Az automatikus osztályozástól a magasabb fokú morfológiákig

Darányi Sándor

1. Előszó

Előljáróban néhány szót szeretnék szólni arról, miként kapcsolódik ez az írás *Horváth Tibor* kandidátusi értekezéséhez.[1]

Minden információkereső rendszer próbaköve a szemantika - az a tény, mennyire értelmes közléseket kapunk vissza a számítógéptől. Horváth Tibor annak a felfogásnak a híve, amely ezt a szemantikát nem a keresőkifejezések közötti logikai kapcsolatok révén közelíti meg (ezeket a *Boole-algebra* definiálja), hanem az indexkifejezések logikai kapcsolatait építi be a rendszerbe a *Chafe-* illetve *Fillmore*-féle mélyesetgrammatika operátoraival. Értekezése ezt a gyakorlatot mutatja be, a PRECIS honosítása példájával.

Az ide vezető gondolatmenet egyik állomása a statisztikai automatikus osztályozás beillesztése az osztályozástörténet vonulatába. Jómagam ezzel kapcsolatban invitálom az olvasót egy kis kitérőre, egyéb, de szintén a könyvtártudomány megfontolásai közé tartozó kérdésekből ajánlva ízelítőt. E dolgozat eredetileg 1990-ben Darmstadtban hangzott el, a *Society for Content and Concept Analysis by Computer* (SCCAC) közgyűlésén.[2]

2. Bevezető megfontolások

Az emberi megismerés mércéinek egyike az, mennyire vagyunk képesek elképzelni és láttatni a láthatatlant. A mikroszkópok és teleszkópok ezen képességünk megbecsült eszközei. Ha mikro- vagy makroszintű koordinátarendszereket figyelünk meg, mind a mikroszkóp, mind a távcső a puszta szemmel észrevehetetlen morfológiákról lebbenti fel a fátylat. Bármely tudományos vizsgálat alapvető feladata az ilyen morfológiák feltérképezése. E térképezés vezet ahhoz, amit rendszerint egy tudományterület fejlődésének szoktunk nevezni. Általánosságban elmondhatjuk, hogy a felfedezett alakzatok megbízhatóan állandó volta a megfigyelés dimenziószámával arányos. Ugyanis minél részletesebben írunk le és hasonlítunk össze két objektumot, annál valószínűbb, hogy az eredmény megközelíti az elméletileg elérhető tökéleteset.

Különösképpen láthatatlan és nehezen osztályozható az emberi tudás, mindazon fogalmak

univerzuma, melyek segítségével a környező világot tükrözzük magunkban. Összhangban a megismerés gyakorlatával, megtehetjük, hogy erre a fogalmi - belső - világegyetemre mint égitestek rendszerére hivatkozzunk, megpróbáljuk feltérképezni együttállásaikat, illetve utazzunk közöttük. Probléma azonban az, hogyan alakítsunk át egy fogalmi sémát olyan alakzattá, amely elég állandó ahhoz, hogy eltűrje a térképezést? A következőkben amellett érvelek, hogy ennek egyik lehetséges módja a Horváth Tibor által is leírt statisztikai automatikus osztályozási módszerek használata. Az ilyen módszerek az input adatoktól független eredményre vezetnek, azaz nem módosítják a megfigyelt jelenséget.[3]

Megjegyzem, ez a dolgozat az információt olyan fizikai entitásként értelmezi, amelyre vonatkozik az entrópia törvénye. Az entrópia mint a bizonytalanság mértéke a rendezetlenség, rendetlenség vagy véletlen formáját ölti. Párhuzamként kínálkozik a statisztikai mechanika esete, melyben a molekulák szerveződése a róluk szóló tudássá alakul át, s az entrópia növekedése e tudás csökkenésével jár.[4] Eléggé figyelemreméltó, hogy mind az információ tudományában, mind a statisztikai mechanikában a rendezetlenség és a káosz, a tudás hiánya akadályozza meg, hogy előre megmondjuk az események kimenetelét.

3. A statisztika mint a megfigyelés eszköztára

3.1. Osztályozás klaszter- és faktoranalízissel

A statisztikai módszerek közül a klaszter- és faktoranalízis osztályozási képességeire, és az eredmények értelmezésére szeretnék összpontosítani. Az osztályozás objektumok olyan csoportosítási folyamatát jelenti, melynek során a csoportok tagjai tulajdonságaikat tekintve hasonlítani fognak egymásra. Míg a klaszteranalízis politétikus, többrétű, rendezetlen osztályokat alakít ki, a faktoranalízisban az osztályozás nem domborul ki, noha benne rejlik az eredményben.[5]

E „benne rejlő", inherens osztályozáson azokat a klaszterokat értem, amelyek mint objektumok vagy megfigyelési változók csoportjai tűnnek fel a háromdimenziós faktortérben (a többi faktor nem láttatható). A csoporttagság tekintetében mind a klaszter-, mind a faktoranalízis eredményei egybevethetők, és egymással többé-kevésbé azonosak. A „többé-kevésbé" kifejezés arra a bizonytalansági szindrómára utal, amelyhez hamarosan visszatérek.

A faktoranalízis előnye a klaszteranalízissel szemben, hogy a faktortér három, egymásra merőleges tengelye, valamint a teret „kifeszítő" objektum-csoportok között inherens oksági összefüggés van. Azt mondhatnánk, a faktorok az okai annak az osztályozásnak, amely a csoportok tagjait egymáshoz rendelte. Ha nem sikerül szavakra lefordítanunk ezeket a faktorokat, az oksági összefüggés nem egykönnyen látható be, ezért az objektumok illetve tulajdonságok klasztereit hozzá szokás rendelni a tárgyalási univerzum tengelyeihez. Ennek egyik lehetséges módja a faktorforgatás[6], a másik a blokk-klaszterálás kiegészítése faktoranalízissel[7].

3.2. Statisztika és okság

Míg a faktoranalízisban könnyű statisztikai okságot látni, hiszen a kikövetkeztetett faktorokból vagy okokból a standardizált input mátrix rekonstruálható, ez a metafora az igazi okságra csak némi fejtörés után alkalmazható.

Azzal érvelünk, hogy a megismerés során a változatlan, invariáns tudás keresése megfelel azon állandó szabályok kutatásának, amelyek a jelenségek létét szabályozzák. Így a tudás az információ okának tekinthető - olyan oknak, amelyről gyarapodó következményei révén szerzünk tudomást.[8]

A faktoranalízis során az okok struktúráját következtetéssel tárjuk fel, az eredmények struktúrája felől. Ilyenkor az input mátrixból végzett számítások minden lépése eredménye az előzőeknek és oka is a rákövetkezőknek. Ugyanakkor a számítások helyességének mércéje, vajon a kikövetkeztetett okstruktúrából a megfigyelési adatok eredeti eloszlása, az input mátrix rekonstruálható-e. Eszerint legalábbis beszélhetünk arról, hogy bemenő, input adatainkat a faktorsúlyok „okozták". Másfelől, az a kérdés, hogy a faktorstruktúra révén elérkeztünk-e a stabil, változatlan tudáshoz a kezdeti információk helyett, voltaképpen azt feszegeti, vajon eredményeink invariánsak-e. Ez csiklandós kérdés, hiszen a különböző klaszteráló és faktorizáló algoritmusok „szeretnek" különböző eredményekkel meglepni minket.

4. Az információ mint világegyetem

Az alábbiakban a faktor- illetve főkomponens-analízissel megmutatott tárgyalási univerzumot olyan *Newton* utáni információs világegyetemnek tekintjük, amelyben a megfigyelő nem képes önnön szubjektivitásától független megfigyelésekre. Ez a függés esetünkben a kutató döntéseként jelenik meg, milyen algoritmust választ az elemzéshez. A választott eszköz kihat az eredményre, amely így nem lesz független a megfigyelőtől.

A matematikai statisztikában e függés megszüntetésének egyik módja az, ha az invariánssal az eredmények statisztikailag szilárd, az algoritmusoktól független, közös részét azonosítjuk. Ekkor különböző algoritmusokat használunk ugyanarra a célra, majd az osztályok azon részét fogadjuk el invariánsnak, amelyek mindegyik algoritmusnál előfordultak. Ez a fogás a megfigyelőtől kevésbé függő eredményekre vezet, csökkenti az osztályozás bizonytalanságait, vagyis a tanulmányozott jelenség entrópiáját.

Az eddigieket összegezve, a faktor- és klaszteranalízis eredményeinek statisztikai szilárdságát fokozván eljuthatunk a jelenségek állandó csoportosításaihoz és ezek invariáns okszerkezetéhez, miközben megőrizzük a bemenő megfigyeléseket mint következmény-struktúrát.

5. A másod- és felsőbb fokú morfológiák fogalma

A behatolás a láthatatlanba, valamint az invariáns keresése hagyományosan rokon próbálkozások. Míg invariánsokat számos tudományterületen mutattak ki - itt meg kell elégednünk néhány hivatkozással (Gould 1979, Ivanov és Toporov 1976, Propp 1975) -, addig a megfigyelő eszköztárának kiterjesztése az optikától a statisztikáig aránylag kései fejlemény. E két erőfeszítés találkozik a különféle méretű és alakú morfok, alakzatok észlelésében.

5.1. Meghatározások

Morfon vagy alakzaton fizikai vagy fogalmi tulajdonságoknak olyan állandó halmazát fogjuk érteni, amelyeket relációk hasonlóan stabil rendszere köt össze. E tulajdonságok legyenek egy hálózat csomópontjai, relációik pedig a hálózat élei! Ezzel az eljárással akár egy test fizikai körvonalai, akár egy fogalom nem-fizikai alakja megrajzolható. Következésképp morfológián alakzatok olyan rendszerét értjük, amelyek elkészíthetők, sokszorosíthatók, csonkolhatók és megsemmisíthetők.

Az első- és magasabb fokú morfológiák egymástól csak csomópontjaik összetettségében különböznek. Az elsőfokú morfológia olyan rendszert jelent, amelyben a háló csomópontjai magányos objektumok vagy fogalmak (megfigyelési esetek), vagy tulajdonságok (megfigyelési változók). Ezt a fajta hálót szokás struktúrának, szerkezetnek is nevezni. A másodfokú morfológia ezek után olyan szuperstruktúra, amelynek csomópontjai az objektumok, fogalmak vagy tulajdonságaik csoportjai, klaszterei. Ugyanezen gondolatot folytatva, harmadfokú morfológián olyan hiperstruktúrát érthetünk, amelynek csomói elsőfokú koordinátarendszerek, univerzumok, az ultrastruktúra hálózatának metszéspontjai pedig maguk is hiperstruktúrák.

Más szóval, egy másodfokú szuperstruktúra a struktúrák struktúrája, a harmadfokú a szuperstruktúrák térszerkezete lesz, és így tovább.

5.2 A magasabb fokú morfológiák ismérvei

Meghatározásánál fogva a magasabb fokú morfológia szerkezet, struktúra tehát, - ugyanakkor nem minden hálózat nevezhető morfnak. Úgy gondolom, az igazi morfot vagy alakzatot legalább a következő négy ismérv együttes előfordulása jellemzi:

(a) láthatatlan,
(b) változatlan, invariáns,
(c) szemantikát hordoz,
(d) dimenziókkal, összehasonlítható tulajdonságokkal rendelkezik.

Láthatatlanságon olyasvalamit értek, ami a megfigyelés mai határain túl esik.

Az invariancia vagy stabilitás egy jelenség azon részét fejezi ki, amely mentes a változástól mint folyamattól.

Ami a szemantika vagy értelem hordozását illeti, a természetes nyelvektől a zenén át a genetikai kódig a lineáris (időbeli) közlés minden for-

májának van egy hordozó komponense, a kommunikáció formális, alaki oldala, meg egy hordozott része, az alaki oldalhoz rendelt szemantika. E két nézetet olyan, nyelvtani (grammatikai) csatoló köti össze, amely egymáshoz rendeli a szavakat, állandósult szókapcsolatokat, mondatokat, meg a nekik tulajdonított jelentést.

A dimenzionalitás arra a tapasztalatra utal, hogy a morfokat tulajdonságaik révén írhatjuk le. Magyarán szólva, tulajdonságai nélkül se az objektum, se a szubjektum nem önmaga. Az ismérvek nélküli lényt nehéz elképzelni.

Ha a vizsgált tárgy ezeknek az ismérveknek megfelel, nézetem szerint magasabb fokú alakzattal van dolgunk.

5.3 Néhány példa

E kritériumok alapján, az ember elsőfokú alakzatok részhalmazainak tekintheti a fogalmi (*conceptual structure*) vagy kognitív struktúrákat (*cognitive structure*), bogokat (*clump*), szemantikai hálókat (*semantic network*), szemantikai térképeket (*semantic road map*), asszociációs pályákat (*association trails*) és idézettségi hálókat (*citation network*). Míg ezek mindegyike láthatatlan és dimenzionált, tulajdonságaik száma tekintetében nagy közöttük a különbség, és közülük számos nem invariáns. A szemantikai struktúrák okságinak tekinthetők - a mélyszerkezet okozza a felszíni szerkezetet -, az idézési struktúrák viszont nem. A hypertext hasonlítható a magasabb fokú morfológiákhoz, amennyiben dimenziói vannak, sokféle szemantikát hordoz, azaz poliszémikus, és esetleg ok-okozati is, de nem invariáns és szerkezetét nem mi csapatjuk ki a láthatatlanból, hanem mesterségesen generáljuk. Mindez az invariancia döntő szerepére mutat rá, melynek az eredmények statisztikai szilárdsága (robustness) jól megfelel.

Eddig elsőfokú morfológiákkal foglalkoztunk. Terjesszük ki most az analógiát magasabb fokúakra is, a csillagászat mint metafora segítségével.

6. A csillagászat mint metafora

Párhuzamként máskülönben láthatatlan morfológiák megfigyelésére, korábban már felhoztam a mikroszkópot és a távcsövet. Most azt állítom, hogy a csillagászati világegyetem szerkezetének nagyságrendjei afféle mérőlécül szolgálhatnak, ha el akarjuk képzelni az emberi tudás nem-fizikai univerzumát.

Idestova száz esztendeje, a csillagászat az égbolt feltérképezésével vesződik, azzal, hogyan dönthetné el az égi alakzatok pontos távolságát a Földtől. Ezért milliónyi felvételt készítettek mind az északi, mind a déli égboltról. A probléma a fényképekkel az, hogy kétdimenziósak, és az általuk szerezhető kétféle alapvető információ, az égitest helyzete és fényereje nem elegendőek ahhoz, hogy elhelyezzük őket a térben, a harmadik kiterjedés mentén. Egy öreg, pislákoló csillag, bármilyen közel van hozzánk, éppoly homályos lehet a fényképen, mint heves és ifjú társa, amely a távolban lobog.

Hogy ezen a nehézségen felülkerekedjenek, a csillagászok növelték megfigyeléseik dimenziószámát. A különböző égi nagyságrendekben különféle indikátorokat felfedezve - úgynevezett *cefeida* csillagokat használnak a galaxisok távolságának mérésére, a színkép vöröseltolódását, a *Doppler-effektust* a nagyobb alakzatok osztálybasorolására sebességük, koruk és helyzetük szerint -, máig mintegy tizenöt milliárd fényévnyi távolságig sikerült felderíteniük az univerzum szerkezetét. Így a dimenzionalitás segített fellebbenteniük a láthatatlan fátylát.

De nem kevésbé segített a látható fény hullámhossza alatti vagy feletti megfigyelés eszköztárának fejlődése. A rádióteleszkópok, infravörös és ultraibolya távcsövek ugyanazon térkoordinátákban a látható tartományból már ismert jelenségek más térbeli eloszlását mutatták, hozzájárulva ezzel az eredmények invarianciájához, mely a megfigyelések közös része lett.

Ezzel az eljárással lépésről lépésre derült ki az ismert világegyetem térszerkezete. Ma egymásba ágyazott nagyságrendek sorozatát különböztethetjük meg, a következőképpen. A *Naprendszer*, 20 fényéven belül eső legközelebbi szomszédaival a *Tejút* nevű galaxis része. Az ilyen galaxisokban, a miénkhez hasonló globuláris klaszterek milliói találhatók. A következő nagyságrendekben a Tejút mindössze egyike annak a néhány tucat galaxisnak, amelyek az *Andromédával* és a *Magellán-felhőkkel* közösen az úgynevezett *Helyi Csoportot* alkotják. A Helyi Csoport azonban ismét csupán egyik eleme a *Helyi Szuperklasztert* felépítő sokaságnak, az is-

mert világegyetemnek majdnem a közepén található, nagyon messze a megfigyelhető eseményhorizonttól, amely zsúfolt a tőlünk sebesen távolodó galaxisokkal és szuperklaszterekkel. Amit pedig a „láthatáron" felismerünk, az a kavazárok, a kvázi-csillagok raja, és az ősrobbanás egyenletes háttérsugárzása.

Megdöbbentő hasonlatosságot látok a világegyetemet egymásba ágyazott hiperstruktúrának láttató csillagászati gyakorlat, valamint a faktor- és klaszteranalízis között, mely utóbbi képes hasonlóan szilárd hiperstruktúrák létrehozására egy másik végtelen és láthatatlan, az emberi megismerés feltérképezése által. A magasabb fokú morfológiák analóg szerkezetükkel analóg célokat fejeznek ki. A két kutatási terület nyilvánvalóan rokon, hiszen a sokváltozós osztályozás módszereit a csillagászat is használja.[9] Ugyanakkor a statisztika is a megfigyelés új szakaszába lépett azzal, hogy a jelenségek magasabb fokú aggregációit, csoportosulásait kezdte vizsgálni.[10]

7. Magasabb fokú morfológiák megfigyelése

Gerald Salton automatikus könyvtára óta ezek a gondolatok a könyvtártudomány számára is relevánsak. Ma már nem új ötlet olyan információs mindenséget építenünk, amelyben a felhasználó - a videojátékok űrsétái és statisztikai eszközök keveréke által - maga kormányozhatja kereséseit egy, a kép nyelvére lefordított adatbázisban. Mind a téridőklaszterálás[11], mind az idősorok elemzésének és modellezésének olyan programcsomagjai, amilyen a LISREL vagy az LVPLS, azt bizonyítják, hogy a mai statisztikai eszközök jóvoltából mind az ilyen mesterséges univerzumokban való utazás, mind „égitesteik" evolúciós vizsgálata lehetséges. Mindazonáltal mielőtt utaznánk, el kell döntenünk, mik között hajózzunk? A jelenségek átalakítása morfokká, alakzatokká tudományterületenként meg kell előzze a későbbi, még érdekesebb vállalkozásokat. Az idézettségi klasztertérképek ezekhez tűnnek előmunkálatnak.[12-13]

Stanislaw Lem egyszer azt írta, a szakember olyan barbár, akinek tudatlansága nem terjed ki mindenre. Ugyanő figyelmeztetett arra is, hogy az információrobbanás egyetlen hathatós ellenszere olyan új gondolatok és eljárások ötvözete lehet, amelyek megrostálják a sokat, és csak a lényegeset tartják meg belőle.

Úgy tűnik, az információ tömkelegéből a tudást kicsapatni képes módszertan kidolgozása a könyvtári informatikára vár. Ha ez igaz, e módszertanban helye lehet a magasabb fokú alakzatokat és relációikat felderítő eljárásoknak is.

Irodalom

1. Horváth T.: „Egy törzsnek az ága..." Ismeretszervezés és szintaxis. Kandidátusi értekezés. Budapest, 1987.
2. Darányi, S. - Kovács, G. - Ábrányi, A.: The Concept and Observation of Higher Order Morphologies. 8th SCCAC Meeting, Darmstadt, 1990. Kiadatlan kézirat.
3. Vierra, R.K. - Carlson, D.L.: Factor analysis, random data, and patterned results. American Antiquity 48, 1981. 272-283.
4. Belzer, J.: Entropy. in: Kent, A. - Lancour, H. (Eds.): Encyclopedia of Library and Information Science, 8. (Dekker), New York, 1972. 126-131.
5. Sparck Jones, K.: Some thougths on classification for retrieval. Journal of Documentation 26, 1970. 89-101.
6. Thurstone, L.L.: An analytical method for simple structure. Psychometrika 19, 1954. 173-182.
7. Darányi S.: Infonautika. Mitológiakutatás és statisztikai informatika. Bölcsészdoktori értekezés. (ELTE) Budapest, 1989.
8. Darányi, S. - Kovács, G. - Ábrányi, A.: Knowledge extraction from information by multivariate statistical methods. in: Hämäläinen, P. - Koskiala, S. - Repo, A.J. (Eds.): Proceedings of the 44th FID Conference and Congress, Helsinki. Participants' edition. 1988. 49-54.
9. Abell, G.O.: The distribution of rich clusters of galaxies. Astrophysical Journal Supplement, 32, (1960). 211-288.
10. Nadel, E. - Lowe, C.V.: A comparison of two approaches to higher-order aggregation of single link bibliometric clusters. in: Hurd, J.M. - Davis, C.H. (Eds).: Proceedings of the 49th ASIS Annual Meeting, Chicago, Ill. 23, 1986. 233-236.
11. Raubertas, R.F.: Spatial and temporal analysis of disease occurrence for detection of clustering. Biometrics 44, 1988. 1121-1130.
12. Bujdosó E.: A tudományos kutatás szerkezetének felderítése: az együttidézési klasztertechnika. Könyvtári Figyelő, 32. 1986. 260-271.
13. McCain, K.W.: The paper trails of scholarship: mapping the literature of genetics. The Library Quarterly, 56, 1986. 258-271.

Jegyzetek

Gould, S.J.: One standard lifespan. New Scientist, 81, 1979. 388-389.

Ivanov, V.V. - Toporov, V.M.: The invariant and transformation in folklore texts. Dispositio, 1, 1976. 263-270.

Propp, V.: A mese morfológiája. (Gondolat), Budapest, 1975.