arra a kérdésre keresi a választ, hogy a mind szélesebb körben elérhető adatbázisok tartalmi feltáró rendszere milyen kihívásokra milyen válaszokat ad, vagyis milyen szinten képesek ezek a források megoldani a minőségi információfeldolgozás problémáját. A közlemény a szerzők 1997-ben végzett nemzetközi vizsgálatáról készített angol nyelvű beszámoló rövidített, átdolgozott változata. (Az eredeti angol szöveg 2000-ben jelent meg az Oxford kiadó gondozásában, lásd az irodalomjegyzékben.) Megállapításai, következtetései időtállónak bizonyultak, ezért úgy véljük, a hazai szakmai közélet érdeklődésére is számot tarthatnak.

A tudományos élet sem vonhatja ki magát azon változások alól, amelyek a közlemények számának exponenciális növekedéséből, a tudásreprezentálás és az információkeresés új eljárásaiból következnek. Egy-egy téma szakirodalmáról leginkább olyan online adatbázisok révén tájékozódhatunk, amelyek az adott ismeretkör dokumentumait dolgozzák fel. A könyvtáraktól ma már azt is elvárják a használók, hogy a gyűjteményük gyarapítása és rendelkezésre bocsátása mellett igény szerint hozzáférést biztosítsanak referensz adatbázisokhoz. A különböző szakterületek közleményeit feldolgozó referensz adatbázisok előállítása, forgalmazása, a bennük való kereséshez szükséges ismeretek átadása óriási üzleti lehetőségeket rejt. Az adatbázis-szolgáltató cégek – Ovid, CSA, Proquest, EBSCO, OCLC, hogy csak a legnagyobbakat említsük – sokszor ugyanazt az adatbázist kínálják más-más formában és szolgáltatásokkal. Mindezen információs tevékenység alapja a közleményekben felhalmozott ismeretek megfelelő reprezentációja, mondanivalójának tematikus feltárása.

A reprezentáció eszközei

A tudás tartalmi reprezentálására a legszélesebb körben használt eszközeink mind a 19. században születtek:

- *Tudományfelosztáson alapuló osztályozási rendszerek:*
  - A felosztás mögött valamilyen tudományrendszertan áll, amelynek alapja filozófiai vagy „általánosan elfogadott" gyakorlat lehet.
  - Az ismeretkörökön belül az osztályok, alosztályok stb. kialakítása logikai alapokon történik (alá-fölé rendeltség, egész-rész viszony). A felosztás szempontját adó lényeges megkülönböztető jegy (differentia specifica) azonban relatív, s egyetlen hierarchikus rendszer sem lehet mentes a következetlenségektől, ellentmondásoktól.
  - Az osztályozó ismérveket általában kódokkal is megjelölik, ami mutatja a hierarchiában elfoglalt helyet.
- *Természetes nyelven alapuló „tárgyszavas" vagy indexelő eljárások:*
  - Az ismérvek egyértelmű jelölése a kulcstényező – a nyelvi sokszínűségből eredő bizonytalanságokat ki kell küszöbölni (homonímia, szinonímia stb.).
  - A téma egyediségének, újdonságának felmutatására lehetőséget ad.
  - Az ismérvek összefüggéseinek (tematikai, logikai) jelzésére a szótárban kiépített utalórendszer szolgál.
- *Szöveges ismertetők, tartalmi összefoglalók* (annotáció, kivonat, referátum stb.):
  - Tartalmi és formai problémákkal egyaránt számolni kell.

Az osztályozás és az indexelés ma is egymás mellett élő eljárások, amelyek kölcsönösen függenek egymástól, és kiegészítik egymást. Az osztályozás feladata, hogy elhelyezze a témát az ismeretkörök között, illetve megadja a helyét a tudomány(ok) rendszerében (l. az Egyesült Államok gyakorlata, LCC, DDC). Az indexelés szolgálja inkább az egyedi ismérvek leírását.

A könyvtári katalógusok többnyire még mindig egyké osztályozási rendszert és valamilyen tárgyszavazási eljárást alkalmaznak (LCSH, DDC, ETO, LCC stb.). Az egyes tudományterületek szakirodalmát feldolgozó adatbázisok viszont a tartalmi feltáró eszközök széles körét használják. A feltáró eszközöket úgy válogatják össze, hogy minél inkább ki lehessen használni előnyös tulajdonságaikat, továbbá egymáshoz illesztve őket, rendszerré szervezik.

A tudásreprezentáció és az adatbázisok

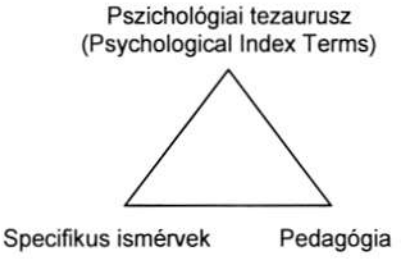
A referensz adatbázisok mindig pontosan meghatározzák a feldolgozás során követett célokat, módszereket, a forrásdokumentumok körét, illetve a reprezentált ismeretterület fókuszát és határterületeit. A szakirodalom megköveteli az egyediség reprezentálásának képességét, ezért az adatbázisok olyan tartalmi feltáró eszközöket alkalmaznak, amelyek megfelelnek az adott tudományterületnek és az érintett dokumentumoknak. Az adatbázis fókuszán kívül eső információs elemek feltárását is az elsődleges szempont határozza meg. Ahogy távolodunk a fókusztól, úgy csökken a reprezentáció hatékonysága és specifikussága. A különböző adatbázisok más és más feltáró eszközöket alkalmaznak, az ezek között létrehozható konkordancia az egyik legizgalmasabb szakmai probléma. Egy tudományterületen a tartalmi reprezentációt három tényező határozza meg *(1. ábra)*.

1. ábra **A tartalmi reprezentációt meghatározó tényezők**

A három tényezőnek összhangban kell lennie, ha azt akarjuk, hogy a tartalmi feltárás valóban hatékony legyen; így egy referensz adatbázisban – pl. a PsycInfóban – a természetes nyelven alapuló feltáró eszköznek illeszkednie kell másik két faktorhoz *(2. ábra)*.

2. ábra **A PsycInfo tartalmi feltárása**

A tezaurusz fogalomkészletének és a közöttük lévő kapcsolatoknak együttesen kell megfelelniük a fenti követelményeknek annak érdekében, hogy megvalósítható legyen a specifikus feltárás a pszichológia területén. A fogalmak és a relációk alkotta rendszer tehát a feltárás céljához és a releváns ismeretekhez igazodik. A PsycInfo tezaurusza tartalmaz ugyan jó néhány pedagógiai vonatkozású deszkriptort, de ezek és a közöttük feltüntetett relációk különböznek egy pedagógiai tezaurusz fogalmaitól, vagyis ezekkel a pedagógiai információt nem lehetne ugyanilyen szinten feldolgozni *(3. ábra)*.



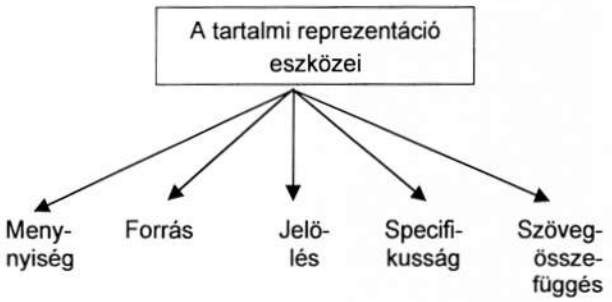
3. ábra **A PsycInfo tezaurusza és a pedagógia**

A kutatási probléma

A vizsgálat során egy 50 adatbázisból álló mintán tanulmányoztuk, hogy milyen eszközöket használnak a tartalmi feltárásra. A cél az volt, hogy számba vegyük, melyek a gyakorlat által is igazolt lehetőségek a tudás hatékony reprezentálására. Különös figyelmet fordítottunk arra, hogy a legnagyobb és legmérvadóbb nemzetközi szakirodalmi adatbázisok hogyan oldják meg ezt a problémát,

hogyan hasznosítják a rendelkezésükre álló eszközkészletet.

Minden egyes szempontot külön-külön figyelembe vettünk, amely a feltárt tétel tartalmára utalt – kivéve az elsősorban azonosításra szolgáló szempontokat (szerző, cím, megjelenési adatok stb.). Így minden, ami a tartalmi feltárást szolgálja: valamennyi mező, amelyben tematikus ismérvek találhatók (pl. személynevek, dokumentumtípusok, felhasználói célcsoportok), minden megoldás, amely további információval szolgál a tartalomról (pl. súlyozás, minősítés, sorrend) a vizsgálat tárgyát képezte. Azokban az esetekben, amikor egy szempontot többféle módon is kifejeznek (pl. kóddal és természetes nyelven is), azt külön eszköznek tekintettük, amennyiben az eltérő megjelölés többletinformációt is adott (pl. a kódolásban manifesztálódó rendszer). Megvizsgáltuk ezeknek a feltáró eszközöknek a típusait, jelölési módjukat, eredetüket (pl. szabad vagy ellenőrzött szótárból származó tárgyszó), jellegüket (generikus vagy specifikus), és azt a módot, ahogyan a szövegösszefüggést ábrázolják. A *4. ábra* a vizsgálat szempontrendszerét mutatja.



4. ábra **A tartalmi reprezentáció eszközeinek vizsgálati szempontjai**

A vizsgált minta

Az összeválogatott adatbázisok többsége a *Dialog* és a *DataStar* kínálatából való, amely cégek az 1990-es évek végén a legjelentősebb szolgáltatók voltak. Arra törekedtünk, hogy valamennyi témakör és tudományterület reprezentálva legyen. A minta nemzetközi, az Egyesült Államokon kívül Európa és Japán is képviseltetik magukat. A vizsgált adatbázisok különböző információs igényekre adnak választ. Öt nagyobb csoportra oszthatók, és minden csoportban a *feltárt objektumok*, illetve az *ismeretkör* határozzák meg a feltárási eszközöket. Az öt csoport a következő:

1. *Címjegyzékek, adattárak*
   - Objektumok: cégek, vállalkozások, személyek.
   - Ismeretkör: üzleti adatok.
   - Adatbázisok: ABCE – German Business and Industry Directory; CZCO – Official Register of Czech and Slovak Organizations; D&B – International Dun's Market Identifiers; DSCL – Disclosure Database; GDDB – Gale Directory of Databases; The McGraw-Hill Companies Publications Online; PLCO – Directory of Polish Companies; WWEB – Who's Who in European Business.
2. *Általános bibliográfiák*
   - Objektumok: meghatározott dokumentumtípusokban megjelenő általános információk.
   - Ismeretkör: általános.
   - Adatbázisok: CBIB – Current Contents Search; DISS – Dissertation Abstracts Online; WTI – World Translations Index.
3. *A „kemény" tudományok referáló adatbázisai*
   - Objektumok: specifikus információk egy megadott dokumentumkörben.
   - Ismeretkör: tudományorientált (kemény tudományok).
   - Adatbázisok: ABI/Inform; Agricola; BIOSIS Previews; CABI – CAB Abstracts; CA Search – Chemical Abstracts; Derwent Drug File; Econlit; EI-Compendex – Engineering Index; EMBASE (Excerpta Medica); Enviro/Energyline Abstracts; FSTA – Food Science and Technology Abstracts; INSPEC; INON – Insurance Information Online; JICST-EPLUS – Japanese Science and Technology; MMKA – Management and Marketing Abstracts; MEDLINE; NTIS – National Technical Information Service; Pascal.
4. *A „puha" tudományok (társadalomtudományok, humaniórák) referáló adatbázisai*
   - Objektumok: specifikus információk egy megadott dokumentumkörben.
   - Ismeretkör: tudományorientált (puha tudományok).
   - Adatbázisok: RILA – Art Literature International; Artbibliographies Modern; ASSI – Applied Social Science Abstracts and Indexes; CELEX – European Union Law; ERIC; Historical Abstracts; ISA – Information Science Abstracts, LLBA – Linguistics and Language Behavior Abstracts; LISA – Library and Information Science Abstracts; PAIS International; Philosopher's Index; PsycInfo – Psychological Abstracts; Religion Index; Sociological Abstracts.

5. *Teljes szövegű adatbázisok – folyóiratok*
   - Objektumok: Egy adott dokumentumban található legspecifikusabb információk.
   - Ismeretkör: válogatott publikációk (a válogatás kvantitatív vagy kvalitatív szempontok szerint történik) – általános.
   - Adatbázisok: AGEN – Agence France-Presse Newswires; AP News; FAZA – Frankfurter Allgemeine Zeitung; FTEE – Financial Times Reports: Eastern Europe; HBRO – Harward Business Review Online; Le Monde; Los Angeles Times.

Kutatási módszer

A tartalmi reprezentáció eszközeit különböző perspektívákból tanulmányozhatjuk (l. 4. ábra), ezek a szempontok adják a vizsgálat fő vonalát. Megszámoltuk, hányféle ismérv található az adatbázisban az egyes szempontoknak megfelelően. A statisztikai elemzés eredménye világosan mutatja az egyes adatbázisok jellemzőit. A következtetések levonásakor elsősorban abból indultunk ki, hogy az egyes adatbázisok hányféle és milyen tartalmi feltáró eszközt alkalmaznak. A tartalmi reprezentációnak sem a minőségét, sem a következetességét nem vizsgáltuk.

Eredmények, megállapítások

A vizsgálat során a tartalmat reprezentáló ismérvek alábbi jellemzőit találtuk:

- Mennyiségi szempontból elkülönítendők az önálló álló, saját adatmezőben megjelenő, illetve a más ismérvekhez kapcsolódó ismérvek (független – függő).
- Az ismérvek forrása lehet maga a dokumentum (pl. lead paragraph, automatikus referátum, CR = Content Representation), vagy származhatnak külső forrásból. Ez utóbbiak lehetnek szabályozatlan ismérvek (pl. free term, key phrase, identifiers), állhat mögöttük ellenőrzött szótár (tezaurusz, dokumentumtípus-lista stb.), vagy lehetnek szabad szöveges leírások (referátum).
- Jelölés szempontjából találtunk természetes nyelvű és kódokkal jelölt ismérveket, illetve nem ritka a kétféle jelölés együttes alkalmazása sem.
- Az ismérvek tartalmát tekintve el kell különíteni a tartalmi elemet hordozó formai szempontokat (pl. dokumentumtípus, tárgyalásmód, intellektuális szint, felhasználói célcsoport), a preferált tartalmi ismérveket (pl. földrajzi név, személynév, ember/állat). Tartalmi szempontból az ismérvek jellege lehet generalizáló vagy individualizáló, illetve hierarchikusan osztályozó vagy leíró.
- Az ismérvek eredeti szövegbeli viszonyának tükrözésére az adatbázisok alkalmazhatnak természetes nyelvű kontextust (pl. referátum, CR, key phrase, lead paragraph), súlyozást (major descriptors, minor descriptors), generikus vagy specifikus determinatívumokat (pl. adalékanyag – titán-dioxid), szerepoperátorokat, szintaxist jelölő linkeket.

Mennyiség – a tartalmi feltáró eszközök száma

Minden olyan adatelem, amely a tartalomra utaló információt hordoz, különálló szempontnak tekintendő. Az ismérvek lehetnek függetlenek, amelyek önállóan szerepelnek (pl. deszkriptorok, osztályozási jelzetek, tárgyszavak, kiemelt tartalmi jellemzők, dokumentumtípusok, kezelési kódok), vagy függhetnek egy másik ismérvtől – módosítják annak jelentését, illetve további információkat adnak a kontextusra vonatkozóan (pl. súlyozás, minősítők, altárgyszavak). A címjegyzékekben és adattárakban a tartalmi feltáró elemek kissé eltérőek, gyakran numerikusak (pl. termékek, alkalmazottak száma, kereskedelmi adatok).

A tartalmi feltáró eszközök között számba vettük a referátumokat, de a teljes szöveget nem. Ha az adatbázis bizonyos adatcsoportokat együtt és külön-külön is kereshetővé tesz (pl. DE[drug], DE[medical], DE[all]), csak a különálló mezőket számoltuk, a közöset nem (DE[drug], DE[medical]). Néhány esetben ugyanazt az ismérvet természetes nyelven és kóddal vagy rövidítéssel is jelölik. Ezeket akkor tekintettük különálló szempontoknak, ha a kód rendszerbeli hovatartozást is jelöl (pl. tárgyszavak és tárgyszókódok két külön ismérvnek tekinthetők, a dokumentumtípusok elnevezése és kódja viszont nem).

A vizsgált adatbázisokra általánosan jellemző a tartalmi feltárás szegmentáltsága, ennek mértéke azonban különböző. Minél nagyobb egy adatbázis, és minél aktuálisabb információs igényekre ad választ, annál többféle ismérvet használ a tartalom reprezentálására. Tíz vagy több ismérvtípus sem ritka. Sokat elárul az ismérvek átlagszáma az egyes csoportokban: 1. csoport: 9,25; 2. csoport: 5; 3. csoport: 9,27; 4. csoport: 5,92; 5. csoport: 6,42. A tartalmi feltárás szegmentáltsága tehát szoros összefüggésben áll a feltárás mélységével, illetve a reprezentálandó információk egzaktságával.

A címjegyzékek és az adattárak relatíve nagy számú ismérvet használnak, azaz igen részletes a

tartalmi feltárásuk. Ezek az ismérvek szinte kivétel nélkül függetlenek, ez következik az adatbázistípus jellemzőiből. Csak a bibliográfiai adatbázisoknak kell megküzdeniük azzal a problémával, hogy hogyan fejezzék ki a tartalomnak azt az aspektusát, amely a szövegösszefüggésben van elrejtve. A keményebb tudományok használják a legtöbb függő ismérvet (főleg minősítőket és szerepjelölőket). Itt a legnagyobb az ismérvek átlagszáma, mivel itt találkozhatunk a legkifinomultabb információs igényekkel is. Az általános adatbázisok használják a legkevesebb ismérvet, a tartalmi feltárás itt a legátfogóbb.

Az ismérvek forrása

A tartalmat reprezentáló ismérvek a dokumentum szövegéből (pl. első bekezdés, kulcsszó) vagy külső forrásból származhatnak. A kívülről vett ismérvek szabályozottak vagy szabályozatlanok lehetnek. A szabályozott ismérvek forrásaként az adatbázisok különféle szótárakat, tárgyszójegyzékeket, tezauruszokat, egyéb ellenőrzött listákat használnak. A referátum nem tekinthető a dokumentum részének, így az is kívülről vett forrásnak számít. Maga a dokumentum, mint az ismérvek forrása, csupán a bibliográfiai adatbázisoknál érdekes. Ezért például az 1. csoportban csak az ellenőrzött listák oszlopába került adat. Ha egy adatbázis kódot és természetes nyelvű jelölést egyaránt alkalmaz ugyanarra az ismérvre, de mindkettőt ugyanabból a szótárból veszi, akkor ezt csak egy szabályozott jegyzéknek számoltuk. Általánosan elmondható, hogy relatíve kevés ismérv származik magából a dokumentumból, ennek magyarázata lehet, hogy az azonosításra szolgáló adattípusok között is vannak olyanok, amelyek tartalmi információt adnak (pl. a cím), és ezek természetesen minden adatbázisban kereshetők.

A források számbavétele bizonyítja, mennyire fontos a szabályozottság a tartalmi feltárásban. A tartalmat fedő pontos cím, a dokumentumok teljes szövegének kereshetővé tétele sem teszi feleslegessé a keresőelemek szabályozását, rendszerezését. Minél specifikusabb egy adatbázis, annál több ellenőrzött listát alkalmaz. E listák száma a kemény tudományok adatbázisaiban a legnagyobb, ahol a legspecifikusabb és legpontosabb információs igényeket kell kiszolgálni. A puha tudományoknál kevesebb a listák száma. Megfigyelhető, hogy még a teljes szövegű adatbázisok is szép számmal alkalmaznak szabályozott jegyzékeket.

Jelölés

A vizsgálat következő szempontja az volt, hogyan jelölik az egyes adatbázisok a tartalmi ismérveket. Leginkább kétféle jelöléssel találkozunk: természetes nyelv és kód. A numerikus adatok ebből a szempontból a természetes nyelvhez sorolódnak. Elég gyakori, hogy ugyanazt az ismérvet többféle jelöléssel is megadják az adatbázisok. A szisztematikus kódok az ismérvek összefüggéseit, rendszerét kívánják leképezni, többnyire hierarchikus szerkezetben. A kód megmutatja az ismérv rendszerben elfoglalt helyét, s közvetlen, szisztematikus keresést tesz lehetővé (pl. a csonkolt keresés módot ad a szintek közötti lépegetésre). Ugyanakkor a természetes nyelvű megfogalmazás közvetlenül informál a témáról. Ezekben az esetekben, bár a két ismérv ugyanazt a fogalmat takarja, a jelölés által közvetített információ más (direkt/indirekt megközelítés).

A természetes nyelv szerepe a tartalmi feltárásban észrevehetően nagyon erős. A referensz adatbázisok – különösen a kemény tudományoknál – kedvelik a kétféle jelölés együttes alkalmazását. A puha tudományoknál jóval kevesebb kódot találunk.

A tartalmi feltáró ismérvek specifikussága

Az ismérvek által lefedett fogalmi rendszernek megfelelően beszélhetünk generikus (átfogó fogalmi kategóriákat lefedő) és specifikus (a téma egyediségét megadó) eszközökről. A kettő között nehéz meghúzni a határvonalat, csak egy adatbázison belül lehet eldönteni egy-egy ismérvről, hogy generikus-e vagy specifikus. Mindkettő lehet osztályozó vagy leíró jellegű.

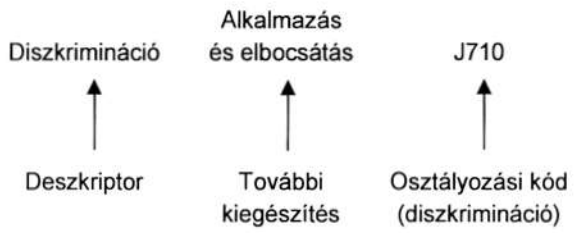
Az osztályozás és a tárgyszavas leírás mellett az adatbázisok széles körben kihasználják a számítástechnika adta lehetőséget, hogy bizonyos tartalmi ismérveket preferált adatmezőkben helyezzenek el. Ezek jól elkülöníthető, pontosan definiált szempontokat írnak le. A kiemelés alapja lehet a feltárásban érvényesülő fazettás elv, de többnyire gyakorlati oka van, például a keresés segítése. Ilyen kiemelt tartalmi elemek lehetnek például a személynevek, földrajzi nevek, anyagnevek, kémiai elnevezések, speciális jellemzők.

A tartalmi leírás szempontjainak elnevezése a különböző adatbázisokban azonos lehet, miközben tartalmuk teljesen eltér (pl. speciális jellemzők = special features). Ugyanaz a kategória az egyik adatbázisban lehet osztályozási rendszer, a má-

sikban kiemelt szempont, s ez mindig az adott adatbázis jellemzőitől függ. A cégadatbázisok például az amerikai SIC kódrendszert alkalmazzák osztályozási célokra, ugyanakkor azonban a gazdasági folyóiratokban a SIC kód kiemelt szempontként szerepel, csupán akkor alkalmazzák, ha a téma szempontjából lényeges.

Az adatbázisok előszeretettel különítik el a témára vonatkozó formai ismérveket: dokumentumtípusokat, cikkek típusait, médiatípusokat, felhasználói célcsoportokat, intellektuális szintet. Ezek árnyalják a témát, tehát külön szempontnak tekintettük őket.

Valamennyi adatbázis nagy súlyt helyez arra, hogy a témát generikus és specifikus ismérvekkel egyaránt leírja. A generikus kategóriák származhatnak osztályozásból, vagy lehetnek tárgyszó jellegűek. Specifikus osztályozást csak ritkán alkalmaznak, ilyeneket leginkább a kemény tudományok adatbázisainak némelyikében találni. A tartalmi reprezentáció éppoly individualizáló, mint az ellenőrzött szótárral, ugyanakkor láthatóvá válik a rendszer is. Az Econlit adatbázisban például a deszkriptorokat osztályozási kódokkal is megadják, hogy a kódok megmutassák a téma rendszerbeli elhelyezkedését. A deszkriptoroknál mód van a téma további szűkítésére, specifikus fogalmakkal való kiegészítésre *(5. ábra)*.



5. ábra

Az adattárak jobban kedvelik a kiemelt ismérveket és az átfogóbb tartalmi kategóriákat. A kemény tudományok adatbázisaiban a specifikus tárgyszavazás és a kiemelt ismérvek a legjellemzőbbek. A puha tudományoknál a specifikus leírás (deszkriptorok, szabad tárgyszavak) bizonyulnak a leghatékonyabb tartalmi feltáró eszközöknek. A teljes szövegű adatbázisok szinte egyáltalán nem alkalmaznak osztályozási rendszereket.

Szövegösszefüggés – a tartalmi feltáró eszközök közötti kapcsolatok

Az aktuális tartalom megragadása minden tartalmi feltárásnak az alapvető célja. Ehhez arra is szükség van, hogy ne csak értelmezzük a különálló ismérveket, hanem a forrásbeli viszonyukat is leképezzük. A kontextusban megjelenő fogalmak további információkat hordoznak, mivel a kontextus is információ. Ezért a vizsgálatunk utolsó szempontja az, hogyan tükrözi az adatbázisok feltáró rendszere az eredeti szövegbeli kontextust.

A legáltalánosabb a szöveges ismertetők, összefoglalók alkalmazása (pl. referátumok, kivonatok). A téma ismérvei és azok forrásbeli viszonyai természetes nyelvi kontextusban jelennek meg. Ugyanezen az elven alapul az az eljárás, amelyben kulcsszóként mondatszerű kifejezéseket adnak meg (key phrase). Egy másik lehetőség, hogy az ismérveket aszerint csoportosítjuk, mennyire meghatározóak, mennyire hangsúlyosak az adott forrásban. Az ismérvek súlyozása nagyon kedvelt a bibliográfiai adatbázisokban (pl. major/minor descriptors).

Az adatbázisok egy részében determinatív kiegészítőket csatolnak a tárgyszavakhoz. Ezek lehetnek minősítők (qualifiers), szerepjelölők, illetve altárgyszavak, amelyek módosítják vagy konkretizálják a fogalom jelentését az adott kontextusnak megfelelően. Vannak általánosan használatos kiegészítők, amikor minden tárgyszó betöltheti bármelyik szerepet, azaz kiegészíthet, módosíthat egy másik tárgyszót. A minősítők egy részét csak meghatározott szakterületen lehet használni (gyógyszerek minősítése, betegségek stb.), ezek erőteljesebb specifikációt jelentenek. Az adatbázisok általában két ismérvet kapcsolnak össze, több szempont viszonyainak leírására ritkán vállalkoznak. A kiegészítőket leginkább a keményebb tudományoknál alkalmazzák. Egy másik lehetőség, hogy az összetartozó fogalmakat láncszerűen összekapcsoljuk (link), ezáltal kerülünk közelebb az aktuális kontextushoz. Érdekes, hogy a teljes szövegű adatbázisok – annak ellenére, hogy náluk elérhető az eredeti szöveg – nagyon kedvelik ezt a megoldást.

Keresőszoftverek

A számítógépes adatbázisok további lehetőségeket kínálnak a tartalmi keresések finomítására, ezek a lehetőségek a keresőszoftverek szolgáltatásaiban rejlenek. Az adatelemek szegmentálása csak akkor hatékony, ha ezeket a szétdarabolt elemeket a keresés során tetszőlegesen lehet kombinálni. Az adatbázisok nyomtatott változatával ellentétben több keresési szempont is érvényesíthető egyszerre. A tartalmi keresést támogató eszközök a következők:

- A számítógép nemcsak ismérveket tud szegmentálni, hanem szavakat és kifejezéseket is. Egy ismérvet jelölő kifejezésen belül is kereshetünk egy szóra, részletre.
- Az indexelési módszerek szintén a tematikus keresést szolgálják. A tartalomra utaló fogalmak kereshetők szavanként, de kifejezésként is. A legtöbb adatbázis arra is módot ad, hogy a keresés során meghatározzuk, a keresett szó önmagában tárgyszó-e vagy egy összetett tárgyszó része (pl. DE[full descriptor], IF[full identifier], FF[full term anywhere]).
- Az adatelemek szegmentálása azt is jelenti, hogy a keresést limitálhatjuk azokra a tételekre, ahol a keresőszó egy bizonyos mezőben, vagy adatmezők egy csoportjában szerepel. A leírás szempontjait szokás a keresés segítése érdekében csoportosítani: alapindex (basic index) és kiegészítő indexek (additional indexes).
- A hagyományos Boole-algebra még mindig a legelterjedtebb eszköz a keresőelemek kombinációjára. Az erőteljesen szegmentált struktúrában azonban a szimpla Boole-operátoros keresés nagyon zajos, nem ad eléggé pontos találati halmazt.
- A legtöbb adatbázis-szolgáltató a pszeudo-Boole operátorok – azaz a közelségi operátorok – széles körét is kínálja. Ezeknek sokféle változata ismert: egy rekordon belüli elemek, egy adatmezőn belüli elemek, egy mondatban szereplő elemek, egy kifejezésben szereplő elemek, egymás után álló elemek. Az ilyen és ehhez hasonló eszközök alkalmazása bizonyíték arra, hogy az információkereséshez struktúrákra van szükség, nem csak szegmentált adatelemekre. Segítségükkel az eredeti természetes nyelvű szövegek is könnyebben kereshetők.
- A csonkolás a teljes szövegű keresést könnyíti, ezáltal sok nyelvi problémára jelent megoldást.

Kihívások és válaszok

A tudás változik, az információfeltáró eszközöknek pedig követniük kell ezeket a változásokat, válaszolni a kihívásokra. Figyelembe kell venni azonban, hogy az adatbázisoknak – mint organikus rendszereknek – tehetetlenségük van. Az új technikai megoldások, a gyors gépek megváltoztatták az osztályozás és az indexelés hagyományos eljárásait, bizonyos elemek erősödtek, mások gyengültek.

Szegmentáltság

Az informatikai fejlesztések egyik legkézzelfoghatóbb hatása a tudásreprezentáció szempontjából az adatelemek maximális szegmentálhatósága. A nagy teljesítményű számítógépeknek köszönhetően ma már könnyű elkülöníteni az információ egyes szegmenseit.

- Már a tudományok területén is nagyfokú specializálódás figyelhető meg. A tudományos ismereteket feltáró referensz adatbázisok is követik ezt a tendenciát, igazodva a meghatározó kutatási programokhoz, a tudás egy-egy szegmensét reprezentálják. Ahhoz is szakértelem kell, hogy megtaláljuk a megfelelő adatbázist.
- Az adatbázisokon belül az információ valamennyi aspektusa és az adatok minden típusa elkülöníthető. Az adatbázis hatékony használata megköveteli szerkezetének nagyon pontos ismeretét. A tartalmi feltárás szempontjából az adatbázisok szerkezete erősen prekoordinált. További problémát jelent, hogy az adatbázis-szolgáltatók különböző formátumokban és szerkezeti elrendezésben készítik el adatbázisaikat.
- A keresőszoftvereknek köszönhetően nemcsak az adattípusok, de a szavak és szószerkezetek is szétválaszthatók. A legtöbb adatbázis lehetővé teszi a szószerkezetek elemeinek, sőt akár szavak részleteinek is a keresését.

A kérdés, hol van a szegmentálás határa. Meddig mehetünk el? Mi az a legkisebb információhordozó elem, amely elkülöníthető, címkézhető, visszakereshető? Egy dolog biztos: a szegmentálás nem mehet a végtelenségig.

Az adatbázisok dokumentumok leírásait tartalmazzák, úgy tűnik, hogy a tudás alapegysége a dokumentum. A tartalmi feltáró ismérvek szintén a dokumentumokra vonatkoznak. A valóságban azonban egy dokumentum több tudásegységet is tartalmazhat. A probléma tehát ezek meghatározása és reprezentálása.

Szabályozottság

Habár a trendek a teljes szövegű adatbázisok elterjedése felé mutatnak, ahol maga a szöveg adja a tartalmi keresés elemeit, a vizsgálat azt igazolja, hogy a szabályozott listáknak és szótáraknak még mindig nagyon fontos szerepük van. A használók gyors és teljes információt akarnak, de ugyanakkor pontosat is. Terminológiai ellenőrzés nélkül ez megoldhatatlan.

A tezauruszok mellett az adatbázisok sok más mezőben is alkalmaznak szabályozott listákat (dokumentumtípusok, földrajzi nevek, osztályozási jelzetek stb.). A listák száma nő, de a szabályozás már nem olyan mély. Ezeknek a jegyzékeknek a

többsége csak az egységességet szolgálja, rendszert már nem ad.

Osztályozás

Az osztályozási rendszerek hierarchiája egyfajta rendet biztosít, ugyanakkor azonban a hierarchia nem mindig jelent alárendelést, jelenthet összefoglalást is. Az adatbázisok egyaránt használják a leíró indexelést és a szisztematikus osztályozást. Ez azt mutatja, hogy a generikus osztályozás még mindig hatékony, ha jól kombináljuk leíró indexeléssel, ezáltal a téma reprezentációja árnyaltabb, és a két szempont a keresésben is jól kombinálható. Az is világos, hogy a hagyományos osztályozási rendszerek (pl. ETO, DDC, LC) nem találhatók meg a nagy adatbázisokban, ezek inkább saját rendszereket használnak.

Az erős hagyományokkal rendelkező adatbázisoknak meg kell őrizniük hagyományaikat. Többségüknek van nyomtatott változata, amelynek a rendszere osztályozási rendszerként megjelenik az online változatban is.

Kontextus

A használók nem szavakat vagy szószerkezeteket keresnek, hanem teljes témákat, ahol a szavak és szószerkezetek valamilyen kapcsolatban állnak egymással. A vezető adatbázisok pontosan tisztában vannak ezzel a követelménnyel. A természetes nyelven megfogalmazott referátumok nem elegendőek, arra is szükség van, hogy a szabályozott szótárakból vett fogalmak között is kapcsolatot teremtsünk. Ennek legkedveltebb eszközei: a legfontosabb fogalmak súlyozással történő kiemelése, illetve kiegészítő, minősítő fogalmak alkalmazása a jelentés konkretizálása érdekében. Két fogalom kombinációja (tárgyszó-altárgyszó) is elég gyakori. A szerepoperátoroknak hasonló szerepük lenne, ezeket azonban ritkábban alkalmazzák.

Az elkülönítés és az összeillesztés az a két alapvető folyamat, amely az információkeresést jelenleg leginkább jellemzi. A tartalmi feltáró eszközök gondoskodnak a szemantikai tér szabályozásáról. A hagyományos osztályozási rendszerek és a prekoordinált tárgyszójegyzékek elérték a saját határaikat, az igények ma arra irányulnak, hogy megőrizzük a tartalmi elemek eredeti szövegbeli összefüggéseit is. A tudományorientált adatbázisokban találunk néhány helyi megoldást az izolált elemek közötti szintaktikai kapcsolatok feltüntetésére (tárgyszóláncok, altárgyszavak, szerepoperátorok, minősítők). Ezeknek a lehetőségeknek a kiterjesztése, és új módszerek kifejlesztése a szintaktikai tér szabályozására lehet a következő lépés a magasabb szintű tudásreprezentáció irányába.

A kérdés az, létezik-e általános megoldás a kontextus reprezentálására. Intenzív kutatások folynak erre vonatkozóan. Általános szintaxis azonban nem létezik, csak elméletben. *Foskett* szerint a szintaxis tükrözésének két alapvető módja képzelhető el: (1) jelezni a tényt, hogy kapcsolat áll fenn, anélkül, hogy minősítenénk azt; (2) konkrétan meghatározni a kapcsolatot. Az első akkor történik, amikor Boole-operátorokkal vagy közelségi operátorokkal keresünk, vagyis meghatározzuk, hogy mely ismérveknek kell együttesen megjelenniük egy rekordban. A kapcsolat konkretizálására több elmélet is született (pl. láncindexelés, PRECIS).

Preferált ismérvek

A szegmentálás eredményeképpen a preferált ismérvek számtalan formája megtalálható. Főként az adattárak kedvelik ezt az eszközt, megjelölve minden lényeges adatelemet, és elkülönítve őket különböző adatmezőkben. A bibliográfiai adatbázisokban azok lesznek preferált ismérvek, amelyeknek az adott tudományterület szempontjából különleges jelentőségük van. Ezek szorosan kötődnek a tudományokhoz, például földrajzi nevek, kémiai nevek, személynevek. Visszakeresésük akkor is fontos, ha nem tartoznak a főtémához. A preferált ismérvek egyre erőteljesebb alkalmazása egyértelműen mutatja a számítástechnika hatását.

Természetes nyelv

A természetes nyelvű eszközök soha nem látott reneszánszát éljük, hiszen ezek nagyon felhasználóbarát módszerek. Ugyanakkor persze a keresés hatékonysága behatárolt a jelentésvariációk miatt, és az ellenőrzés mindig időleges lehet. További problémákat okoz, hogy a tudományok egyedi szakkifejezéseket használnak, és az adatbázis-előállítók szóhasználata is eltérő.

Az adatbázisok alapvetően szöveges információkat tartalmaznak, ritkák az ábrák és a grafikonok. A szakértői rendszerek és a mesterséges intelligencia témakörében folyó kutatások is azt mutatják, hogy a természetes nyelvű tudásreprezentáció és információkeresés a közeli jövőben nem fog veszíteni vezető helyéből. Ezzel egy időben erősödik a kódrendszerek alkalmazásának tendenciája is. Az általunk vizsgált adatbázisok bizonyítékul szolgálnak a kétféle megközelítés harmonikus egymás mellett élésére.

Konklúzió

Megvizsgálva a jelenleg elérhető tudásreprezentáló eszközöket és módszereket, egyértelmű, hogy két alapelv érvényes: *szedd szét, és illeszd össze.* A dokumentumok témáját előre meghatározott szempontok szerint elemeire kell bontani, a fellelhető tartalmi elemeket fel kell darabolni, majd közöttük megfelelő kapcsolatokat létrehozni. Kezdetben a gyűjtemény állt a háttérben, később a tudományterületek átvették a vezető szerepet. Azóta a tartalmi feltáró eszközök csak jól körülhatárolt ismeretkörben tudnak hatékonyan funkcionálni.

Minél kidolgozottabb egy adatbázis, annál több megközelítési szempontot alkalmaz. A cél, hogy megmutassuk egy adott tudományterület egyediségét, különbözőségét. Az információ mennyiségi növekedésének ez az egyértelmű hatása. Következésképpen a felhasználóknak megalapozott tudással kell rendelkezniük az adatbázisokról, ha hatékonyan akarnak információt keresni.

A ma ismert eszközök csak tudástöredékeket tudnak kezelni. A dokumentumokban meglévő össztudás és a dokumentum tényleges üzenete csupán közvetett módon van reprezentálva. A mennyiségi növekedést minőségi váltásnak kell követnie.

Irodalom

COUSINS, Shirley Anne: Enhancing subject access to OPACs: Controlled vocabulary vs natural language. = Journal of Documentation, 3. sz. 1992. p. 291–309.

CSÍK Tibor: Ismeretek és könyvtári osztályozás. = Könyv, Könyvtár, Könyvtáros, 4. sz. 1995. p. 13–24.

CSÍK Tibor–VARGA Katalin: Knowledge and information processing. = Library automation in transitional societies. Lessons from Eastern Europe. Ed. by Andrew Lass and Richard E. Quandt. New York: Oxford Univ. Press, 2000. p. 293–312.

FARRADANE, J. E. L.: A scientific theory of classification and indexing. = Journal of Documentation, 6. sz. 1950. p. 83–99., 8. sz. 1952. p. 73–92.

FOSKETT, A. C.: The subject approach to information. London: Clive Bingley, 1982.

INGWERSEN, Peter: Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. = Journal of Documentation, 1. sz. 1996. p. 3–50.

LIN, J.: Integration of weighted knowledge bases. = Artificial Intelligence, 2. sz. 1996. p. 363–378.

LANCASTER, F. W. et al.: Evaluation of interactive knowledge based systems: Overview and design for empirical testing. = JASIS, 1. sz. 1996. p. 57–69.

McMURDO, G.: How the Internet was indexed. = Journal of Information Science, 6. sz. 1995. p. 479–489.

ROBERTSON, S. E.–BEAULIEU, M.: Research and evaluation in information retrieval. = Journal of Documentation, 1. sz. 1997. p. 51–57.

VICKERY, Brian: Conceptual relations in information systems. = Journal of Documentation, 2. sz. 1996. p. 198–200.

VICKERY, Brian–VICKERY, Alina: Information science in theory and practice. London: Bowker-Saur, 1987.

VICKERY, Brian: Knowledge representation: A brief review. = Journal of Documentation, 3. sz. 1986 p. 145–159.

VICKERY, Brian: Knowledge discovery from databases: An introductory review. = Journal of Documentation, 2. sz. 1997 p. 107–122.

WEINBERG, Bella Hass: Complexity in indexing systems – abandonment and failure: Implications for organizing the Internet. = ASIS 1996 Annual Conference Proceedings (19 October 1996).

Beérkezett: 2005. V. 12-én.

Csík Tibor
az Országos Pedagógiai Könyvtár és Múzeum tudományos titkára,
az egri Eszterházy Károly Főiskola oktatója.
E-mail: *Csik.Tibor@opkm.hu*

Varga Katalin
az Országos Pedagógiai Könyvtár és Múzeum könyvtárának vezetője, főosztályvezető.
A Pécsi Tudományegyetem könyvtártudományi tanszékének egyetemi adjunktusa.
E-mail: *kvarga@hu.inter.net*