INFORMÁCIÓCSILLAGÁSZAT AZ INTERNETEN: ELMÉLET ÉS GYAKORLAT

Darányi Sándor, *daranyi@kazy.elte.hu*
ARIST BT, Budapest

Abstract

Information retrieval on the Internet suffers from insufficient indexing, the opaqueness of retrieval models, and the misconcept of navigation. In order to enable true three- or four-dimensional navigation, we must construct semantic universes first, which represent the spatial arrangement of domain-specific knowledge. Such spatial content maps can be constructed following the guidelines of Gerard Salton's dynamic library model. This is a recursive model which applies to any system of classifications changing over time, and can be used for the grouping of electronic documents as well. I suggest that by replacing cluster analysis with principal component analysis in the original model, information visualization becomes possible. The results, robust distributions of both documents and keywords, resemble stellar configurations and pave the way for a postulated information astronomy.

1. 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 nõ, a növekvõ tömegû 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 kettõ azonban egyik sem. A második robbanás tehát bizonyos gyakorlati és elméleti kérdéseket egyaránt felvet. Az mindenki számára világos, hogy egy ott-lap, ftp archivum vagy adatbázis esetében a tartalomnak adunk lokalizálható formát. Az elektronikus címhez kötött elektronikus tartalom azonban mára az érdeklõdés középpontjába állítja a hálózati információforrások indexelését és visszakeresését, hiszen minél nagyobb tömegû 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 headertõl 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 minõségére visszahat. A megoldás

[1] Ld. az Internet Society statisztikáját (1995. aug. 2). 1995 elsõ 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 becsülhetõ.

tehát csakis valamiféle indexelés lehet, tartalmi sûrítményekkel, felettes fogalmakkal megcímkézett elektronikus dokumentumok és dokumentum-fájlok létrehozása, automatikusan gyarapodó szövegegyüttes esetében nyilván automatikus indexeléssel [2].

2. Mi a probléma?

Amennyire ez a világhálózat fejlõdésének ma még kusza és feldolgozatlan történetébõl kiszûrhetõ, az elmúlt esztendõkben népszerûvé 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 lehetõségét tesszük hozzá, megkapjuk az ftp-t; ezt menükkel és kereszthivatkozásokkal kiegészítve, a gopher-hez jutunk; a kereszthivatkozásokat hipertextként kezelve létrejön, majd - grafikus felületen - szalonképes külsõt ölt a WWW.

Valamennyi felsorolt szolgáltatás visszakeresési oldala olyan software-t 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 egyhelyütt csakis akkor képes találni, ha azok elõzõleg valamiféle csoportosításnak lettek alávetve. Ez a csoportképzés lehet kézi (pl. a Yahoo-nál az automatikus html-gyûjtést kézi bekötéssel egészítik ki, ami a gopher *subject tree* "webesített" változata), gépi (pl. az AltaVista a gyakran használt oldalakat nagyobb valószínûséggel 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 tényezõ kölcsönhatásán múlik. Ezek: az indexelés kérdése, a keresõmodell problémája, a navigáció mint fogalmi eltévelyedés és a hiányzó tartalmi támpontok ügye. Az elsõt 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 valószínûségi 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 szereplõ ott-lapok tömkelegét 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 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 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 kereshetõ, és a keresés eredménye grafikusan láttatható, vagyis a felhasználó a keresés végsõ szakaszában "robotpilóta" helyett "kézi vezérlésre" térhet át. Egy ilyen modell részint egyszerûvé tenné a bonyolultat, másrészt áttekinthetõvé a ma még áttekinthetetlent - létrehozná azokat a tartalmi csillagképeket, melyek a négy keresési modell valamelyikével bejárhatók.

3. Saltontól a tartalmi térképezésig

Az információháztartásnak azt a modelljét, mely a tartalmi bõvülést vagy tágulást rekurzióval egyszerûsíti, 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 elõbb néhány szóban ezt ismertetem, majd - általánosítása után - megmutatom, miként használható hálózati információ gyarapodásának leírására. Végül pillanatfelvételeket mutatok be adatbázisok információtartalmának eloszlásairól.

3.1 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 megfigyelhetõ egyik tulajdonságnak. Aszerint, hogy a szóban forgó ismérv jellemzõ-e az adott dokumentumra, a mátrixba 0-t vagy 1-t írunk[3]. A mátrix sorai a dokumentumvektorok, oszlopai a kulcsszó- (tulajdonság-) -vektorok, rokon dokumentumok vagy összetartozó indexkifejezések keresése tehát egyaránt a vektortér-modellhez vezet.

Ezekrõl a módszerekrõl elegendõ általánosságban annyit mondani, hogy esetükben a csoportelemzés különbözõ válfajairól van szó. Miként lehet egy csoport struktúráját magából az anyagból, tehát a megfigyelõ elõzetes ítéletalkotása nélkül megismerni? Állhat-e a csoport nagyon sok egyedbõl, és osztályozhatjuk-e ezeket nagyon sok tulajdonságuk alapján? Ezekre a kérdésekre válaszol a klaszteranalízis is, az elemzett sokaság, például dokumentumok hasonlóságait és különbségeit 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 megváltozik

(dokumentumokat adunk hozzá vagy veszünk belõle el), az információbevitel vagy -veszteség következtében mind a dokumentumcsoportok szerkezete, mind a keresõkifejezések összetartozása megváltozhat. Másszóval a klaszterek súlypontja, centroidja áthelyezõdik. A rendszer két állapotának különbsége a centroidok egymástól mért távolságával arányos.

Mindebbõl két dolog következik. Elõször: nemcsak a dokumentumok, hanem a keresõkérdések is klaszterálhatók, a keresési szempontok változása pedig a keresõkérdések tematikus csoportjainak súlypontját mozdítja el. Végeredményben tehát olyan modellhez jutottunk, amelyben minden dinamikus, a rendszer osztályai 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.

3.2 A modell továbbfejlesztése

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

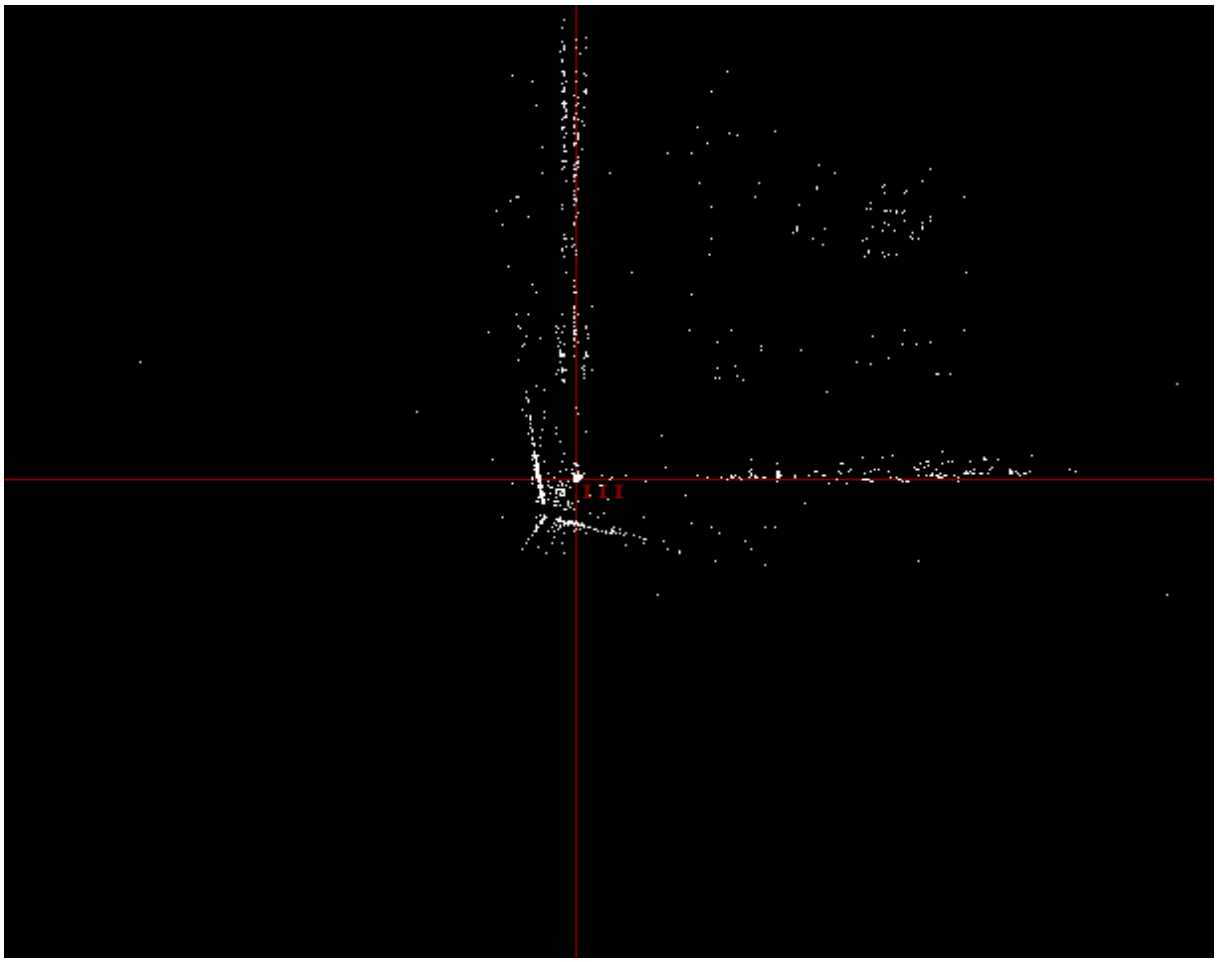
---

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

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

Norbert Wiener óta ismeretes, hogy az információ négy koordináta, *x*, *y*, *z*, és *t* megadásával definiálható [5]. Mivel *t* az idókoordináta értéke, mellyel most nem foglalkozom, a kérdés az, van-e olyan sokváltozós módszer, amely az *x*, *y*, *z* szemantikai koordináták kiszámítására képes [5]. Tapasztalataim szerint a fõkomponensanalízis ilyen eljárás, ez pedig megnyitja az utat akár egyes adatbázisok, akár az Internet tartalmi térképezése felé, amennyiben képes létrehozni a tájékozódáshoz szükséges tartalmi konstellációkat.

Az eredeti input mátrixot szorzatnak tekintve, a fõkomponensanalízis 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 kereshetõ, azaz „hajózható” (1., 2. ábrák).



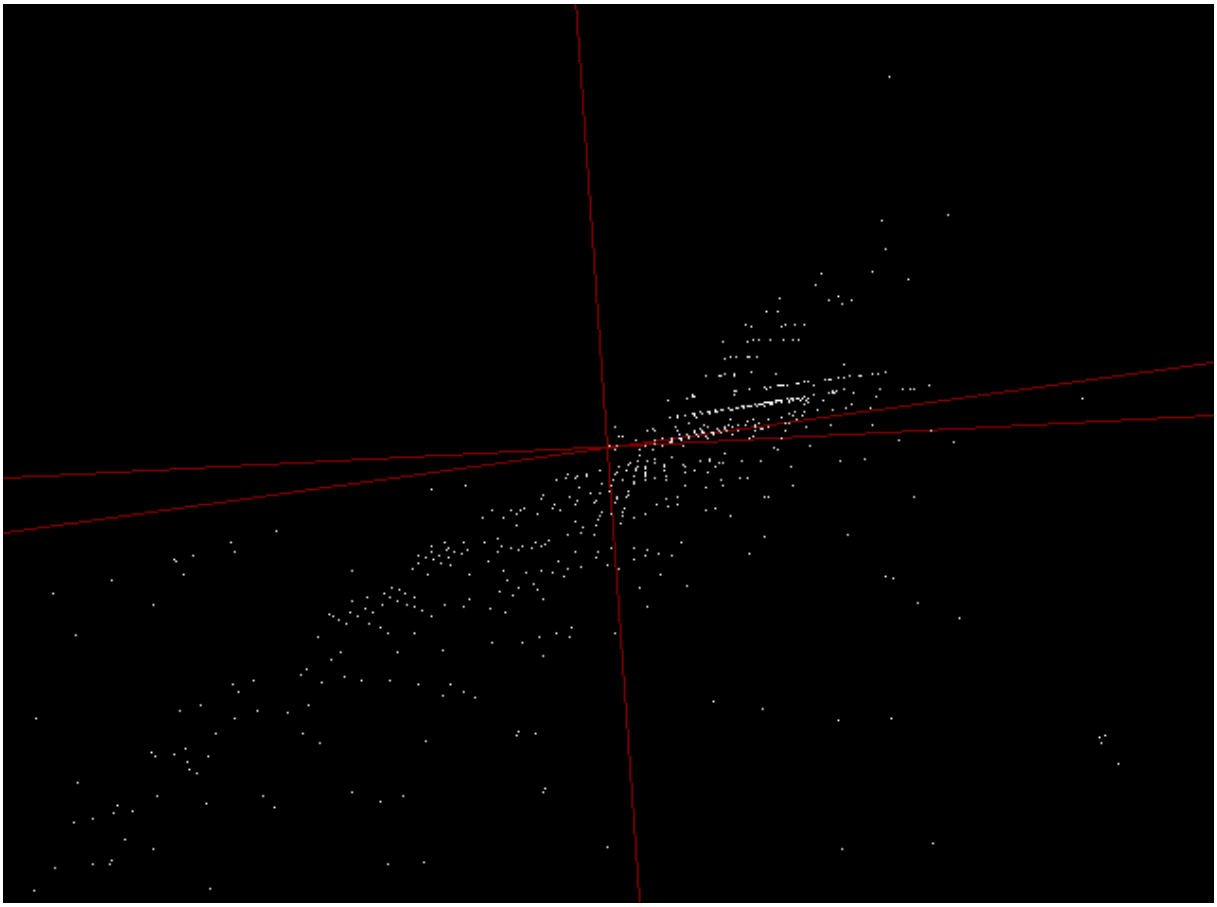
1. ábra: 1389 dokumentum és 1839 kulcsszó tartalmi térképe (legyezõszerû ponthalmaz az I-II tengelyek körül, illetve háromszögû eloszlás az origóban) [*Sophia* adatbázis, I = mûvészet, II = történelem / földrajz, III = filozófia]

Milyen lesz az az információs tér, amely egynél több adatbázist tartalmaz? Hogy ezt elképzeljük, ahhoz jó támpont a Világegyetem szerkezete, mely egymásba ágyazott nagyságrendek-kel láttatható. Eszerint

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

Naprendszerünk a Tejút nevû 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 - csillagszerû objektumokhoz - képest [6]. Ahogyan ebben a mintegy harmincmillió fényév átmérõjû, táguló gömbhalmazban égitestek állandó, csillagképeknek nevezett konstellációit látjuk, ugyanúgy bontakozik ki a szemantikai térképezés során az összetartozó dokumentumok számos, egymásba ágyazott nagyságrendje. Ezeket nevezem elsõ-, másod- illetve felsõbb fokú morfológiáknak. Evolúciójuk, alakulásuk a saltoni modellel követhetõ [7].

Mindebbõl következik, hogy ha hagyományos dokumentumok helyett pl. ott-lapok 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 idõponthoz. Ekkor a tartalmi térkép változásának két osztályozás különbsége felel meg. Egy ilyen, táguló szemantikai térben az információkeresés a videojátékokûrutazásaira fog hasonlítani [8].



2. ábra: Kulcsszavak csoportosulása a fogalmi térben

4. Más táguló modellek

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

5. Kitekintés

A javasolt modell három további elõnyét szeretném kiemelni:

1. A Hoyle-féle kozmológia, népszerû nevén „õsrobbanás-elmélet" ellenpárjává Plótinosz ismeretelméletét teszi: a kozmológiában egybõl, a kezdeti szingularitásból keletkezik sok, a megismerésben sokból egy („a megismerés ugyanis olyan látás, amely a kettõben látja az Egyet") [10]. A természettudományokban mindennapos ez a sokat a kevésre, a jelenségeket okukra, a variálódást néhány vagy egyetlen invariánsra visszavezetõ szemlélet.

2. A javasolt eljárás a szabályindukció révén kapcsolódik a tudás- vagy adatbányászathoz [11], illetve a szakértõi rendszerek alkalmazásához. Így olyan hibrid rendszerek hozhatók létre [12], amelyek adatbázisokra vagy az Internetre egyaránt alkalmazhatók, ám ma nincs vizuális komponensük.

3. A láttatás lehetõségeinél fogva a tartalom és a virtuális valóság közötti szakadék áthidalható [13].

5. Köszönet

Köszönöm Dr. Szabó Sándornak és munkatársainak (ELTE TFK Könyvtár Tanszék), hogy e munka megírásához a feltételeket biztosították, Kokas Károlynak (JATE Központi Könyvtára) a hálózati információkeresés módszereirõl folytatott beszélgetést.

6. Irodalom

[1] Darányi S. (1995): Quo vadis, bibliothecarius digitalis? In: Bajza J. - Tóth B. /Szerk../: Networkshop'95 konferencia anyag (IIFP) Budapest, 72-73.

[2] Dempsey, L. (1994): Networking for Libraries. A Seminar at Libtech International '94 (Learned Information) London, Appendix V.

[3] Salton, G (1968): Automatic information organization and retrieval. (McGraw - Hill) New York.

[4] Salton, G. - McGill, M.J. (1983): Introduction to Modern Information Retrieval. (McGraw - Hill) New York.

[5] Hauffe, H. (1981): Die Informationsgehalt von Theorien. (Springer) Wien, 10 [11].

[6] Marik M. /Szerk./ (1991): Csillagászat. (Akadémiai Kiadó) Budapest

[7] Darányi S. (1991): A z automatikus osztályozástól a magasabb fokú morfológiákig. Könyvtári Figyelõ 1=37(3), 418-422.

[8] Korfhage, R.R. (1986): BROWSER - A concept for visual navigation of a database. IEEE Computer Society workshop for visual languages (IEEE) Washington, 143-148.

[9] Ferris, T. (1985): A vörös határ. A Világegyetem szélének kutatása. (Gondolat) Budapest.

[10] Plótinosz (1986): Az Egyrõl, a szellemrõl és a lélekrõl. (Európa Könyvkiadó) Budapest, 231.

[11] Piatetsky-Saphiro, G. - Frawley, W.J. /Eds./ (1991): Knowledge Discovery in Databases. (AAAI Press - The MIT Press) Menlo Park, Ca. - Cambridge, Ma.

[12] Bielawski, L. - Lewand, R. (1991): Intelligent Systems Design: Integrating Expert Systems, Hypermedia, and Database Technologies. (Wiley) New York.

[13] Darányi, S. - Zawiasa, R. - Hajnal, Z. (1996): Conceptual Mapping of a Database in the Humanities: First Results of an Experiment with *Sophia*. Journal of Documentation,

52, 1, 86-99.