Darányi Sándor
School of Journalism, Library and Information Science, Oslo College, Norvégia

Információcsillagászat az Interneten: elmélet és gyakorlat

Az információkeresést az Interneten három tény korlátozza: az elégtelen indexelés, a keresőmodellek tisztázatlanságai, valamint a navigáció nevű félreértés. A valódi három-, illetve négydimenziós navigáláshoz előbb szemantikai univerzumokat kell készíteni, amelyek a szakterületek ismereteit térben ábrázolják. Az ilyen tartalmi térképezés irányelvei Gerard Salton dinamikus könyvtára elképzelésében keresendők. Ez a rekurzív modell bármilyen, folyamatosan változó osztályozási rendszert fejlődésében képes leírni, és elektronikus dokumentumok kezelésére is alkalmas. Az eredeti modellben a klaszteranalízist főkomponens-analízissel helyettesítve, az információ láttatása is lehetővé válik. A helyettesítés eredményeként dokumentumok és kulcsszavak olyan stabil eloszlásait kapjuk, amelyek csillagképekre emlékeztetnek, egy ma még nem létező információcsillagászat felé egyengetve az utat.

Bevezetés

Úgy tartják, hogy az információs társadalom az információrobbanás következménye (vagy az lesz). Ezt a robbanást persze senki sem szó szerint érti, hanem egy olyan tágulási folyamatra utal vele, melyet az információ katalizál. E katalízis lényege az, hogy a tudás mennyisége az adatokból kivont információéval arányosan nő, a növekvő tömegű ismeret pedig egyre nagyobb teret foglal el. A körfolyamat gyorsul, a tágulás ezért hasonlít explózióra.

Ugyanakkor a metafora egy másik értelemben is megállja a helyét. Az Internet növekedési statisztikái azt bizonyítják, hogy újabb, sokkal kevésbé képletes információrobbanás játszódik le a szemünk láttára[1], amely dokumentumok új típusait hozta létre [1].

Egyszeri esemény lehet véletlen vagy csoda, ugyanabból kettő azonban egyik sem. A második robbanás tehát bizonyos gyakorlati és elméleti kérdéseket egyaránt felvet. Az mindenki számára világos, hogy egy ottlap, ftp-archívum vagy adatbázis esetében a tartalomnak adunk lokalizálható formát. Az elektronikus címhez kötött elektronikus tartalom azonban mára az érdeklődés középpontjába állítja a hálózati információforrások indexelését és visszakeresését, hiszen minél nagyobb tömegű adatból kell az algoritmusnak keresnie, a találatok pontossága annál inkább veszélyben forog[2].

Mivel a kezdet kezdetén semmiféle egyezmény nem kötötte ki a dokumentumok relevanciájának jelölésmódját, a keresés többnyire a html szabvány headertől headerig tartó mezejében, a teljes szövegben vagy az IP címtartományban történik. Ugyanakkor e modern dokumentumok nincsenek kulcsszavakkal indexelve, hiányoznak a keresés finomabb, elvontabb támpontjai, ami a találati halmaz minőségére visszahat. A megoldás tehát csakis valamiféle indexelés lehet, tartalmi sűrítményekkel, felettes fogalmakkal megcímkézett elektronikus dokumentumok és dokumentumfájlok létrehozása, automatikusan gyarapodó szövegegyüttes esetében nyilván automatikus indexeléssel [2].

Mi a probléma?

Amennyire ez a világhálózat fejlődésének ma még kusza és feldolgozatlan történetéből kiszűrhető, az elmúlt esztendőkben népszerűvé vált szolgáltatások – kis módosításokkal – rendre ugyanazokat az ötleteket használták, ezek a kis változtatások azonban evolúciójukhoz vezettek. Ha egy távoli gépnek szabványos IP-címet adunk, az eredmény a telnet lesz; ha ehhez a fel- és letöltés lehetőségét tesszük hozzá, megkapjuk az ftp-t; ezt

[1] Lásd az Internet Society statisztikáját (1995. aug. 2.): 1995 első félévében a növekedés 37%-os volt, a hostok száma elérte a 6,6 milliót. 14 negyedév növekedési rátáját alapul véve, az ezredfordulóra ez 101 millió gép bekapcsolását jelentené.

[2] A találati halmazzal ugyanis arányosan növekszik a „zaj" halmaza is. Manapság ez a naponta elemzett szövegvagyon 22–23 millió oldalra, 8–10 milliárd szóra becsülhető.

menükkel és kereszthivatkozásokkal kiegészítve, a Gopherhez jutunk; a kereszthivatkozásokat hipertextként kezelve létrejön, majd – grafikus felületen – szalonképes külsőt ölt a WWW.

Valamennyi felsorolt szolgáltatás visszakeresési oldala olyan szoftvert használ, amely tartalmi osztályok törzsfáját járja be, helyben vagy idegen gépeken. Ezeket a törzsfákat azonban el kell készíteni: a kúszómászó – a *crawler*, *spider*, *search engine* stb. – egynél több dokumentumot egy helyütt csakis akkor képes találni, ha azok előzőleg valamiféle csoportosításnak lettek alávetve. Ez a csoportképzés lehet kézi (pl. a Yahoo-nál az automatikus html-gyűjtést kézi bekötéssel egészítik ki, ami a Gopher *subject tree* „webesített" változata), gépi (pl. az AltaVista a gyakran használt oldalakat nagyobb valószínűséggel sorolja a kérdésre relevánsak közé) vagy vegyes technikájú (pl. az Eunet Galaxy gyakorisági alapon épít tartalmi fastruktúrát).

Ilyen körülmények között az információkeresés sikere legalább négy tényező kölcsönhatásán múlik. Ezek: az indexelés kérdése, a keresőmodell problémája, a navigáció mint fogalmi eltévelyedés, és a hiányzó tartalmi támpontok ügye. Az elsőt már vázoltam. A másodikhoz legfeljebb annyit kívánok hozzátenni, hogy a keresés ma ismert négy modellje közül – ezek a Boole-, a pontatlan logikai, a vektortér-, illetve a valószínűségi modell – a felhasználó számára egyetlen percre sem világos, melyik szolgáltatásban melyik érvényesül, vagy inkább melyek keveréke. A keresési folyamatot ez áttekinthetetlenné teszi, a találati listán szereplő ottlapok tömegét pedig esetlegessé. Említést érdemel az a háromdimenziós olvasás is, amelyet rejtélyes okból navigációnak neveztek el, s amelynek állandó emlegetése azt az érzetet keltheti, mintha úgy lennénk urai a helyzetnek, ahogyan *Tengerész Henrik* kortársai voltak urai a tengereknek. Valójában azonban a hajózás már a molukkák idején, szextánssal és asztrolábiummal is biztonságosabb révbe vezetett, mint az infonautika manapság. Mindennek közös oka a negyedik hiányosság: nevezetesen, a hajósoknak volt Sarkcsillaguk, és voltak csillagképeik, amelyekhez haladásukat mérhették, nekünk viszont nincsenek tartalmi konstellációink.

Mindez együttesen felveti, lehet-e az Interneten szaporodó információ leírására olyan rekurzív modellt találni, amely ugyanakkor az automatikus indexelés technikáival összhangban kereshető, és a keresés eredménye grafikusan láttatható, vagyis a felhasználó a keresés végső szakaszában „robotpilóta" helyett „kézi vezérlésre" térhet át. Egy ilyen modell részint egyszerűvé tenné a bonyolultat, másrészt áttekinthetővé a ma még áttekinthetetlent – létrehozná azokat a tartalmi csillagképeket, melyek a négy keresési modell valamelyikével bejárhatók.

Saltontól a tartalmi térképezésig

Az információ-háztartásnak azt a modelljét, mely a tartalmi bővülést vagy tágulást rekurzióval egyszerűsíti, a közelmúltban elhunyt *Gerard Salton* fogalmazta meg. Mivel elgondolásait dokumentumok automatikus indexelése és osztályozása során dolgozta ki, modellje dinamikus könyvtár néven vált ismertté. Az alábbiakban előbb néhány szóban ezt ismertetem, majd – általánosítása után – megmutatom, miként használható hálózati információ gyarapodásának leírására. Végül pillanatfelvételeket mutatok be adatbázisok információtartalmának eloszlásairól.

A dinamikus könyvtár

Salton elgondolása az volt, hogy a dokumentumok tartalmi feltárását is gépesítse, majd erre alapozza mind tárolásukat, mind visszakeresésüket [3, 4]. Erre a sokváltozós statisztika egyik módszerét, a klaszteranalízist használta.

Sokváltozós módszerek alkalmazásához az input adatokat mátrixban ábrázoljuk, melyeknek egy sora felel meg pl. egy dokumentumnak, egy oszlopa pedig a dokumentumhalmazon megfigyelhető egyik tulajdonságnak. Aszerint, hogy a szóban forgó ismérv jellemző-e az adott dokumentumra, a mátrixba 0-t vagy 1-et írunk[3]. A mátrix sorai a dokumentumvektorok, oszlopai a kulcsszó- (tulajdonság-) vektorok, rokon dokumentumok vagy összetartozó indexkifejezések keresése tehát egyaránt a vektortérmodellhez vezet.

Ezekről a módszerekről elegendő általánosságban annyit mondani, hogy esetükben a csoportelemzés különböző válfajairól van szó. Miként lehet egy csoport struktúráját magából az anyagból, tehát a megfigyelő előzetes ítéletalkotása nélkül megismerni? Állhat-e a csoport nagyon sok egyedből, és osztályozhatjuk-e ezeket nagyon sok tulajdonságuk alapján? Ezekre a kérdésekre válaszol a klaszteranalízis is, az elemzett sokaság, például dokumentumok hasonlóságait és különbségeit metrikus térbeli viszonyokra, közelségre és távolságra fordítva le. Az így készült összehasonlító ábrán két dokumentum minél közelebb esik egymáshoz, a tartalmuk annál hasonlóbb, és viszont. Ha viszont indexkifejezések térbeli viszonyait vizsgáljuk, a közelség fogalmi összetartozást takar. Mindez azonban csak a rendszer egy bizonyos állapotára igaz, mert ha a rendszer meg-

[3] Léteznek nem bináris technikák is, ezekkel azonban itt nem foglalkozom.

változik (dokumentumokat adunk hozzá vagy veszünk belőle el), az információbevitel vagy -veszteség következtében mind a dokumentumcsoportok szerkezete, mind a keresőkifejezések összetartozása megváltozhat. Más szóval a klaszterek súlypontja áthelyeződik. A rendszer két állapotának különbsége a kulcsszavakból képezett centroid vektorok egymástól mért távolságával arányos.

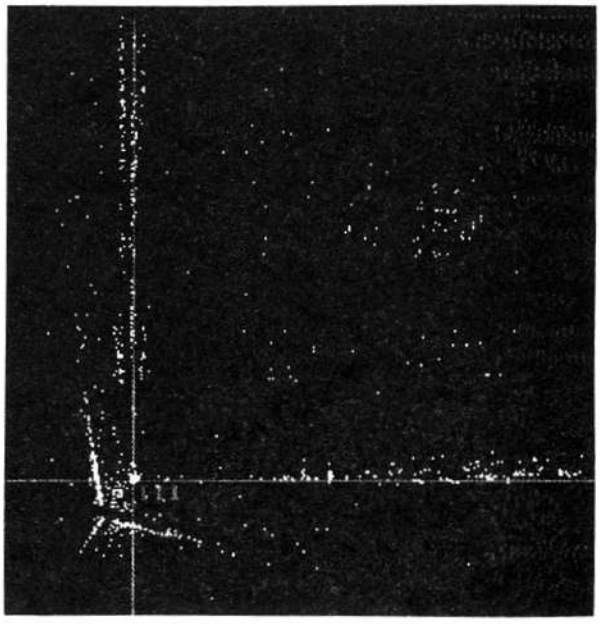
Mindebből két dolog következik. Először: nemcsak a dokumentumok, hanem a keresőkérdések is klaszterálhatók, a keresési szempontok változása pedig a keresőkérdések tematikus csoportjainak súlypontját mozdítja el. Végeredményben tehát olyan modellhez jutottunk, amelyben minden dinamikus, a rendszer „osztályai" (azaz klaszterei) mindenkor híven tükrözik az adott állapotot, ugyanakkor mindezt emberi beavatkozás nélkül, ami a tárolást és a visszakeresést az osztályozással és az indexeléssel egy logikai alapra helyezi. (Mindezt a könyvtárra mint intézményre vonatkoztatva, a gyarapodás változásai állapottér-változásokká alakulnak át, a változó tartalom változó térviszonyok képében jelenik meg, melyeket a keresések szintén változó térstruktúrájával kell megfeleltetnünk.) Másodszor: folyamatos gyarapodást feltételezve, a centroidok kiszámítása rekurzív módon, ugyanazokat a lépéseket ismételve történik.

A modell továbbfejlesztése

Az imént az alapkérdések során nem emeltem ki a csoportviszonyok láttatását, mely a statisztikai programcsomagoknak nem a legerősebb oldala. A saltoni modell is ebből a szempontból fejleszthető. Ezen a területen világszerte megélénkült a kutatás[4].

Norbert Wiener idősorelemzései óta ismeretes, hogy az analóg információ négy koordináta, *x*, *y*, *z*, és *t* megadásával definiálható [5]. Mivel *t* az időkoordináta értéke, mellyel most nem foglalkozom, a kérdés az, van-e olyan sokváltozós módszer, amely az *x*, *y*, *z* szemantikai koordinátáknak, mint egy folyamat pillanatnyi állapota kiszámítására képes[5]. Tapasztalataim szerint a főkomponens-analízis – bizonyos megszorításokkal, melyekre alább visszatérek – ilyen eljárás, ez pedig megnyitja az utat akár egyes adatbázisok, akár az Internet tartalmi térképezése felé, ha képes létrehozni a tájékozódáshoz szükséges tartalmi konstellációkat.

Az eredeti input mátrixot szorzatnak tekintve, a főkomponens-analízis kiszámítja a szorzandó, valamint szorzó mátrixot. Az egyiket a dokumentumok, a másikat a kulcsszavak eloszlásának tekintve, megkapjuk a keresett térkoordinátákat. Vagyis olyan fogalmi teret alakíthatunk ki, amelyben a dokumentumok csoportjai az egyes tételek szemantikai viszonyait tükrözik, kulcsszavaik csoportosulásai nemkülönben. Az így kialakított szemantikai tér a vektormodellel kereshető, azaz „hajózható" *(1., 2. ábra)*.

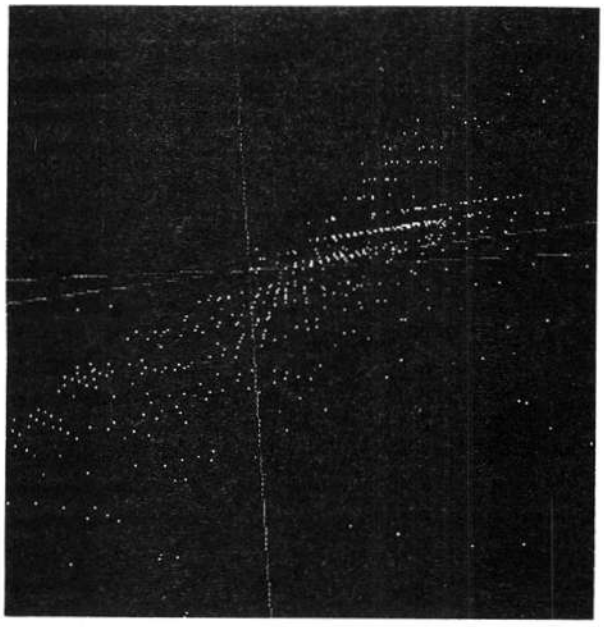


1. ábra **1389 dokumentum és 1839 kulcsszó tartalmi térképe** (legyezőszerű ponthalmaz az I-II tengelyek körül, illetve háromszögű eloszlás az origóban) [*Sophia* adatbázis, I = művészet, II = történelem/földrajz, III = filozófia]

Milyen lesz az az információs tér, amely egynél több adatbázist tartalmaz? Hogy ezt elképzeljük, ahhoz jó támpont a világegyetem szerkezete, mely egymásba ágyazott nagyságrendekkel láttatható. Eszerint Naprendszerünk a Tejút nevű galaxisban található, az viszont – mintegy húsz másik spirálköddel – az úgynevezett Helyi Csoportot alkotja. A Helyi Csoport azonban csupán töredéke a Helyi Szuperklaszternek, amely a megfigyelt univerzum közepe táján helyezkedik el, a peremvidéken észlelt kvazárokhoz – csillagszerű objektumokhoz – képest [6]. Ahogyan ebben a mintegy harmincmillió fényév átmérőjű, táguló gömbhalmazban égitestek állandó, csillagképeknek nevezett konstellációit látjuk, ugyanúgy bontakozik ki a szemantikai tér-

[4] Az érdeklődő az alábbi lapok bármelyikéről elindulhat: *http://www.cc.gatech.edu/gvu/softviz/infoviz/infoviz.html, http://websom.hut.fi/websom/, http://www.lis.pitt.edu /~isdept/faculty.html.*

[5] Ezt nem átvitt értelemben gondolom, hanem szó szerint. Mivel a sokváltozós módszerek bármilyen, tehát nem nyelvi eredetű vizsgálati anyag csoportjait is távolságviszonyok által fejezik ki, ezek értelmezése (szemantikájuk) az *x*, *y*, *z* koordinátahármas függvénye.



2. ábra **Kulcsszavak csoportosulása a fogalmi térben**

képezés során az összetartozó dokumentumok számos, egymásba ágyazott nagyságrendje. Ezeket nevezem első-, másod-, illetve felsőbb fokú morfológiáknak. Evolúciójuk, alakulásuk a saltoni modellel követhető [7].

Mindebből következik, hogy ha hagyományos dokumentumok helyett pl. ottlapok tartalmát írjuk le az input mátrixban, az *x*, *y*, *z* koordinátahármas kiszámolásával elvben az egész Internet tartalmi tere létrehozható. A negyedik, *t* koordináta a rendszer változásait köti időponthoz. Ekkor a tartalmi térkép változásának két osztályozás különbsége felel meg. Egy ilyen, táguló szemantikai térben az információkeresés a videojátékok űrutazásaira fog hasonlítani [8].

Más táguló modellek

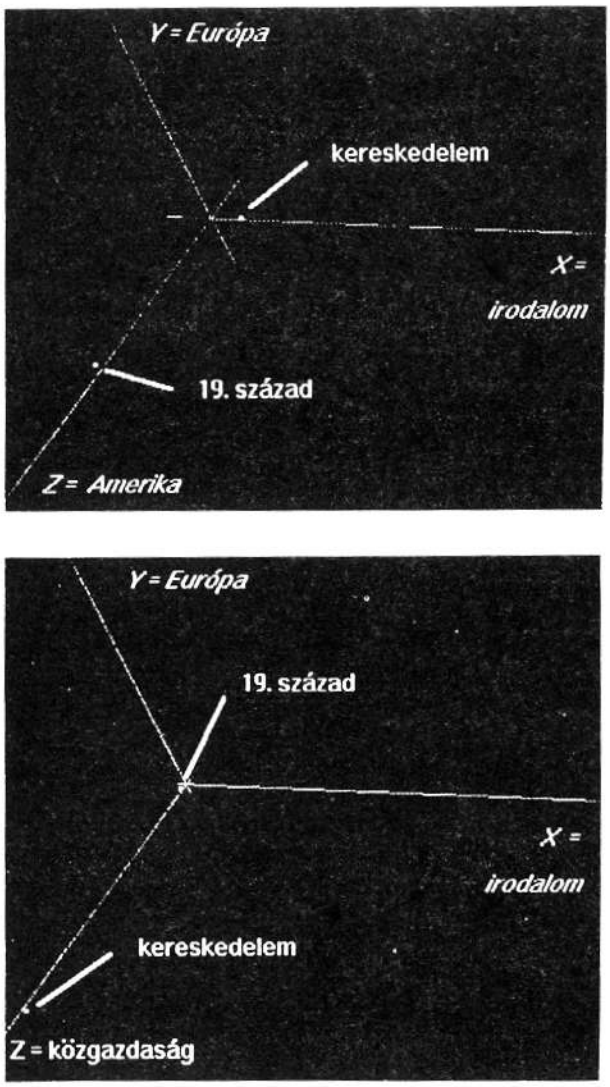
Az információs tér láttatásából általában következik, hogy a dinamikus könyvtár egybevethető a táguló világegyetem kozmológiai modelljeivel [9]. Ebben az értelemben a tartalmi galaxisok térképezését tekinthetjük az információcsillagászat előmunkálatainak. Ezt az elnevezést azonban csak metaforikusan használom; további vizsgálatoknak kell eldönteniük, vajon az érdekes hasonlóságok takarnak-e valódi, mélyebb összefüggéseket. Egy másik, öngerjesztő tágulási folyamat az emberi megismerés, ha a tartalom síkjából folyton a tartalom kontextusába lépünk ki, majd kezdődik minden elölről.

Kitekintés

A javasolt modell további előnyei:

1. A *Lemaître*-, majd *Gamow*-féle kozmológia, népszerű nevén „ősrobbanás-elmélet" ellenpárjává *Plótinosz* ismeretelméletét teszi: a kozmológiában egyből, a kezdeti szingularitásból[6] keletkezik sok, a megismerésben sokból egy („a megismerés ugyanis olyan látás, amely a kettőben látja az Egyet") [10]. A természettudományokban mindennapos ez a sokat a kevésre, a jelenségeket okukra, a variálódást néhány vagy egyetlen invariánsra visszavezető szemlélet.
2. A javasolt eljárás a szabályindukció révén kapcsolódik a tudás- vagy adatbányászathoz [11], illetve a szakértői rendszerek alkalmazásához. A szakértői rendszerek ismereti magja, úgynevezett tudásbázisa gyakran használja a produkciós szabálynak nevezett, „ha–akkor" feltételformalizmust: ilyen „ha–akkor" szabályok húzódnak meg egy-egy konkrét osztályozás hátterében is – *ha* egy objektum ilyen és ilyen ismérveknek megfelel, *akkor* ebbe és ebbe az osztályba tartozik. Osztályozás után ezt a feltételrendszert „kicsapatva", olyan hibrid rendszerek hozhatók létre [12], amelyek adatbázisokra vagy az Internetre egyaránt alkalmazhatók, ám ma nincs vizuális komponensük.
3. A láttatás lehetőségeinél fogva a tartalom és a virtuális valóság közötti szakadék áthidalható [13]. Ennek módja akár a VRML-szabvány (*Virtual Reality Modelling Language*), akár más szabványértékű konvenciók, például a Hyper-G követése.
4. A számításmenet problémáit, illetve ezek megoldásait nemrégiben taglaltam [14]. Háromdimenziós, vagy alacsony dimenziószámú megoldás esetén a tartalom ún. „fekete lyukakba" hull, több kulcsszó esik azonos koordinátákra. Túl magas dimenziószámú megoldás viszont a tartalmi viszonyokat zilálja szét. A *3. ábrán* olyan, $n = 100$ dimenziós megoldás eredményeit szemléltetem, amelyek tartalmi hűsége megfelelő, és háromdimenziós egyedi felhasználói alterekben mégis láttathatók. A kulcsszavak viszonyait megváltoztatja, hogyha pl. a „19. század", illetve a „kereskedelem" keresőkérdéseket „Európa" és az „irodalom" kontextusában „Amerika" vagy a „közgazdaság" felől szemléljük.

[6] A szingularitás a kozmológiában is matematikai absztrakció, valóságos megfelelőjéről keveset tudunk – megfelelője az őspont, amelyből minden keletkezett.



3. ábra *Sophia* adatbázis, alterek keresése

Köszönet

Köszönöm *Szabó Sándor*nak és munkatársainak (ELTE TFK Könyvtár Tanszék), hogy kutatóévemhez és e munka megírásához a feltételeket biztosították, *Kokas Károly*nak (JATE Központi Könyvtára) a hálózati információkeresés módszereiről folytatott beszélgetést, *Horváth Tibor*nak (OPKM) a lektori jelentésében javasolt módosításokat, melyek mondandóm pontosabb előadásában segítettek.

Irodalom

[1] DARÁNYI S.: Quo vadis, bibliothecarius digitalis? = Bajza J.–Tóth B. (Szerk..): Networkshop'95 konferencia anyaga (IIFP) Budapest, 1995. p. 72–73.
[2] DEMPSEY, L.: Networking for Libraries. A Seminar at Libtech International '94 (Learned Information) London, Appendix V. 1994.
[3] SALTON, G.: Automatic information organization and retrieval. McGraw – Hill, New York, 1968.
[4] SALTON, G.–McGILL, M. J.: Introduction to Modern Information Retrieval. McGraw – Hill, New York, 1983.
[5] HAUFFE, H.: Die Informationsgehalt von Theorien. Springer, Wien, 1981. p. 10 [11].
[6] MARIK M. (Szerk.): Csillagászat. Akadémiai Kiadó, Budapest, 1991.
[7] DARÁNYI S.: Az automatikus osztályozástól a magasabb fokú morfológiákig. = Könyvtári Figyelő, 37. köt. 3. sz. 1991. p. 418–422.
[8] KORFHAGE, R. R.: BROWSER – A concept for visual navigation of a database. = IEEE Computer Society workshop for visual languages (IEEE) Washington, 1986. p. 143–148.
[9] FERRIS, T.: A vörös határ. A Világegyetem szélének kutatása. Gondolat, Budapest, 1985.
[10] PLÓTINOSZ: Az Egyről, a szellemről és a lélekről. Európa Könyvkiadó, Budapest, 1986. p. 231.
[11] PIATETSKY-SAPHIRO, G.–FRAWLEY, W. J. (Eds.): Knowledge Discovery in databases. AAAI Press – The MIT Press, Menlo Park, Ca. – Cambridge, Ma. 1991.
[12] BIELAWSKI, L.–LEWAND, R.: Intelligent systems design: integrating expert systems, hypermedia, and database technologies. Wiley, New York, 1991.
[13] DARÁNYI, S.–ZAWIASA, R.–HAJNAL, Z.: Conceptual mapping of a database in the humanities: first results of an experiment with *Sophia*. = Journal of Documentation, 52. köt. 1. sz. 1996. p. 86–99.
[14] DARÁNYI, S.–PREMINGER, M.–HAJNAL, Z.: Bubble retrieval: A geometric approach to automated classification and information visualization. Az ALA SIGIR '97 konferenciára (Philadelphia, USA) beküldött kézirat.

Beérkezett: 1997. III. 5-én.