

DISSZERTÁCIÓ

Automatikus orvostudomány

1979

Horváth Tiber

A U T O M A T I K U S O S Z T Á L Y O Z Á S

Bölcsészdoktori disszertáció

Horváth Tibor

BUDAPEST

1979

T A R T A L O M

	Oldal
Bevezető	3
I. OSZTÁLYOZÁSI RENDSZEREK	9
II. AUTOMATIKUS NYELVI ELEMZÉS	23
1. Szövegminták problémája	27
2. Kulcsszavak növekedésének kérdései	34
3. Szavak tőalakjának automatikus felismerése	35
4. Szókapcsolatok elemzése	43
5. Gyakorisági vizsgálat	46
III. SZÓELEMZÉSTŐL A KLASZTERÁLÁSIG	53
IV. AUTOMATIKUS OSZTÁLYOZÁS	59
1. A klasszikus logika problémája	61
2. A távolság meghatározása	67
3. Hasonlósági függvények	71
4. Klaszterek kialakítása	83
5. Klaszterek típusai	90
6. Az alkalmazás területei	92
Irodalom	

B e v e z e t ő

Disszertációm alapja a Könyvtári Figyelő 1978. 5. számában közzétett tanulmányom, amelynek témája az automatikus osztályozás volt. Az osztályozási rendszerek tipológiáját szintén egyik előző publikációból vettem át /10. és 11. bibliográfiai tétel/. Azért tartom kíváncsúnak mindezt a bevezetőben megjegyezni, mert a disszertáció szövege helyenként megegyezik az idézett tanulmányokéval.

A disszertációval több célt kívántam elérni.

1. Az informatikai osztályozásnak - a könyvtári osztályozást is ideértve - jelenleg három, egymástól eléggé különböző alapokon felépített elmélete alakult ki. A könyvtári illetve dokumentációs osztályozás lényegében a klasszikus logikára épül, másik ágon a tudományrendszertan kétezer éves diszciplinájába ereszti gyökerét. Az 50-es években a számítástechnika megjelenésével az osztályozás is új alapokat keresett, így született meg a nyelvészeti alapozású elmélet, amely az osztályozási rendszereket formalizált nyelvként kezelte, amelynek meghatározott szókinccse /lexikai egységek/, szemantikája és szintaxisa van. Az "osztályozási rendszer" műszó helyett információkereső nyelvről beszéltek, a terminológiai változtatással is érzékeltetve az új koncepció lényegét. Ez az elmélet eredményezte

egyfelől az információs teauruszokat, másrészt az olyan fejlett szintaxissal rendelkező nyelveket, mint a franciák SYNTOL-ja. Harmadikként született, szintén az 50-es és 60-as években, az osztályozás statisztikai elmélete. Mindhárom elmélet önmagában is igen szerteágazó utakon járt, nagyon különböző osztályozási nomenklatúrákat eredményezett. Gondoljunk arra, hogy Taube UNITERM koncepciója és az információs teauruszok - bár elméleti alapjaik igen közeliek - nem hasonlítanak jobban egymásra, mint mondjuk egy tárgyszórendszer az ETO-hoz. Az elméleteknek ebben a bábeli tornyában égető szükséggé vált egységesebb alapvetést alkotni. H.Borko, a kaliforniai egyetem könyvtáros professzora /UCLA/ kiáltványnak is beillő cikkben követelte az egységes elméletet. Kézenfekvő ugyanakkor, hogy az egységesebb elmélet kidolgozása nem lehet végbe valamilyen összegezéssel, hanem úgy, hogy a problémák mélyebb elemzésre, átgondolásra kerülnek.

Disszertációmban megkíséreltem egy kifejezetten statisztikai elméletnek - a cluster analízisnek, klaszterálásnak - értelmezését megadni a klasszikus logika alapján, rámutatva arra, hogy a két elmélet nemcsak illeszthető, hanem egyik leírható a másik terminusaival.

L.von Bertalanffy, a rendszerelmélet "atyja" szavaival élve, talán igazolni lehetett a két elmélet izomorfiaját,

egyben azt, hogy mélyebb kapcsolataik vannak a felszíni különbözőségek ellenére. Amennyiben ez sikerült, akkor a disszertáció egyben újszerű vonással is gazdagította a kérdés szakirodalmát. Másfelől, hasonló indítékok alapján a nyelvt statisztikai módszereknek jogosságát igyekeztem bemutatni néhány osztályozási probléma megoldásában. Ezek azonban ma már szakmai közhelyek, a disszertáció célja itt a magyar nyelvészeti kutatások igen nagyvonalú összegzésére törekedett s inkább a figyelmet kívánja felhívni a nyelvészek és informatikusok közös feladataira, mert egyik táborban sem tudatosult eléggé bizonyos problémák fontossága a másik tudomány számára.

A disszertáció ebben a tekintetben úgy fejleszthető tovább, hogy a két, szintézisre hozható elmélet mellé a harmadiknak is kijelöli helyét egy egészségesebb rendszerben.

2. Második cél volt bemutatni egy valódi interdiszciplináris témát. Az informatika valóban új tudomány, s mint ilyen, számos hagyományos diszciplína határán alakult ki. Önállósága azonban ma már aligha vitatható. A disszertáció témája is a logika, nyelvtudomány, matematika, statisztika, számítógéptudomány határán mozog, maga a probléma azonban egyértelműen informatikai-könyvtártudományi kérdés, az említett tudományok egyikének sem tartozik

vizsgálódási körébe. Az informatikai problémák megközelítésének aligha tartható módszere az, hogy ezeket a kérdéseket csak valamelyik, már polgárjogot nyert tudomány - a problémákhoz képest mindig egyoldalú - módszereivel és elméleteivel igyekeznek megoldani, egyben igazolni ezzel valamiféle tudományosságot is. Természetes azonban, hogy az informatika önálló voltának hangsúlyozása együtt jár azzal, hogy azokat a mély és valódi kapcsolatokat feltárjam, amelyek az informatikát a többi tudományhoz kötik. Ebben a tekintetben a disszertáció egy érv, egy adalék, egy példa kívánt lenni.

3. A harmadik célkitűzés szorosan kapcsolódik az oktatáshoz. Viták vannak arról, hogy az elektronikus számítógépek milyen szerepet töltenek be a társadalomtudományokban, milyen mélyen kívánatos számítástechnikát tanítani a könyvtárosképzésben vagy akár más szakokon. A szembenálló felek érvei szinte párhuzamosan zúgnak el egymás mellett. Abban egyetértés uralkodik, hogy a gép csak technika, a problémák megfogalmazása adja az igazi feladatokat. A számítógépes szakemberek hangoztatják, hogy a valódi feladat a gépi munkálatok előkészítése. De miben áll ez az előkészítés? Hogyan kell a problémákat megfogalmazni? Kell-e új ismeret a nem számítógépes szakember számára ahhoz, hogy számítógéppel dolgozhassék?

Valóban, a gép "csak cifra szolga". Jelentősége azonban abban áll, hogy segítségével olyan feladatok is megoldhatók, amelyek nem voltak lehetségesek ennek a technikának megjelenése előtt. A számítógép lényegesen kiterjeszti bármely tudományban a megoldható problémák körét, mint a távcső - amely szintén csak technika - a csillagászatban. A disszertációban végig számítógépekről van szó, anélkül azonban, hogy magát ezt a technikát szükségképpen idézni kellett volna, mert néhány problémának a megfogalmazását nyújtja a számítógép számára. Azt kívánja megmutatni, hogy egy nagyon régi kérdés, az osztályozás milyen lehetőségekkel bővült, milyen új módszerek váltak alkalmazhatóvá, egyszóval, miben áll egy évezredes szakmának, a könyvtárosságnak kérdéseit újra fogalmazni, miben áll megújítása. Ebben a kérdésben a disszertáció elkötelezte magát. Aki pedig idegenkedik a "technika" előretörésétől, annak nem szükséges ezt a technikát észre vennie: a problémák új megfogalmazása e technika nélkül is elég intellektuális feladatot, élményt nyújt. Ha a szakmai gondolkodás ennek következtében merészebbé, egyben pontosabbá válik, ha új távlatok nyílnak a gondolkodás számára, akkor ez önmagában is eredmény, hagyhatjuk a kifejezetten technikai jellegű kérdéseket másra.

4. Szintén az oktatással függ össze az is, hogy világosan kell látni, a könyvtáros- és információs képzésben milyen alaptudományi studiumok kívánatosak, s miért.

A disszertációban számok hivatkoznak az idézett szakirodalmi tételekre, amelyek a dolgozat végén kerültek felsorolásra. Ha az idézet nem a szűkebb téma szakirodalmából való, akkor a hivatkozás forrása szöveg közben található.

Budapest, 1979. október 31.

I. OSZTÁLYOZÁSI RENDSZEREK

Automatikusnak tekintünk egy osztályozást, ha

- a/ az osztályozási kifejezések a közlemények eredeti szövegéből /természetes nyelv/ kerülnek meghatározásra;
- b/ ha az így nyert kifejezésekkel az összes további művelet /csoportosítás, rendezés, kapcsolatok meghatározása stb./ automatikusan megy végbe.

Az automatikus osztályozás teljes problémakörét tehát két, egymástól meglehetősen eltérő kérdéscsoport alkotja. A statisztika, taxológia, orvostudomány, továbbá számos más olyan tudomány, amely az automatikus osztályozás módszereit alkalmazza, csak a b/ alatti problémát veti fel. Érthető módon, hiszen a csoportosításra szánt adatok, objektumok, élő szervezetek, betegségek, stb., már rendelkeznek megadott ismérveikkel, tünetcsoportjaikkal, amely tulajdonságokat előzetesen meghatározták kísérleti vagy más módszerekkel. A feladat csupán ezeknek az objektumoknak, adatoknak természetes osztályait vagy csoportjait meghatározni. Az informatikában azonban a vizsgálat tárgya természetes nyelven megírt szöveg, itt nem áll rendelkezésre a dokumentumokra vonatkozó ismérvhalmaz, amellyel a közlemények tartalmi, módszertani, stb. tulajdonságai leírhatók. Ha pedig rendelkezésre áll egy osztályozási rendszer, no-

menklatura, információ-kereső nyelv, rubrikátor, akkor nem lehet szó automatikus osztályozásról, hanem valamilyen másról.

Meg kell vizsgálni tehát azt a kérdést, hogy az automatikus osztályozás milyen helyet foglal el más elveken nyugvó osztályozások között. Erre azért is szükség van, hogy a terminológiai elhatárolás élő tartalommal teljék meg, továbbá, hogy alkalmas rendszerezés alapján össze lehessen hasonlítani legalább a fő osztályozási típusokat, ha egybevetésükre sor kerül.

Jelen tanulmánynak nem témája az osztályozási rendszerek tipológiáját megadni kimerítő részletességgel, ezért csak egyetlen felosztást idézünk itt a lehetségesek közül. A tárgyalt felosztás a /11/ sz. alatti publikációból való és jellegezetessége, hogy négy, egymástól független felosztási alapon nyújt dichotomikus típuspárokat, megengedve azonban azt, hogy köztük átmeneti fajták is legyenek.

Első felosztás

Az osztályozási /információkereső/ nyelv szabályozottsága, elemzettsége szempontjából természetes nyelvű és szabályozott típusok vannak. Helytelen ezt a felosztást a nyelv "eredeté"-re visszavezetni és a természetes nyelvűt szembe-

állítani a "mesterséges" vagy formalizált rendszerekkel. Ugyanis minden osztályozási nyelv eredete a természetes nyelv. Régi rendszerek még előálltak a tudományok felosztásának eredményeként, ma ez az út csak kivételesen járható, és csak különleges célú osztályozások esetében, mint amilyen a BSO vagy a legfelső szintű rubrikátor /az utóbbi az előbbinek megfelelője a KGST együttműködésben/. Ezeknél a cél nem is maga a dokumentumok osztályozása, hanem a tudományok és ismeretek felosztása: az információk átfogó csoportjait kívánják kialakítani.

Ez a felosztás nem azonos Babiczky Béla két típusával, amely szerint elválnak egymástól a természetes és mesterséges osztályozás. Természetes osztályozás nála az objektumok lényeges ismérvei, belső lényege szerinti osztályok kialakítását jelenti, amely szemben áll az objektumoknak valamely gyakorlati célú osztályokba sorolásával. Az utóbbit tekinti mesterségeseknek. /2/. Ügyeljünk tehát a terminológiára: a természetes osztályozás nem azonos a természetes nyelvüvel. Mindkét felosztásnak megvan a jogosultsága, de más-más felosztási alapon nyugszanak, s így függetlenek egymástól.

A természetes nyelvű osztályozás azt jelenti, hogy az osztályozási-indexelési kifejezések az eredeti szövegből kerülnek meghatározásra, kivonásra, és az így nyert kife-

jezések előfordulásuk alakjában /pl. ragozva, jelekkel, toldalékokkal együtt/ kerülnek alkalmazásra. Természetes nyelvűnek tekintjük azt az osztályozást is, ha a szóalakokat egységesítjük a ragozott, toldalékolt alakok összevonásával, sőt, ha elemzést végzünk a valódi szinonimák tekintetében. Az így kapott osztályozási kifejezéseket kulcsszavaknak hívjuk.

A legelső természetes nyelvű osztályozás, H.P.Luhn KWIC indexe volt és egy bibliográfiai kiadványban, a Chemical Titlesben látott napvilágot. Azóta ez az osztályozási technika nagy utat tett meg, mai fejlettebb változatai már nemcsak a dokumentum címére korlátozzák a kulcsszavak meghatározását, hanem a teljes szövegre, vagy a teljes szövegből vett kisebb terjedelmű mintára. A kulcsszómeghatározás a természetes nyelv statisztikai és morfológiai elemzésén nyugszik /a nyelvtudomány szolgáltatja a megfelelő algoritmusokat/. Ilyen módszerrel minden dokumentumhoz nagyszámú kulcsszó határozható meg.

A természetes nyelvű osztályozásoknak filozófiája az a felismerés, hogy az eredeti szövegben előfordulnak azok a szavak, szószerkezetek, kifejezések, amelyek alkalmasak a mű témáinak, módszereinek, stb. leírására. Sőt, - miután a nyelvi közlés statisztikus törvényeknek van alávetve - e kifejezések jellemző gyakorisággal és jellemző statisztikával

तिकai eloszlásokat követve fordulnak elő. Természetes, az ilyen fajta osztályozás akkor alkalmazható, ha egyben automatikus és ha a kiválasztott kulcsszavakkal és kifejezésekkel számos más műveletet lehet végezni.

A természetes nyelvű osztályozások tehát valószínűségi törvényeknek engedelmeskednek: nagy valószínűséggel határozzák meg a mű témáját, módszerét, stb. Az ebből fakadó bizonytalanság kisebb, mint a szabályozott osztályozások esetében a "konzisztencia" /következetesség/ elkerülhetetlen megszegéséből eredő hibák okozta bizonytalanság.

A kulcsszavas osztályozásnak vannak hátrányai. Ilyen korlát a nyelvhez kötöttség. Terjengősebben lehet velük osztályozni. Gyakran pedig hiányoznak a megfelelő algoritmusok is.

A természetes nyelvű osztályozással szemben állnak a szabályozott osztályozási nyelvek. Ezek úgy keletkeznek, hogy a kulcsszavak több elemzésen esnek át és minden elemzés egyfajta szabályozást eredményez. Az osztályozó nyelv így egyre egyértelműbb, következetesebb, tömörebb lesz, alakilag, jelentésbelileg egyaránt. A kapott szabályozott, elemzett osztályozási kifejezések egymás közti kapcsolatai is kiépíthetők, s végül a relációk ugyanolyan fontos részeivé válnak az osztályozásnak, mint maguk a kifejezések.

A szóban forgó elemzések az alábbiak lehetnek.

- Alaktani és lexikográfiai elemzés, ennek során kerülnek összevonásra a szóalakok, egységesül az írásmód, stb.
- A nyelvi kétértelműség elemzése /homonimia/ és ennek kiküszöbölése.
- Elemzés szinonimák szempontjából.
- Logikai kapcsolatok /főlé-alárendelés/ elemzése.
- Szerkezeti kapcsolatok /egésze-része/ elemzése.
- Egyéb kapcsolatok elemzése /eszköz-rendeltetés, hasonlóság, dolog-tulajdonság, stb./

Ha az osztályozási nyelv ezeknek az elemzéseknek eredménye, és ha a relációk beépítésre kerülnek, akkor tezaurusznak nevezzük, lexikográfiai egységeit, tehát a benne szereplő osztályozási kifejezéseket pedig deszkriptoroknak. A tezauruszok egyben a szabályozott osztályozási rendszerek legjobb példái.

A szélsőségek között átmeneti típusok vannak. Ilyenek pl. az ismert tárgyszavas osztályozás, amely már nem természetes nyelvű, de nem is kimerítően elemzett. A valóságban vannak tehát átmenetek is, Sőt, a felhasználásban sincs merev elkülönülés a kettő között. Mert megfelelő konkordanciákkal természetes nyelvű osztályozásról akár automatikusan át lehet térni egy szabályozottra.

Második felosztás

Az osztályozási rendszereket más alapon automatikus és nem automatikus /emberi?/ csoportra lehet felosztani; köztük helyezkednek el a félautomatikus rendszerek.

Az automatikus osztályozás a fentebb található definíció szerint két feltételnek tesz eleget. Ha csak az egyik feltétel teljesül, akkor félautomatikus.

Az automatikus osztályozásnak legfejlettebb változata az ún. klaszterálással érhető el. /Cluster = halom, csomó, egy rakás valamiből. A szónak terminusként elfogadott magyar megfelelője nincsen./ A klaszter analízis bármilyen objektumok halmazának automatikus csoportosítására, klaszterálására szolgáló eljárás, ha adottak az objektumokhoz rendelt ismérvek. Így a legkülönbözőbb tudományok alkalmaz-
zák, statisztika, szociológia, orvostudomány, biológia, stb. Az eljárás megértéséhez figyelembe kell venni, hogy a tájékoztatásban minden dokumentumhoz az ismérvek /kulcs-
szavak, deskriptorok, stb./ egy halmaza, sorozata tartozik. Ezek az információs rendszereket modelláló dokumentum-
ismérv mátrix egy-egy dokumentumvektoraként foghatók fel. E vektorok között "hasonlóság" számítható. Ha a hasonlóságot kifejező érték - amely egyébként 0 és 1 közé eső szám - egy bizonyos küszöb fölött van, akkor ez szorosabb összetartozást, nagyobb hasonlóságot fejez ki, más szóval az

ismérvek közt több azonosat jelez. Azok a dokumentumok alkotnak egy klasztert, amelyek közt számított hasonlósági együttható e küszöbérték fölött van, amelyek elég magasan hasonlítanak egymáshoz. Az elmondottak persze nagyon leegyszerűsítik az eljárást; a klaszterálásnak fejlett elmélete, differenciált eljárásai vannak, majdnem önálló diszciplinává vált. Azt kell mindenekelőtt kiemelni, hogy az aristotelesi logika "osztálybasorolási" eljárásával szemben határozottan előnyben van azzal, hogy egyidejűleg veszi figyelembe a csoportba sorolásnál az összes ismérvet, s nemcsak egy ismerv alapján sorol klaszterbe. Erre szokták mondani, hogy ez az eljárás "objektív".

A klaszterálás szinte kivétel nélkül a kulcsszavak, ismérvek automatikus meghatározásával párosul /tehát természetes nyelvű automatikus eljárással/, így a klaszterálás teljesen automatikus osztályozás. Az egyre finomodó klaszterálási eljárások, számítások a legmagasabb színvonalú, legmélyebb emberi osztályozásokkal vethetők egybe, sőt, lassan bármelyik más osztályozásnál jobb hatásfokúak. Ehhez járul az a határozott előny, hogy automatikusak.

Hátrányaként jelentkezik magas technikai követelménye, gépigénye, nagy kapacitású tárolókkal.

A klaszterálásról még több is igaz. Az irodalomkutatás is a klaszterálás szabályai szerint mehet végbe: a keresőkép /profil/ is egy ismérvsorozat, azaz vektor. Ezt a vektort kell összehasonlítani a klaszterek un. centroidjával. Ez a centroid a klaszterbe tartozó dokumentumok vektorainak olyan átlaga, amely a klaszterben a közöst mutatja meg. Az összehasonlítási eljárásban tehát először a megfelelő klasztert kell kiválasztani, ezután kerülhet sor a klaszterbe sorolt dokumentumok vektoraival való egybevetésre.

Jelen disszertáció az itt felvetett kérdéseket kívánja részletezni a továbbiakban, s az itt adott szempontok nemcsak ennek az osztályozási fajtának a típusba-sorolásához szükséges jellegzetességeit adták meg, hanem egyben a később kifejtendő kérdéseknek rövid összegzését is adták.

Harmadik felosztás

Éggyakrabban azzal a felosztással lehet találkozni, amely szerint a létező osztályozási rendszerek mellérendelő és hierarchikus /föle-alárendelő/ fő típusokra oszthatók, azon az alapon, hogy az osztályozási kifejezések közt a rendszer kialakítása során függetlenséget vagy logikai függőséget határoztak meg. Ezt a szempontot azonban pontosítani kell. A két tipushoz - miután ismert fajtákról van szó - néhány megjegyzést kívánatos tenni.

Az első az, hogy ebben a felosztásban sem helyezhető el minden, mert pl. az információs teauruszok - fő részüket tekintve - mellérendelőek, de tartoznak hozzá egyéb részek, így hierarchikus rész is. Ez a felosztás is egyszerűsít.

A másik megjegyzés az, hogy a hierarchikus kapcsolatok logikai fölé-alárendelés, illetve egésze-része szerkezeti kapcsolatok alapján jöhetnek létre. Lényegük az, hogy e kapcsolatokat az osztályozási rendszerben teljes mélységben építik fel, a legátfogóbb, legáltalánosabb fogalmaktól indulva a legspecifikusabbakig; vagy ellenkező irányban a legszűkebbtől indulva a legáltalánosabbakig. Ha bármely, a hierarchiába tartozó fogalmat tekintünk alapnak, akkor ez lehet elágazási pont - mert indulnak belőle fajfogalmak, illetve "része" fogalmak,- ellenkező irányba pedig csomópont, mert összefoglalója, genusa, egésze valamiknek; és fölötte még összefoglalóbb, még általánosabb csomópontok vannak. Minden fogalom sokféle szempontból bontható fel, tagolható tovább, másrészt sokféle szempontból lehet "fölöttes" fogalma. E sokféle szempont közül a hierarchikus rendszerben mindig egyet lehet kiválasztani. Ezért minden egyes fogalom /amely tehát egyrészt elágazás, másrészt csomópont/ egyben döntési pont is a fogalmi hierarchiában. Ha nagy a hierarchia, túl sok döntést kell hozni, s minden döntés valaminek a kiemelését, más szem-

pontok elhagyását jelenti. Gyakran egyenlő esélyű lehetséges további bontások közt kell választani. Nem kell különösebben kommentálni, hogy egy hierarchikus rendszer a végén hogyan viszonylik ahhoz a valósághoz, amit tükröznie kell.

A hierarchia valahogyan a "rend" benyomását kelti, s néha lenyűgöz bennünket a konstrukció logikája, elegáns megoldásai. A hierarchikus osztályozás mögött kimondatlanul is ott lappang egy előfeltevés, hogy ez a hierarchia követi a valóság hasonló logikán nyugvó rendjét. Feltételez egy a priori rendet. Hogy a valóság így létezik-e, annak eldöntése a filozófiára tartozik. Az osztályozás kérdése az, hogy ha létezne is ilyen valóság, követhető lenne-e egyetlen, akár a legjobban megépített osztályozási rendszerrel!

Ezzel szemben a mellérendelő osztályozás egyáltalán nem azt jelenti, hogy logikai kapcsolatok a rendszerbe nem építhetők. Sőt, többszörösen is beépíthetők, más-más felosztási alapon kapott speciesek és genusok, és ezt éppen a mellérendelés teszi lehetővé. Mivel minden fogalom egyenrangú, sokoldalú kapcsolataik feltüntethetők, nem kell egy relációtípust kiválasztani a hierarchia kedvéért. Ennek az ára, hogy nincs bennük teljes mélységű kapcsolatrendszer, mindig csak egy-két fokozatú az adott fogalomhoz viszonyítva. Ezzel egyúttal pontossá vált a "függőség" tartalma is, amely fentebb definícióként szerepelt.

Végül meg kell mondani, hogy ha konkrét hierarchikus osztályozásokat vizsgálunk, akkor a hierarchia kedvéért olyan kapcsolatok is szerepelnek benne, amelyeknek nincs közük a fölé-alárendeléshez, sem a szerkezeti relációkhoz. Mert teljes hierarchiát nem lehet szigorúan logikai és szerkezeti kapcsolatokból felépíteni.

A ~~k~~ritikai észrevételek nem a hierarchikus osztályozási rendszerek ellen szólnak. Hierarchikus rendszerekre is szükség van. A fogalmak közti relációknak az osztályozási rendszerekben való kiépítését a tudományokban gyakran fellelhető dilemma jellemzi, amely akkor lép fel, ha két fontos követelményt csak úgy lehet teljesíteni, hogy egyik a másik rovására érvényesül. /A fizikában klasszikus esete ennek Heisenberg határozatlansági relációja, vagy az informatikában a teljesség-pontosság egymásnak ellentmondó érték kategória./ Esetünkben ez a dilemma úgy szól, hogy vagy teljes mélységben érvényesül valamely relációtípus, konkrétan az alá-fölérendelési kapcsolat, de akkor más kapcsolat-fajtákat mellőzni kell, vagy ellenkezőleg, minden fajta reláció kiépítésre kerül, de akkor le kell mondani ezek teljes fokozatú kiépítéséről és csak a legközelebbi kapcsolatok jelenhetnek meg. Az ellentmondást az információs teauruszok úgy oldották fel, hogy a teaurusz lexikai egységeit, kifejezéseit több egyenértékű részben, más-más szerkezetben megismétlik.

Negyedik felosztás

Abból a szempontból, hogy az osztályozás az egyedi információk mélységében, vagy ellenkezőleg, átfogóan megy végbe, ismét két fő típust lehet meghatározni, az individualizáló és generalizáló osztályozásokat. Az előbbi az egyedi információk leírására alkalmas, az utóbbi csoportok, osztályok, átfogóbb témák, nagytömbű információk meghatározására. Vegyünk egy példát, Lyka Károlynak A művészetek története c. könyvét. Hogyan lehet ezt osztályozni? Ha jól meggondoljuk, a legtöbb helyen így: képzőművészet, történet. Aki részletesebben kívánja ezt a feladatot elvégezni, a fenti két fogalom mellé még az alábbiakat is odaitéli: építészet, festészet, szobrászat. Aki még ennél is részletesebben, az megnevezhet 10-12 stílusjegyet illetve korszakot, mint reneszánsz művészet, barokk, stb. Egyre növekszik a feltártság, egyre több helyről lehet a könyvet visszakeresni, de ezzel párhuzamosan egyre szűkebbek az osztályozó fogalmak, egyre inkább kisebb részleteket neveznek meg. Ezen az úton elérhetünk olyan finom részletekig, mint pl. a gótika első, XIII. század elejéről származó magyarországi építészeti emlékei. Ez is benne van a könyvben. Most már az egyedi információk mélységében vagyunk, és ebben a mélységben másként kell osztályozni. Teoretikusabban fogalmazva az információkereső nyelvek lényeges tulajdonsága az, hogy kifejezései, fogalmai, megnevezései hogyan darabolják fel a valóságot apró, vagy nagyobb részletekre.

Amilyen átlagos méretű ez a feldarabolás, olyan darabok szerint lehet visszakeresni is. Az osztályozásnak a kérdések gyakorlatához kell igazodniok, ezek pedig általában az egyedi információk mélységében merülnek fel. Jelen pillanatban a generalizáló osztályozási rendszerek esetében kívánnak meg nemzetközi egyöntetűséget, hiszen céljukat az információk kölcsönös cseréjében kell keresni. Ezért mind a BSO, mind a Legfelsőbb Szintű Rubrikátor nemzetközi normatív előírásként jelent meg. Az ETO-hoz hasonló osztályozások megkísérlik a kettőt egyeztetni éppen a hierarchia segítségével. De a hierarchia természete olyan, hogy inkább a generalizáló osztályozás felé tolja el az indexelő nyelveket.

Itt is két véglet került meghatározásra. A kettő közt van a legtöbb gyakorlati osztályozási rendszer, valamilyen átmenetet képezve. Ilyenkor az a jellemzőjük, hogy inkább az egyikhez, vagy inkább a másikhoz állnak-e közelebb.

II. AUTOMATIKUS NYELVI ELEMZÉS

Az automatikus osztályozás első problémaköre tehát az, hogy megfelelő módszereket kívánatos keresni az osztályozási kifejezések meghatározására, ha e kifejezések forrása a közlemények eredeti, természetes nyelvű szövege. Továbbá, mivel automatikus eljárásokról van szó, e módszereket megfelelő algoritmusok formájában kell leírni, lehetővé téve ezzel magasszintű adatfeldolgozási technika alkalmazását.

Az automatikus kulcsszó meghatározás az informatikában valóságos diszciplinává nőtte ki magát attól kezdve, hogy H.P.Luhn ma már klasszikusnak számító publikációiban /1958/ megvetette e diszciplína alapjait. /15/ Általános megfontolásoktól eltekintve az egész kérdéskör azonban - a dolog természeténél fogva - minden egyes nyelvben külön-külön problémákat vet fel. A magyar nyelv ilyen szempontú vizsgálata néhány részlet kivételével nem indult meg. Ezen a téren tehát bőségesen adódnak kutatási feladatok.

^a
Luhn óta természetes nyelvű szövegek elemzése kulcsszavak megállapításának céljaira statisztikai nyelvelemzésen nyugszik. Jóllehet e módszerek számos tekintetben finomodtak, alapjai változatlanok maradtak.

G.Salton professzor szerint a statisztikai nyelvelemzés két okból került a nyelvelemzés középpontjába azt követően, hogy a számítógépek betörtek a nyelvészeti kutatások területére is. /20/

- Mindenekelőtt azért, mert igaznak bizonyult az a feltevés, hogy e mennyiségi elemzés eredményeként nyert kifejezésekkel le lehet írni a közlemények tartalmát.
- Másodszor azért, mert ezzel az eljárással lehet meghatározni az "entitások csoportjai"-t /igy Salton/, pl. szavakét vagy szótövekét, amelyek egy tezaurusz, vagy bármely más - nem feltétlenül automatikus - osztályozási nyelv lexikai egységeit alkotják.

A Saltoni megfontolások lényegesen kiterjesztik e módszerek alkalmazhatósági körét is, ti. nemcsak az automatikus osztályozás, hanem bármely más indexelő nyelv, információkereső nyelv számára a kifejezések gyűjtésének egyik eljárását határozza meg. Valójában bármely modern információkereső nyelv deklarálja azt az elvet, hogy kifejezései, lexikai egységei a természetes nyelvből erednek, azaz az élő, mai tudomány nyelvéből.

Salton céljait tekintve a statisztikai nyelvelemzésnek két fajtáját különbözteti meg.

Az első a szótípusok elemzése /word type/, amelynek célja szövegek /pl. két dokumentum/ megkülönböztetése, és maga az elemzés a szónak a szóban forgó szövegben való használatán nyugszik, és nem vizsgálja a szónak /vagy nyelvtani szerkezetnek/ egyéb használatát. Pl. az egyik szöveg következetesen komputer-t, a másik computer-t használ. Ilyen célú vizsgálatokkal főleg a nyelvtudomány él, pl. szövegek szerzőségének megállapítására. Magyarországon ezt a módszert a kriminalisztikai nyelvészet alkalmazta.

A másik célú elemzés a szókép elemzése /word token/, az egyedi szó előfordulásán nyugszik. Alapja a szavak gyakorisági vizsgálata, és ez volt Luhn módszere is. Az informatika számára ez bizonyult fontosnak. Jelen dolgozat is erre korlátozódik.

Már ezen a ponton érdemes azonban e módszer legnagyobb korlátját is megemlíteni, nehogy az a téves hit alakuljon ki, hogy e módszer önmagában elegendő, hogy a gyakoriság-vizsgálat önmagában megoldotta a kulcsszó-meghatározás problémáját. Hiszen e módszer atyja, Luhn is tovább lépett finomabb módszerek felé. Ez a korlát pedig - itt nem részletezendő - szemantikai nehézségekből ered, nevezetesen abból, hogy a gyakoriság-vizsgálat nincs tekintettel a kontextusra, a szöveg-környezetre, márpedig a kontextus befolyásolja a szó jelentését. Vegyünk egy példát. A "kérdés"

szó előfordul az alábbi három kontextusban:

Az iparfejlesztés kérdései...

A kérdéslogika vázlata.../Reichenbach logikája/

A kérdés az orosz nyelvben.

A kérdés szó jelentése meglehetősen különböző. Ennél nehezebb problémákat, mint a homonimia, a szinonima kérdés, stb., még nem is említettünk.

A gyakoriság-vizsgálat finomítása két további módszer alkalmazásával mehet végbe, amelyek szintén automatikus eljárások. Az egyik a szavak állandósultabb kapcsolatainak elemzése /szintagma elemzés/, a másik a szavak és dokumentumok asszociációjának vizsgálata, amelyekkel azok a szavak nyerhetők, amelyek a dokumentumokat karakterizálják. Az utóbbi kerül alkalmazásra Salton elképzelései nyomán a SMART automatikus visszakereső rendszerben.

Maga a statisztikai szövegelemzés - s ezen a továbbiakban szűkebben automatikus kulcsszó-meghatározást értünk - az alábbi részletkérdéseket veti fel.

1. A szövegminták problémáját.
2. A kulcsszavak növekedési törvényszerűségeit; a teljes szótári készlet és a kulcsszavak közti kapcsolatokat; a kulcsszavak eloszlásának problémáit.
3. A szavak főalakjának automatikus felismerését.
4. A 2. és 3. probléma megismétlődik szintagmák kapcsán.
5. Gyakoriság-vizsgálat és a gyakorisági intervallumok kérdését.

1. A szövegminták problémája

Ideális esetben a közlemények teljes eredeti szövege kerül elemzésre. Ezt a gyakorlatot sehol sem követik, nemcsak azért, mert lényegesen kisebb terjedelmű, az eredetit reprezentáló származék-szövegek ugyanolyan eredményt produkálnak, hanem mert technikailag sem oldható meg. Több tízezer, vagy százezer publikáció teljes szövege puffertárak alkalmazásával, vagy szakaszokban feldolgozva is meghaladja a legnagyobb számítógépek tárolási lehetőségeit. Az elemzés ezért korlátozódik.

- a közlemények címére,
- referátumra vagy kivonatra,
- az eredeti szövegből származó mintákra.

A dokumentumok cimén nyugvó elemzés - a permutált indexek testesítik meg - a hagyományos osztályozási eljárások hatásfokánál nem eredményez mélyebb osztályozást, s így ezeknek csupán indexművek készítésénél van jelentősége. Nem tipikusan az automatikus osztályozás tárgykörébe tartozó kérdés. Szabályozott osztályozási rendszerek szógyűjtésére inkább alkalmas, de automatikus osztályozás alapszövegeként jelentősége korlátozott.

A referátumok elemzése általánosan elfogadott gyakorlat. A szöveg tömör, a kulcsszavak halmozottan találhatók bennük. A referátumok azonban emberi tevékenység termékei, magukban hordozzák a referátumok készítőinek szubjektív

ítéletét arra vonatkozóan, hogy mit tartanak érdemesnek kiemelni az eredeti közleményből, ezért a kívánt mértékig nem tárgyilagosságok. Referátumból nem mindig derül ki az eredeti közlemény újdonságértéke, módszertani megoldása, megbízhatósági kritériumai, stb. Éppen azokat a szempontokat csempészi vissza tehát, amelyek elkerülése az automatikus osztályozás révén érhető el. Más a helyzet azonban az automatikus kivonatokkal. Csakhogy éppen úgy szövegstatistikai elemzésen nyugszanak, mint az osztályozás, így előbb megy végbe a szövegstatistikai elemzés, mint ahogyan a kivonat megszületik. Ennek folytán nem alkalmazható ez a referátum típus sem a kívánt célokra, mert az elemzésnek egy későbbi származéka, és nem kiindulása.

Legcélravezetőbb és legtárgyilagosabb a megfelelő szövegminták előállítása. A minta nagyságára vonatkozóan a kulcsszavak növekedési görbéjének vizsgálata nyújt támpontot, előzetesen erről semmi sem mondható.

Az elemzett szövegekkel kapcsolatosan néhány minőségi kritériumot támasztunk. Alkalmas szövegben a szógyakoriság, kulcsszó-gyakoriság eloszlása törvényszerűségeket mutat. Ismeretterjesztő szöveg, esszé stílusú tanulmány, s egyáltalán: nem egészen szakszerű tárgyalásmód nem eredményezheti a várt eredményt. Cherry írja: "Ha nem találjuk a szógyakoriság pontos eloszlását, akkor ...a pontos

gyakoriságtól való eltérésre sem bizhatjuk magunkat, amelyet a szignifikancia mérése szolgáltatott". /5., 107.1./ Ennél tovább megy Meadow, ismert könyvében /16., 100.1./, aki egyenesen definíció értékű kritériumnak tartja, s a kulcsszavak meghatározására tartja fent. "Szavak tárgyszóként akkor szignifikánsak, ha arányban állnak aktuális és várható gyakoriságaik" - írja.

Kulcsszavak meghatározására magyar nyelvű szöveg esetében - kevés kivétellel - nem sokszor került sor. A szerző - a Könyvtártudományi és Módszertani Központ megbízásából - más típusú szövegelemzés melléktermékeként kizárólag címekre vonatkozó adatokat nyert. Az alábbi táblázatok ennek eredményeit tartalmazzák /az adatok még nem kerültek publikálásra/. Tudományonkénti bontás mutatja, hogy egy-egy átlagos cím hány szignifikáns szót eredményezett, s ebből mennyi a magas informatív értékű szó. A tudományok bontása követi a forrásként felhasznált Magyar Folyóiratok Repertóriumának szakcsoportjait. A táblázat a nyert kulcsszavak számának csökkenő sorrendjében mutatja a tudományokat.

1.sz. táblázat

Egy tételre jutó összes releváns szómátlaga
tudományonként

1.	Kémia	6,23
2.	Könnyűipar	5,87
3.	Mezőgazdaság	5,66
4.	Építőipar	5,62
5.	Műszaki tudományok	5,40
6.	Vegyipar	5,39
7.	Orvostudomány	5,23
8.	Üzemtan	5,15
9.	Statisztika	5,11
10.	Közgazdaságtan	5,03
11-12	Matematika	5,00
11-12	Közigazgatás	5,00
13.	Geológia, paleontológia	4,98
14.	Jogtudomány	4,94
15.	Népjólét, biztosításügy	4,93
16.	Lélektan	4,89
17.	Műszer- és mérés technika	4,88
18.	Biológia	4,82
19.	Fizika	4,81
20-21	Kereskedelem, közlekedés	4,80
20-21	Hadtudomány	4,80
22.	Történettudomány	4,73
23.	Pedagógia	4,58

24.	Tudományok általában. Könyv- tártudomány	4,73
25.	Politikai tudományok	4,53
26.	Társadalomtudomány, általános szociológia	4,35
27.	Művészetek általában	4,25
28.	Csillagászat	4,24
29.	Néprajz	4,12
30.	Fényképészet	4,03
31-32	Sport	3,76
31-32	Földrajz	3,76
33.	Filozófia	3,67
34.	Nyelv- és irodalomtudomány	3,66
35.	Zene	3,59
36.	Vallás	3,41
37.	Etika	3,38
38.	Képzőművészet	3,37
39.	Esztétika	3,29
40.	Biográfia	2,99
41.	Természettudomány általában	2,75

Átlag: 4,83 releváns szó/tétel

2. sz. táblázat

Az egy tételre átlagosan eső magas relevanciójú szavak
tudományonként

1.	Kémia	4,02
2.	Könnyűipar	3,83
3.	Építőipar	3,79
4.	Mezőgazdaság	3,72
5.	Vegyipar	3,51
6.	Üzemtan	3,36
7.	Műszer és mérés technika	3,35
8.	Orvostudomány	3,33
9.	Műszaki tudományok	3,32
10.	Fizika	3,24
11.	Hadtudomány	3,20
12.	Közigazgatás	3,17
13-14	Néprajz	3,16
13-14	Közgazdaságtan	3,16
15.	Geológia, patontológia	3,13
16.	Kereskedelem, közlekedés	3,12
17.	Matematika	3,11
18.	Jogtudomány	3,10
19.	Történettudomány	3,04
20.	Politikai tudományok	3,01
21.	Biológiai tudományok	2,97
22.	Tudományok ált.Könyvtártudomány	2,96
23.	Pedagógia	2,87

24.	Művészetek általában	2,84
25.	Statisztika	2,82
26.	Lélektan	2,77
27.	Társadalomtudomány, általános szociológia	2,76
28.	Fényképészet	2,75
29.	Csillagászat	2,73
30.	Biográfia	2,65
31.	Népjólét, biztosításügy	2,58
32-33	Filozófia	2,52
32-33	Zene	2,52
34.	Nyelv- és irodalomtudomány	2,52
35.	Képzőművészetek	2,50
36.	Vallás	2,46
37.	Etika	2,44
38.	Sport, játékok	2,43
39.	Természettudomány általában	2,00
40.	Esztétika	2,14
41.	Földrajz	1,89

15

Átlag: 3,11 magasan releváns szó/tétel

2. Kulcsszavak növekedésének kérdései

A statisztikai szóelemzés eljárása igen egyszerű algoritmust követ. Az elemzésre kijelölt szöveg szavait rendre egymás után feljegyezzük, s minden szó vizsgálatakor /ezt a matematikai statisztikából származó kifejezéssel "esemény"-nek hívjuk/ megállapítjuk, hogy előfordult-e. Az elemzés során nyerjük az ún. teljes szótári készletet, amelyben minden szó előfordul, a névelőket, névutókat, stb., is beleértve. Új eseménynek tekintjük azt, ha bármely szó először jelenik meg a szöveg kezdetétől számítva. Maga az elemzés annyi "kísérlet"-ből áll, ahány szó vizsgálatra kerül, akár előfordult már előzőleg, akár nem.

A szavak elemzésének sorrendje egyébként közömbös.

A kulcsszavak a teljes szótári készlet részeként jelennek meg.

A teljes szótárkészlet, s benne a kulcsszavak növekedésére az alábbi feltételek érvényesek. /A tárgyalás alapja V.V.Nesitoj tanulmánya a Kibernetika 1977. évfolyamában közzétett tanulmánya, amelyben a probléma matematikai elméletét dolgozta ki imponáló teljességgel. /18/

- Az új esemény megjelenésének valószínűsége az első kísérletben eggyel egyenlő, azaz biztos új esemény jelenik meg.

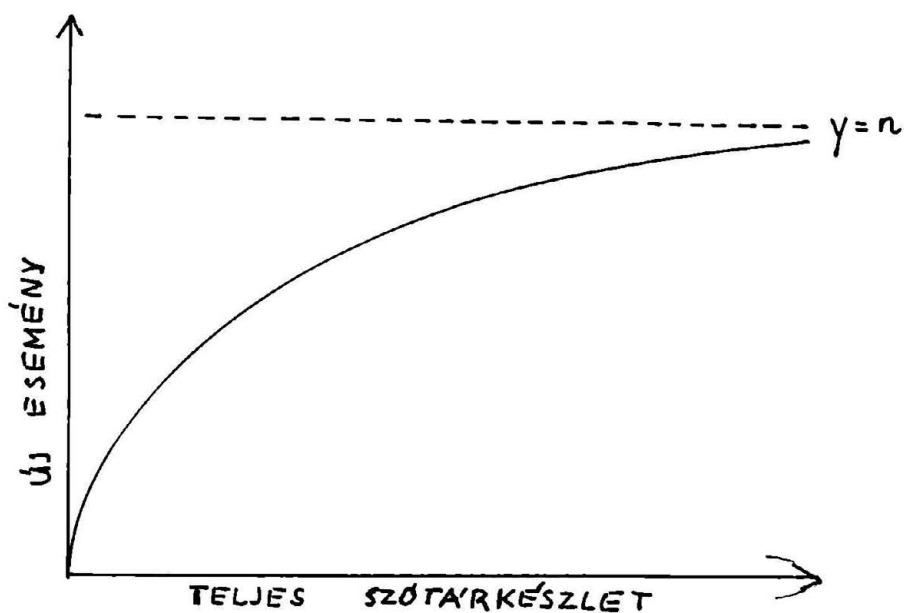
- Új esemény megjelenésének valószínűsége zérus, ha az elemzett szöveg végtelen mennyiségű, pontosabban, ha a szöveg nagysága minden határon túl növekszik.
/Matematikailag itt arról van szó, hogy az új esemény megjelenését leíró valószínűségi függvény határértéke tart a 0-hoz, ha a kísérletek száma tart a végtelenhez./
- Az új események növekedési görbéje konvex és töréspont nélküli, az $y=n$ egyenest asszimptotikusan közelíti
/1. ábra./

Megjegyzendő még, hogy a kulcsszavak növekedése a teljes szótárkészlet függvényében az adott szaknyelv függvénye is, bár a görbe szerkezete nem változik.

3. Szavak tőalakjának automatikus felismerése

Nyelvészeti szempontból kétségtelenül a legnehezebb kérdés a szavak tőalakjának felismerése. A szavak ugyanis ragozva, jelekkel ellátva, képzett formában, másfelől különböző szóösszetételek elő- vagy utótagjaként szerepelnek. Valamely szó előfordulásának megszámlálása csak úgy lehetséges, ha a szó alakváltozatait kiküszöbölve a tőalakok állnak rendelkezésre. Számítógépeink különbözőnek tekintenek minden két szót, amely akár egyetlen karakterben /jelben/ eltér.

További nehézségeket támaszt a homonimák problémája, ti. ha valamely szókép több fogalmat jelöl. /Pl. ár, kormány, fej, stb./ Az utóbbi esetben el kell dönteni, hogy az adott kontextusban melyik töröl van szó.



1. A'BRA

Az algoritmusok kialakítását nyilván azon az úton kell keresni, hogy minden toldalékokkal ellátott szót elemeire kell bontani, külön-külön meghatározni, s a kívánt tőalakot ily módon előállítani. Közelítően jó megoldásokat lehet találni erre a feladatra, minden tekintetben kifogástalan megoldásoktól azonban - a magyar nyelv tekintetében - még távol állunk.

Ez a probléma eredetileg a gépi fordításokkal kapcsolatos kutatásokban jelentkezett, s részleges eredményekhez vezetett. Kielégítő algoritmus azonban önmagában a nyelv morfológiai elemzésétől nem várható, hanem csak a szintaktikai és szemantikai nyelvelemzéssel összefüggésben álló vizsgálatoktól. Pl. a homonímia tipikusan olyan kérdés, amely kizárólag szóelemzés útján nem oldható meg.

Az orosz-magyar gépi fordítás kapcsán Melcsuk kísérel-
letei voltak az első próbálkozások a szóalakok felismerésére. Mint ismeretes, Melcsuk a magyar nyelvet közvetítő nyelvként kívánta alkalmazni, éppen "rossz" grammatikai természete miatt. Módszere az un. leghosszabb tövek felismerése volt. Abból indult ki ugyanis, hogy a magyarban igen gyakoriak az összetett szavak, ugyancsak gyakoriak a halmazott ragok. Az elemzés segéd-eszköze egy, számítógépben tárolt tőszótár és végződés-szótár.

Leegyszerűsítve az eljárást, az alábbiakban foglalhatjuk össze.

A számítógép a szó elejéről két, három,... karakterből álló csoportokat vág le, s hasonlítja a tőszótárba. Meghatározza, hogy tő-e.

A maradékot karakterenként elemzi, s hasonlítja lehetséges végződésekhöz. Ha eközben valamely karakter vagy karaktercsoport nem azonosítható, visszaállítja a teljes maradékot, s az elemzés kezdődik előlről.

Pl. az emléke szó töve emlék, a maradék -e ezt követően azonosítható. Bonyolultabb esetben jobban érzékelhető az algoritmus nehézsége. A példa legyen a nyelvemlékek szó. /Hell György nyomán/

Megtalálható leghosszabb tő: nyelv. Maradék: emlékek.
Maradék első végződése: -e. Azonosítható. Új maradék: mlékek.

A mlékek első végződése: -m. Azonosítható. Maradék: lékek.

Több végződés a maradék elején nem azonosítható, ti. végződésként. Ezért vissza kell állítani a teljes maradékot: emlékek.

A visszaállított teljes maradék - miután végzések halmazaként nem azonosítható - újra a tőszótárba kerül összehasonlításra. Új tő: emlék. Maradék: -ek. Azonosítható. Ezzel az eljárás befejeződött, a számítógép felismerte a szóösszetétel mindkét tagját és a végzések.

Maradtak azonban esetek, amikor egy összetett szó második összetevője végződés-halmazként is értelmezhető. Kónyi Sándor példája: vasakarat. Tő: vas. Lehetőséges végzések: -a, -k, -a, -ra, -t, birtokos személyjel, többesjel, határozó, illetve tárgyrag. A számítógép nem ismeri fel az akarat tövet.

Hasonló problémák adódnak a karóra /kar + óra vagy karó + ra/.

A KLTE Orosz Filológiai Tanszéke Papp Ferenc vezetésével az előbbinél árnyaltabb megoldásokat dolgozott ki, de csak a magyar főnevek szintézisére. Melcsuk sugallta eredmények mellett 12 ragozási típust határoztak meg, amelyek részletes kidolgozása során figyelembe vették a szavak végződését is, sőt, a hangrendet. Továbbra is nyitottak maradtak bizonyos ragozási típusokkal nem elemezhető esetek. Pl. a bába és bábja különválasztása, mindkettő ugyanis azonos ragozási típus, és a báb + birtokos személyjelként értelmezhető automatikus elemzésben. /Talán az analóg esetek, mint láb + a segít, hiszen "lábja" nincs a magyarban./

Az eddigi legbiztosabb eredményeket az un. véges állapotú grammatikával a Budapesti Műszaki Egyetem Nyelvi Intézetében érték el. A kutatások vezetője Hell György.

Hell az alábbiakból indul. A magyar grammatika szerint a magyar szóalakok összetevői a tövek, képzők, ragok, jelek, igekötők. Automatikus elemzés számára az összetevőket háromra csökkentette: a jeleket beosztotta a ragok közé, az igekötőket pedig a tövekhez. Három segéd-eszközt épített: a szótövek szótárát, ragok szótárát és képzőkét. /Köztük tehát a jelek és igekötők is./ A szótövekhez iktatott néhány ragot is a tőlem, neked,... szavak elemzéséhez, mivel ezek az éntőlem, teneked,... rövidített alakjai. Hangugratás, v-s, egyáltalán a változó töveket több alakban kell szerepeltetni. Pl. hó és hav-, bokor és bokr-, stb.

Az elemzésnél nemcsak a töveket és maradékokat azonosítja, hanem tekintetbe veszi ezek kapcsolódásának szabályait is. Felállított 18 lehetséges transzformációs /Hell terminológiájával átírási/ szabályt, amely meghatározza, hogy bármely magyar szó milyen állapotból milyen másik állapotba kerülhet.

A Hell-féle algoritmusok segítségével még az olyan esetek is elemezhetők, amikor egy összetett szó két eleme közé ékelődik rag, mint tetemrehívás, urambátyám, stb.

A tisztán morfológiai elemzés - mint ennek a pontnak bevezetőjeként említettük - nem oldhatja meg maradék-

talán^x azokat a problémákat, amelyek csak szintaktikai és szemantikai analízissel együtt lehet. Ezek általában akkor állnak elő, ha valamely összetevő eleve többféle elemzést tesz lehetővé. Fentebb a karóra példája mutatta ezt. Hasonló a helyzet bizonyos morfémák esetében, mint -ének, -ikre, -ekéire, -okkal, stb.

A gyakorlati alkalmazhatóságot nem zavarja az a körülmény, ha az esetek csekély számú töredéke nem elemezhető egyértelműen. E dolgozatban ugyanis a statisztikai szóelemzés célja kulcsszavak meghatározása és gyakoriságuk megállapítása. Ha a nyelvi morfológiai elemzés ennek a célnak elérésére korlátozódik, akkor az ismertetett eljárásokból eredő hibák még a tűrési határokon belül maradnak.

A szótövek kezelésének létezik a fentiektől teljesen különböző módja is. Működő visszakereső rendszerek kénytelenek a fenti igen nehéz nyelvészeti problémát megkerülni, vagy egyszerűen nem érnek, illetve nem értek rá megvárni, míg a nyelvészet meg is oldja ezeket a gyakorlati alkalmazhatóság szintjén. Számos visszakereső rendszer a tárolt tételek címeiben és referátumokban lehetővé teszi a kulcsszavak szabad keresését. Ezek az ún. szabad deszkriptoros osztályozások. Más esetben - kötött deszkriptoros megoldásban - megadják azt a

szótárat, amelynek kifejezéseit a szövegekben keresni lehet. A referátum és cím szintén toldalékolt formában tartalmazza a szavakat, s így ezek megtalálása nehézségekbe ütköznék. A keresőprofil tehát vagy felsorolja a lehetséges összes alakváltozatot - ami lehetetlen, mert a behasonlítási műveletek számát akár megszásszorozhatja, vagy az ún. maszkolás módszeréhez folyamodik. A maszkolás a szavak olyan leárnyékolását jelenti, amely csak a tövet hagyja meg, sőt, gyakran a csonkított tövet. A tő megkeresése hasonlóan indul ahhoz, ahogyan a leghosszabb tövek módszerénél láthattuk: a szavak elejéről különböző hosszú karakter-sorozatok kerülnek a megadott tő vagy csonka tövel összehasonlításra.

Egy példa illusztrálhatja az eljárást. Ha a keresőprofilban megadott tő a "ház-", akkor a keresés az összes referátumot /címet/ kijelzi, amelyben a ház, illetve ház- előfordul. Pl. a ház, háza, hazafi, házmaster, ház-nagy, házkezelőség, háztartásbeli, házi,... szavakat is.

A profilszerkesztőnek módjában áll a tövet másként meghatározni, pl. házi, vagy akár haza alakban.

Szabályozott nomenklatúrához szokott gondolkodásunk azonnal tiltakozik az ilyen zajos eredmény láttán.

A hikák azonban korántsem ilyen mértékűek, mivel a keresőprofilok sok tőből állnak, s ezek közt csak kevés a logikus kapcsolat. Ha pl. a profil második töve a tervezés /terv- alakban/, akkor ez már nem kapcsolódik logikusan a hazafi, házas, házmaster, stb., alakokhoz, eleve nincs olyan **tétel, amelyekben ilyen** kapcsolatok véletlenül adódnának. A profil további tövei csak a valódi kapcsolatok kiszűrését eredményezi. Bizonyos hiba azonban mindig marad, ha a megadott több tő többféle logikus kapcsolatot tesz lehetővé.

A maszkolás révén keletkező hibák elvi oka az, hogy a természetes nyelvben - amely információelméleti vagy szemiotikai értelemben is jelrendszer, kód - a tövek nem prefix tulajdonságúak. Azaz, valamely tő származtatható úgy is, hogy meglévő tőhöz újabb karakterek hozzáadásával más tő keletkezik. A "fü" tő kiegészül egy "l" karakterrel, "fül" keletkezik, s ennek már nincs semmi köze a fü-höz. Hasonlóan áll elő a függ, független, függöny, fürdő, fürész,... stb.

Mindenesetre a maszkolás a profilkészítők különleges ügyességét tételezik fel, nagy gyakorlatot kíván.

4. Szókapcsolatok elemzése

A téma eddigi tárgyalásából az tűnt ki, hogy egytagú osztályozási kifejezéseket nyertünk, szavakat, Sőt, a

tőalakok felismerése során olyan eljárások bevezetése vált szükségessé, amely az összetett szavakat is összetevőire bontja. Osztályozási szaknyelven szólva unitermeket kaptunk. Az uniterm egyike a lehetséges információkereső nyelveknek, de szélsőséges változataiban, mint Taube uniterm I. néven ismert szélsőségesen mellérendelő és postkoordinált variánsa az információk keresésében veszélyeket rejt magában a magas információs zaj miatt.

Az állandósult szókapcsolatok vagy a gyakori szókapcsolatok meghatározása kiegyensúlyozottabb osztályozási kifejezéseket eredményez.

A tartós szerkezetben gyakori szintagmák meghatározása relative könnyű, de munkaigényes. A szavak páronkénti kiírása során igen erőteljes gyakorisággal ugranak ki. Másfelől a szókapcsolatok döntő többsége az osztályozásban hasznavehetetlen. Míg pl. az "orvosi műszer", "fejlett szocializmus", "munkavédelmi eljárás" stb. szerkezetek magas gyakoriságuk miatt ebben a szerkezetben határozhatók meg, addig a "szocializmus építése" a "pedagógus szerepe", "kiemelkedő eredmény" stb. szintagmák - magas gyakoriságuk ellenére - aligha alkothatnak tárgyszót, bennük ugyanis nyelvi sztereotípiák jelennek meg.

A szerkezetek meghatározásának finomítottabb módszere figyelembe veszi az ún. közbeékelődő szavakat és kifejezéseket is. Pl. "a pedagógusok és mások szerepe", "irányított - így mondják - demokrácia", két összetevője közé tolakodott szavakra nincs szükség. Ha ezeket nem határozzuk meg, csökken a szószerkezet gyakorisága. A közbeékelődés kiiktatása viszont erősen növeli az elemzési munkát. Első menetben meghatározott, magas gyakoriságú szókapcsolatokra a szavak távolabbi környezetét újra át kell fésülni, hogy a második, harmadik... szomszéd nem az állandó szerkezet másik tagja-e. /Mindenesetre a gyakorlati elemzésben a távolabbi összetevőt nem honorálják oly mértékig, mint a közvetlent. Csak fél előfordulásnak számítják./

Szintagmákban gyakori eset, hogy két, egyenértékű determináns jelenik meg állandósult kapcsolatokban. "Kommunista és munkáspártok", "könyvtári és tájékoztatói rendszer" stb. a példák erre. A szerkezetek meghatározása tetszés szerint árnyalható, rendszerint azonban tetemes munkanövekedéssel jár, amely nem áll arányban az eredménnyel.

A tartós szerkezetek elemzésének más szempontból is jelentősége van. Az összetevők külön-külön lehetnek ugyanis alacsony értékűek gyakoriságuk és relevanciájuk

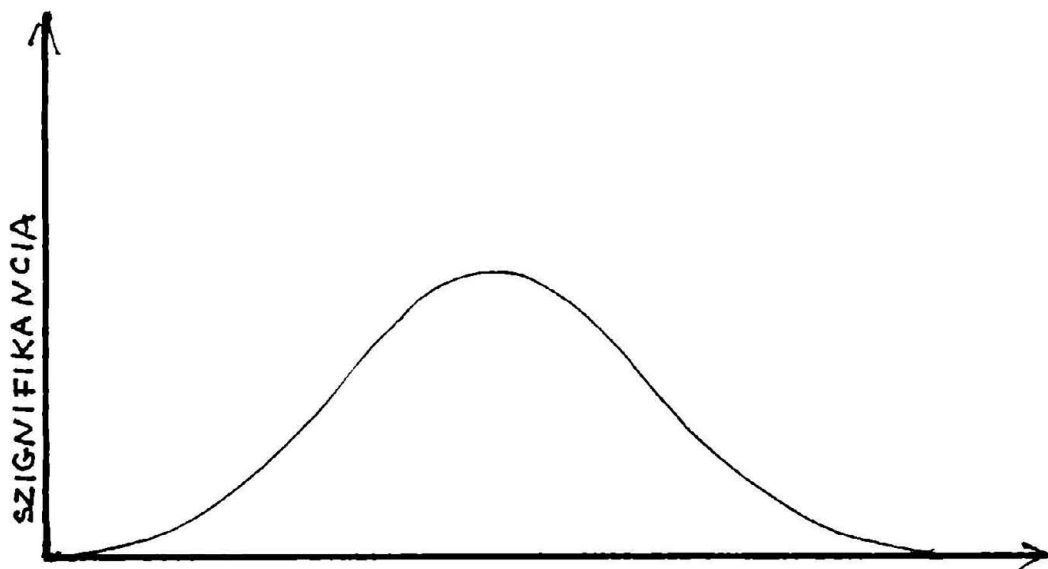
alapján. Együtt azonban markáns terminus technicust jelenthetnek. "Formális nyelv" - legyen az a példa - alkotói külön-külön nem minősíthetők magasan relevánsnak, együtt azonban pontos tárgykört jelölnek.

A szintagmák elemzésének hatásfokát javítani lehet különböző szójegyzékek - negatív szótárak, stoplisták - segítségével úgy, hogy a velük alkotott szerkezetek, szópárok nem kerülnek sem megszámlálásra, sem kiírásra. Kötőszók, névutók és bizonyos szavak, mint "problémái", "kérdései", "némely", "vonatkozásában", "szerepe", "különös tekintettel", stb. típusú szavak alkotják a negatív szó pár szótári egységeit.

5. Gyakorisági vizsgálat

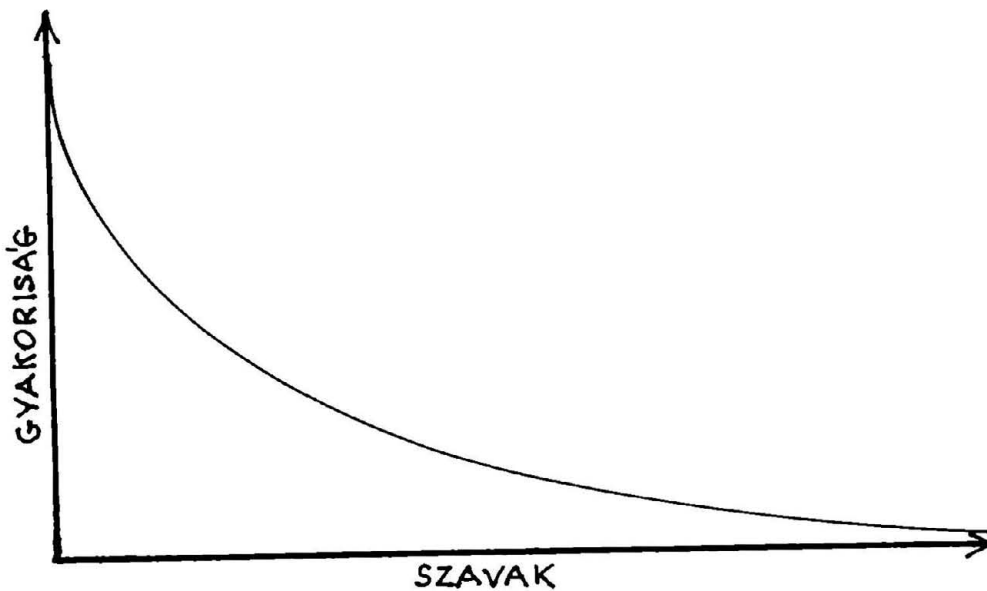
A teljes szótári készlet magában foglalja a dokumentumok osztályozására alkalmas kulcsszavakat, /releváns szavak, szignifikáns szavak, tárgyszavak/. A szavak előfordulásának gyakorisága törvényszerű kapcsolatban áll szignifikanciájukkal - ez H.P.Luhn felfedezése. A teljes szótári készlet egységeit tehát súlyozni kell gyakoriságukkal, és a szótárat e gyakoriság szerinti rendben kell megjeleníteni. Majd alkalmas módszerrel az a gyakorisági tartomány határozható meg, amely a kulcsszavak intervalluma.

De milyen összefüggés áll fenn a gyakoriság és a szavak relevanciáját kifejező érték közt? Belátható, hogy bármely nyelv leggyakoribb szavai osztályozásra alkalmatlanok. A magyar nyelv leggyakoribb szavai: az, a, és, de, meg, előtt, után,... Hasonló módon látható az is, hogy egyszer, vagy nagyon ritkán előforduló szavak is alkalmatlanok osztályozásra. Ezekben egyes szerzők egyéni szóhasználata nyilatkozik meg. Luhn felírta azt a görbét, amely a gyakoriság és szignifikancia közt fennáll, s az ismert harang-görbét kapta:



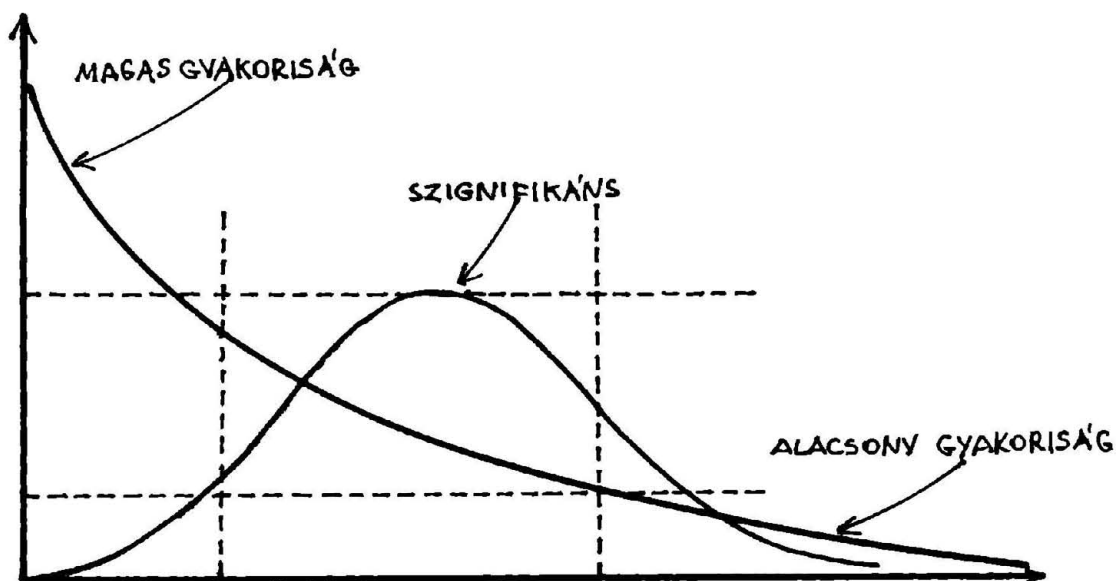
2. A'BRA

A szavak gyakoriságának eloszlása is jellegzetes görbét mutat. Az ábrán a gyakoriság csökkenő sorrendjében látható az eloszlást reprezentáló görbe:



3. A'BRA

Ha e két görbét ugyanazon koordinátarendszerben írjuk fel, látható, hogy kijelölhető egy gyakorisági tartomány, amely a szignifikáns szavakat tartalmazza. Ezt az intervallumot úgy lehet nyerni, hogy a legalacsonyabb és legmagasabb gyakoriságú szavakat kirekesztettük.



4. ÁBRA

Az ábra szerint azokat a szavakat érdemes kulcsszóként kiemelni, amelyek a szignifikanciát ábrázoló görbe csúcsa körül helyezkednek el. Meadow, később Borko számos konkrét eloszlási görbe tanulmányozása nyomán kimutatta, hogy a kivágott intervallumban is különböző típusú kulcsszavak találhatók bizonyos szabályosságot tükrözve. Magasabb gyakoriságot mutattak azok a kulcsszavak, amelyek a témákat általánosabban jellemezték, több dokumentum közös jegyei ezek, és a tárgykörök leírásánál elsősorban arra szolgálhattak, ami a tárgykörben fontos. Alacsonyabb gyakorisági tartományokban - természetesen

a megadott tartomány határain belül - azok a kulcs-szavak helyezkedtek el, amelyek a dokumentumok, tárgykörök, témák megkülönböztetésére szolgáltak. Így vált lehetségessé definiálni valamilyen "közös" ismérveket, másfelől a "megkülönböztető" jegyeket. Kézenfekvő, hogy az utóbbiak inkább specifikusabb jelentésű szavak voltak.

Ennélfogva a szignifikanciának is két meghatározást adtak. Egyrészt a szó szignifikáns lehet, mert a közös sajátosságot fejezi ki más tárgyi kategóriákkal, s ez egyben azonosnak bizonyult a fontosságot, újdonságértéket kifejező szavakkal. Másfelől a szó lehet szignifikáns azért, mert a megkülönböztető jegyeket hordozza, a tárgykörök individuális tulajdonságait. Ez utóbbiak pedig egybeestek az elméleti görbe alapján várható gyakoriságú szavakkal.

Ez az igen finom, sőt, igen imponáló megfigyelés távolra mutató igazság, s az osztályozás valódi mélységeit tárja fel. Azért is, mert ennek a merőben új osztályozási koncepciónak, amely az automatikus osztályozásban jelentkezik és "csak" statisztikus összefüggésekre mutat rá, megmutatja mély kapcsolatait a többi osztályozással. De egészen más úton jut el ugyanahhoz az igazsághoz. Ismeretes ugyanis, hogy az emberi osztályozásban általá-

ban a dokumentum témájánál egy -két fokozattal magasabb szinten megy végbe az osztályozás. Ennek révén válik lehetővé pl. struktúrák létrehozása is tárolási rendszerekben. Mert nemcsak individuum van, hanem osztály is, logikai értelemben.

Ez a megfigyelés azért jelentős, mert támpontot nyújt az automatikusan meghatározott kifejezések további automatikus kezeléséhez, akár hierarchikus kapcsolatok meghatározásához.

Harmadszor, igen lényeges adalék az informatika relevancia elméleteihez, amely elméletek távolról sem megoldott, de az informatika központi problémájára vonatkoznak.

A vizsgált két eloszlási görbe, úgy tűnik, kínálja önmagát, szinte mechanikusan lehet kivágni az adott célú intervallumot, s meghatározni az ide tartozó kifejezéseket. Valójában azonban itt is csupán statisztikus összefüggésekről van szó. Éppen ezért a gyakorlati kulcsszó-meghatározások eseteiben gyakran alkalmazzák az un. nem rögzített szignifikancia módszerét. Ez abból áll, hogy bizonyos szavakról, vagy valamely szóról előre el lehet dönteni, hogy fontosak, s a szignifikancia görbe csúcsát mesterségesen e szóhoz tartozó gyakorisághoz

helyezzük. /A szignifikancia görbét úgy toljuk el, hogy csúcsa e kívánt helyre essék./ Ezzel módosul a gyakorisági intervallum is, amely szintén eltolódik. Az eljárás azonban azt is lehetővé teszi, hogy kisebbre szabjuk az intervallumot, tehát általában kevesebb kulcsszó meghatározását teszi lehetővé, úgy, hogy biztonságosabban lehet a kulcsszavakat értékelni a fölösleg /redundancia/ csökkentésével.

III. SZÓELEMZÉSTŐL A KLASZTERÁLÁSIG

Mit tud kezdeni az információs technológia az előző fejezet módszereivel meghatározott osztályozási kifejezésekkel? Számos kérdés megválaszolatlan még. Az válthatja ki elsősorban a kritikát, hogy bár automatikusan meghatározható az az ismérvhalmaz, amely a dokumentumokhoz rendelve objektíven tükrözi annak tartalmát, de a szövegek előkészítése nagyobb munkát igényel, mint ugyanezeknek a dokumentumoknak emberi osztályozása. A technikai jellegű munkák mennyisége is óriási: az elemzendő szöveget számítógépes input formájában kell elkészíteni, ez a szokásosnál nagyobb mennyiségű lyukasztást jelent. További kérdés, hogy mit érünk olyan automatikus eljárásokkal, amelyek csak egyetlen nyelvű szövegre vonatkoznak, esetünkben a magyar nyelvűekre. Van-e a magyar nyelvű szakirodalomnak akkora jelentősége, hogy kedvéért érdemes a gyakorlati hasznosítást is biztosító kutatásokat elvégezni, algoritmusokat kidolgozni? És mi az eredmény? Egyelőre egy ismérvhalmaz, szavak és szerkezetek, köztük semmiféle összefüggést nem lehet felfedezni, a kifejezések csoportjait csupán az tartja össze, hogy ugyanarra a dokumentumra vonatkoznak. Lehet-e ezekkel a - majdnemhogy ömlesztett - ismérvekkel olyan tárolási megoldásokat találni, amelyek biztosítják adott kérdésre a relevans dokumentumok szelektálhatóságát?

Ismérveink jótulajdonságai abban foglalhatók össze, hogy "természetesek". Az élő tudományos közlés kifejezései, arányaikban is követik a szaktudományok kérdésfelvetéseit, mélységben /specifikusság/ szintén igazodnak a kutatási témákhoz.

Ami a nyelvi elemzés algoritmusait illeti, azt kell belátni, hogy jelentőségük nemcsak az informatikában van. Fő haszonélvező a nyelvtudomány. A leíró grammatikának olyan részletes és pontos kidolgozását eredményezi, amely a számítógépes nyelvészet előtt fel sem merülhetett. A nyelvtudomány is operál gyakoriságokkal: segédeszközként ezért készülnek általános gyakorisági szótárak, amelynek bázisán szaknyelv, írói nyelvek vizsgálhatók, stilisztikai elemzések végezhetők. Szótárszerkesztők munkáját könnyíthetik meg. Egysszóval, nyelvészeti kutatások egyre inkább támaszkodhatnak az automatikus nyelvelemzés módszereire és eredményeire. Ezek az algoritmusok máshol sem az informatika sugalmazására születtek, hanem a nyelvészeti kutatások öntörvényű fejlődése kívánta meg megalkotásukat. Az informatika csak betársult, igaz, hogy ipari méretekben támasztott igényeket. Amikor az informatika érdeklődése a nyelvészet felé fordult számos nehézségére keresvén megoldást - a 60-as évek Amerikájában - a kívánt algoritmusok jórészt készen állottak. Más tudományok újszerű módszerei is támaszkodni kívánnak a nyelvi

elemzésre. A tudománymetria egy-egy tudományt, diszciplínát reprezentáló szakkifejezések /= kulcsszavak/ statisztikai analízisével kíván a tudományokra következtetéseket levonni. Hasonló módszereket alkalmaz a szociológia is, amikor tartalomelemzést végez. A példák igazolják, hogy az automatikus nyelvi elemzés számos ágon kapcsolódik más tudományokhoz, s többé-kevésbé azok valamilyen nagyon korszerű módszerének alapját alkotja.

Egyúttal annak igazolását is látni ebben, hogy az informatika a többi tudománnyal szoros szinbiózisban fejlődik, s nem szigetelődik el önjáró problémaköreivel, mint ahogyan a tradicionális könyvtártudomány esetében számos vonatkozásban ez volt tapasztalható.

Ilyen összefüggésekben valószínű, hogy megéri a szükséges algoritmusokat kidolgozni. Nemcsak az informatikának követelménye ez, hanem a korszerű társadalomtudományi kutatások előfeltétele is. Olyan módszerről van szó továbbá, amelyet nem kell külön-külön kidolgozni valamennyi információs rendszernek, valamennyi tudománynak, hanem egyszer kell megszületnie.

Ami a kulcsszavak további sorsát illeti, több utat lehet vázolni. Ugy kell ezeket tekinteni, mint több információfeldolgozási folyamat alapját, mint amiből számos további elágazás ered. Idézni lehetne újra Saltont, aki a kulcs-

szavakat nemcsak az automatikus osztályozás számára szánt kifejezés-gyűjteménynek tekintette, hanem bármely más információ-kereső nyelv lehetséges szótári egységeinek, továbbá számos szótári segédeszköz induló szókincsének, de különféle elemzések alapadatainak is. Információs teauruszok leghathatósabb szógyűjtési módszere is a természetes nyelvű szövegek analízisén nyugszik. A legegyszerűbb további felhasználás az, hogy pozitív, vagy negatív szótárak /stoplisták/ formájában állnak rendelkezésre a legkülönbözőbb információs feldolgozási folyamatokban. Indexművek ezeket már évtizedek óta használják. A negatív szótár /stoplista/ - amely értelemszerűen a legmagasabb és legalacsonyabb gyakoriságú szavakat tartalmazza - kizárat kifejezéseket a további feldolgozásból. Példázzák ezt a szövegkörnyezetes kulcsszó-indexek. Ami e kizárás után megmarad, az a további feldolgozás tárgya. A pozitív szótárak éppen ellenkezőleg, azoknak a kifejezéseknek jegyzékei, amelyeket a további feldolgozás megtart, vagy ezeket kell felismerni.

/Pl. szövegkörnyezetes tárgyszóindexek./ A példák szaporíthatók lennének.

Jelen tanulmány nem kívánja ezeket a lehetséges utakat bejárni. Egyetlen további problémát tárgyal, az automatikus osztályozást, tehát azt a kérdést, hogy milyen algoritmusok vannak ezeknek a kifejezéseknek olyan fajta kezelésére, amelynek eredménye összetartozó csoportok,

logikai értelemben vett osztályok. Még hozzá olyan eljárásokat kell találni, amelyek mind az objektumok, dokumentumok, mind pedig az ismérvek, kulcsszavak összetartozó csoportjainak meghatározására valók. Ha a csoportosítási probléma megoldására találunk alkalmas módszert, akkor ennek olyannak kell lennie, hogy megfelelő absztrakt reprezentációval lehessen a problémákat felírni, automatikus algoritmusok ugyanis ezen a szinten adhatók meg.

A kérdés kulcsa valóban a reprezentáció. Ez általánosabban is igaz. D.M.MacKay híres könyvében /Information, Mechanism and Meaning/ egyenesen az információ-elmélet tárgyaként határozza meg az információ reprezentációját. Álláspontjában sok igazság van. Megfelelő reprezentáció a probléma világos megfogalmazását jelenti és olyan ábrázolását, amely biztosítja az elvi megoldást.

A megfelelő reprezentáció a probléma felírása absztrakt jelek formájában. Ez nem feltétlenül matematikai ábrázolást jelent, bár hajlamosak vagyunk mindent matematikainak tekinteni, ami absztrakt. A tudományok mindig is ezzel a módszerrel éltek. Ju.Sz.Sztyepanov kitűnő könyvében /Szemiotika/ idézi a XVI. század nagy orvosát, Paracelsust, aki ezt az alábbi hitvallásában fogalmazza meg. "Mi emberek a jelek és külső hasonlatosságok segítségével tárjuk fel azt, amit a belső magába zár; ekképpen rátalálunk a füvek és kövek összes sajátosságaira. Nincs

semmi a tengerek mélyén, sem a mennybolt magasán, amit az ember ne volna képes feltalálni. Nincs olyan hegy, bármilyen magasán is lett légyen, amely elrejtethné méhét az ember tekintete előtt, az egymásnak megfelelő jelek köntösében megmutatja magát".

A következő fejezet ezt a reprezentációt keresi meg, a jeleket, amelyben megmutatja magát az, amit a belső magába zár.

IV. AUTOMATIKUS OSZTÁLYOZÁS

Az előző fejezetben keresett reprezentációt az un. információs mátrixban találhatjuk meg. Bármely visszakereső rendszer két legfontosabb alkotója az objektumok, dokumentumok és a jellemzésükre szolgáló ismérvek, kulcsszavak, osztályozási kifejezések. Minden dokumentumhoz az ismérvekből kiválasztott részhalmoz, az ismérvek egy sorozata tartozik. Minden ismérvhez pedig a dokumentumok egy sorozata.

Ennek alapján lehet definiálni az információs mátrixot, amely könyvtárak, bibliográfiák, szakirodalmi visszakereső rendszerek, stb. modellálására szokásosan alkalmazható és amely további vizsgálódásaink kiinduló pontja.

Mátrixon meghatározott elemeknek táblázatos formában /sorokban és oszlopokban/ való elrendezését értjük. Esetünkben olyan táblázat alkotja a mátrixot, amelynek minden sora egy dokumentumot képvisel, minden oszlopa egy-egy ismérvet, egy osztályozási kifejezést, kulcsszót. Neve ezért dokumentum-ismérv mátrix. Egyszerűség kedvéért példánkban legyen hat dokumentum és tíz osztályozási kifejezés, amelyet tehát az alábbi módon írunk fel.

	Állomány	Bibliográfia	Elemzés	Kölcsönzés	Könyvtár	Mutató	Osztályozás	Számítógép	Szerkesztés	Történet
1. dokumentum	0	1	0	0	0	0	0	0	0	1
2. dokumentum	0	1	0	0	0	1	0	1	1	0
3. dokumentum	1	0	1	0	1	0	0	0	0	0
4. dokumentum	0	1	1	0	0	0	1	0	0	0
5. dokumentum	0	0	0	1	1	0	0	1	0	0
6. dokumentum	0	1	0	0	0	0	0	1	0	1

Ha egy dokumentum megkap egy kulcsszót, akkor az adott dokumentum sorának és a kulcsszó oszlopának találkozási helyére, pozíciójába 1 számot írhatunk, egyébként zérust. A fenti mátrix minden sorát el tudjuk olvasni, pl. az 1. számú dokumentum a bibliográfiák történetével foglalkozik. A sorok tehát egy-egy dokumentum képét adják, jellemző jegyeinek összeségét, amelyet egy 0 és 1 jelekből álló jelsorozat, az un. vektor reprezentál. Minden oszlop pedig egy jellemző jegy, ismerv képe. Arra vonatkozóan nincs előírás, hogy a mátrix elemei csak 0 és 1 értékeket vehetnek fel. Súlyozott osztályozás esetén az 1-es helyén állhat 1, 2, 3,.. szám, jelezve, hogy a dokumentumot jobban, vagy kevésbé jellemzi az adott kulcsszó.

1. A klasszikus logika problémája

Az "osztályozás" lényegesen szélesebb körű intellektuális művelet annál, hogy leszűkíthetnénk a dokumentumok osztályozásának problémakörére. J. Piaget szerint az intellektuális struktúrák három alapra vezethetők vissza: osztályozási, viszony- és topológiai struktúrákra. Az elsőnek az a kérdése: mi mibe tartozik bele, illetve, hogy mi mit tartalmaz. Az aristoteleszi logika kidolgozott fogalomtanában objektumok osztályba sorolását úgy oldotta meg, hogy a tartalmi jegyek közül egyet /vagy néhányat/ kitüntetett, s az objektumok e kitüntetett jellemzők alapján kerültek osztályokba.

J. K. Vojšvillo az objektumok és tulajdonságaik viszonyát a következőként határozza meg. "A logikában ismertetőjegynek nevezzük azt, aminek alapján valamely osztály tárgyait kiemeljük, és ami lehetővé teszi azt a következtetést, hogy valamely tárgy egy adott osztályhoz tartozik hozzá. Ezzel az illető tárgyat az adott tárgyakkal azonosítjuk, vagy pedig megkülönböztetjük az ilyen minőségű tárgytól... Az "ismertetőjegy" szó használatakor a tárgyak minőségére, tulajdonságára, más tárgyakhoz való viszonyára stb. gondolunk. Ezeket a vonásokat a megismerés folyamatában szintén kiemeljük, általánosítjuk és a nyelv terminusaiban rögzítjük." A fogalom. Bp. Gondolat, 1978. 206. l.

Érthetőbben: egy vagy több tulajdonsággal "osztályok" definiálhatók, s minden objektum, amely az adott tulajdonsággal rendelkezik, az osztályba besorolásra kerül. Csakhogy minden objektumnak több tulajdonsága van /elméletben végtelen számú/, így minden objektum más-más tulajdonságai alapján más-más osztályba kerülhet. A tudományos tevékenység gyakorlatában ezért mindig igyekeztek megkeresni azokat a "fontos", "lényegi" sajátosságokat, amelyekkel osztályok generálhatók.

Atekintetben, hogy melyek a "lényeges" sajátosságok, tudományos iskolák, irányzatok csaptak össze /de ezek a kérdések már kívül estek a logika hatáskörén/.

Annyit azonban meg kell jegyezni, hogy a "lényeg" logikai értelemben is két dolgot jelöl. Egyik jelentésében "valami szubsztanciális", amelyet már Aristoteles is alapvető kategóriái közé sorolt, méghozzá legelsőnek. Másik jelentése: ami meghatározza egy osztály minőségét. Annyi bizonyos, hogy a "lényeg" történetileg is változik.

Pl. Darwin rendszerezése az élővilágról a fajok eredetén alapult, míg Linné a morfológiai hasonlóság alapján alkotta meg rendszerét. Akárhogyan alakult is azonban az osztályba sorolás szempontja, egy hátrányt nem lehetett leküzdeni: az objektumok valamely tulajdonságának kiemelése az osztályba sorolás kedvéért a

többi sajátosság negligálásával jár együtt. Ha például a "korona" besorolása került a koronázási ékszerek osztályába, akkor egyben elveszett számos más tulajdonsága, pl. hogy ötvösművészeti termék, nemesfémből készült termék, fejdísz, stb. Az osztályok terjedelmét változtatni, bővíteni-szűkíteni lehetett azáltal, hogy meghatározásuk kevesebb-több tulajdonságon nyugodott, de az alapprobléma változatlanul megmaradt.

A kérdés tehát úgy szól: megalkothatók-e tárgyak, objektumok csoportjai olyan eljárással, amely nem valamely tulajdonság /vagy tulajdonságok/ kiemelésén alapul, hanem egyidejűleg veszi figyelembe az objektumok valamennyi tulajdonságát /mivel e sajátosságok száma végtelen - mint tisztáztuk fentebb - megelégszünk azzal, ha elegendően nagyszámú tulajdonságát/, - továbbá, hogy nem részesíti előnyben egyik vagy másik tulajdonságot, hiszen ez a megítélések szubjektív forrásává válhat.

Az előző pont mátrixát tekintve tehát az a kérdés, lehet-e eljárást találni e dokumentumok természetes csoportjainak kialakítására úgy, hogy valamennyi kulcsszót egyidejűleg veszünk figyelembe, s nem preferáljuk egyiket sem. A mátrixot figyelembe véve ez az eljárás a dokumentumvektorok /a mátrix sorai/ alkotóelemeinek vizsgálatán nyugodhat.

Mit is jelent egy sorvektor? Az $\underline{a} = /1000/$ négy komponensű sorvektor pl. azt, hogy az \underline{a} vektor által reprezentált dokumentumot az első ismértv jellemzi, logikailag azt, hogy beletartozik az első ismértvvel meghatározható osztály tárgyai, objektumai közé. Ha a $\underline{b} = /10100/$ sorvektort tekintjük, akkor a \underline{b} vektor által reprezentált dokumentumot az első és harmadik ismértv határozza meg, beletartozik az első és harmadik ismértvvel definiált osztályba. De az első ismértv közös, tehát mind \underline{a} , mind \underline{b} ugyanabba az osztályba tartozik, de \underline{b} -t még egy ismértv jellemez, tartalmi jegyei tehát bővebbek, ezért közöttük fölé-alárendelési kapcsolat van, azaz \underline{b} dokumentum \underline{a} -nak egy részletezőbb, specifikusabb témájával foglalkozik. Ugyanigy belátható, hogy $\underline{c} = /111000/$ és $\underline{d} = /101010/$ által reprezentált dokumentumok közt részleges tartalmi fedés van, mert vannak azonos jegyeik, de mindkettőnek van olyan ismértve is, amely a másiknak nincs. Logikailag metszetviszonyt fedezhetünk fel. Tegyük fel, hogy dokumentumainkat n komponensű vektorokkal reprezentáljuk, tehát n kulcsszó alkotja az osztályozási szótárt. Hány osztály alkotható a komponensek vizsgálata alapján? Nyilván osztályt alkotnak ^{csak} a dokumentumok, amelyeket jellemez az első ismértv, azok, amelyet a második ismértv,... azok, amelyet jellemez az n -edik ismértv. Osztályt alkotnak azok, amelyeket az első és második ismértv együttesen jellemez, amelyeket az

első és harmadik ismerv együttesen,...az ~~első~~ és n -edik együttesen. Osztályt alkotnak azok, amelyeket az első, második és harmadik kulcsszó jellemez együttesen, s.i.t. Összesen 2^n osztály adódik elméletben, annyi, ahányféleképpen az n ismerv variálható.

Az osztályalkotásnak ez a módszere automatizálható, de még megmaradtunk a klasszikus logika határain belül.

Az oszlopvektorok az ismérvek tartalmi rokonságát tükrözik. A 2. fejezetben a statisztikai kulcsszó-elemzés lehetséges kiterjesztései közt szerepelt Saltonnak az a módszere, hogy az ismérvek olyan asszociációit vizsgálta, amelyek azt mutatták, mely ismérvek járulnak ugyanazokhoz a dokumentumokhoz. Ahol ez az asszociáció szorosnak mutatkozott, ott az ismérvek tartalmi rokonságát tételezte fel. Illusztrálásul tekintsük az alábbi oszlopvektorokat.

<u>a</u> =	1	<u>b</u> =	1	<u>c</u> =	1	<u>d</u> =	0
	0		0		0		1
	1		1		1		0
	0		0		0		1
	1		0		1		0

Az a és c vektorok kivétel nélkül ugyanazokhoz a dokumentumokhoz tartoznak. Salton köztük szinonima

viszonyt határozott meg. Azt kell belátni, hogy az automatikus kulcsszóelemzés által meghatározott ismérvek közt szerepelnek a szinonimák is, hiszen a kulcsszavak közt még semmiféle szabályozás nem ment végbe. A hipotézis jogos, a SMART rendszerben kísérletileg is igazolódott. Ha most a és b vektort tekintjük, köztük valamilyen tartalmi rokonság van, hiszen ugyanazokhoz a dokumentumokhoz tartoznak, de a ezen túlmenően még plusz dokumentumhoz is. Ha a gyakorlatban adódó több száz, vagy több ezer komponensű vektornál az azonos pozícióban álló komponensek döntő többsége azonos, akkor még lehet őket szinonimáknak tekinteni. Így lehet a szinonimiának egy, a nyelvészeti szemantikától eltérő, statisztikus definícióját megkapni. Nyilvánvaló, hogy d vektor olyan ismerv, amelynek jelentése idegen, összeegyeztethetetlen a többitől.

Az informatikában a sorvektorokkal való manipuláció a dokumentumok csoportbasorolását eredményezi, míg az oszlopvektorokra vonatkozó vizsgálatoknak az osztályozási rendszerek alkotásában, az ismérvek analízisében van szerepe.

A dokumentumok csoportbasorolásában jelölhető meg a fő célkitűzés. Ennek eléréséhez azonban az eddigi eljárásoknál lényegesen finomabb módszerekre van szükség.

2. A távolság meghatározása

Alkalmasnak látszik erre a célra a dokumentumvektorok közti un. "távolság" illetve un. "közelség" meghatározása. Ebből kettős nyereség származik. Egyrészt eleget teszünk a dokumentumok osztályozásával szemben támasztott követelménynek, mely szerint az osztályozás fő célja a hasonló dokumentum hozzárendelése a hasonlóhoz. Másrészt alkalmas számítási eljárást nyerhetünk a hasonlóság - vagy az ezzel analóg távolság és közelség - mérésére.

Mégegyszer tudatosítanunk kell, hogy a dokumentumvektorok elemei /komponensei/ egy tulajdonságghalmaz /osztályozási kifejezések/ elemei, s e tulajdonságok egy sorozata - esetünkben bináris értékeket véve fel - alkotja a vektort.

A távolság meghatározására több módszer létezik. Az un. Hamming-féle távolság /R.W.Hamming matematikusról, az információelmélet kiválóságáról elnevezve/ az egybe nem eső, a különböző komponensek számát veszi figyelembe. Ha $\underline{a} = 000$ és $\underline{b} = 010$, akkor a köztük lévő távolság 1, mert 1 komponensben különböznek. Ha visszatérünk a példaként adott információs mátrixra, akkor az 1. és 2. dokumentum közti távolság 4, mert négy pozícióban állnak különböző elemek. Az 1. és 3. dokumentum

távolsága 5, és így tovább. De a példa vizsgálatából az is kiderül, hogy a távolság csak nagyon gorombán tükrözi a tartalmi különbségeket, mert függ olyan tényezőktől is, mint pl.: milyen az indexelés mélysége, azaz átlagosan hány kulcsszóval osztályozzuk a dokumentumokat.

A k ö z e l s é g azokat a pozíciókat veszi figyelembe, ahol mindkét vektor azonos pozíciójában nem zérus elem áll. Az előzőekkel ellentétben itt éppen az e g y e z ő k o m p o n e n s e k számát kívánjuk meghatározni, nem a különbség, hanem az azonosság mértékét. Más szóval: azt kívánjuk meghatározni: hány-szor szerepel ugyanaz a kulcsszó az összehasonlított két dokumentum leírásában. Ezt a mérőszámot úgy kapjuk meg, ha a két vektor azonos pozíciójában álló elemeket összeszorozzuk. Az így kapott nem zérus elemek összege adja azt a számot, amely az azonos pozícióban álló értékek számát mutatja, tehát azt, hogy a két dokumentum hány-szor kapott ugyanolyan kulcsszót. Ha a példaként adott mátrixot nézzük, látjuk, hogy az 1. és 2. dokumentum csupán egyszer kapta ugyanazt a kulcsszót /többi elemében különbözik/, a 2. és 3. dokumentum egyetlen egyszer sem kapta ugyanazt, viszont a 2. és 6. dokumentumnál két ízben fordul elő ugyanaz a kulcsszó /"bibliográfia" és "számítógép"/.

A számítási eljárás tehát a következő. A két vektor azonos komponenseit összeszorozzuk, s a kapott értékeket összeadjuk. Ha a két vektort \underline{a} és \underline{b} jelöli /komponenseik a_i és b_i , ahol i értéke az n komponensen fut végig, - a példában 1-től 10-ig/, akkor a hasonlóságot kifejező $h_{\underline{a}, \underline{b}}$ függvény

$$h_{\underline{a}, \underline{b}} = \sum_{i=1}^n a_i b_i$$

Végezzük el a számítást a példa 2. és dokumentumán

\underline{a} /2.dok./ = /0100010110/

\underline{b} /6.dok./ = /0100000101/

$\underline{a_i b_i}$ = /0100000100/ /komponensenkénti szorzat/

$$\sum_{i=1}^{10} \underline{a_i b_i} = 2 \text{ /komponensek összege/}$$

Ennek a módszernek két gyengéje van. Az első, hogy két dokumentum abban is hasonló lehet, hogy mely kulcsszavakat nem kapták meg egyszerre /ezt a fenti érték nem jelzi/. Valami "egyikre sem jellemző" tulajdonság ez. A másik hátrány az, hogy a közös tulajdonságok számát kifejező függvény értékét abszolút

számban kaptuk meg, ezért bizonytalanul tudunk íté-
letet alkotni a hasonlóság mértékéről. Jobban szeret-
jük az olyan mérőszámokat, amelyeknél ismerjük a felső
és alsó határokat, azokat az értékeket, amelyeket a h
függvény egyáltalán felvehet. Az első probléma megol-
dását röviden vázoljuk, a másik probléma pedig azokhoz
az eljárásokhoz vezet, amelyekkel már gyakorlati rend-
szerekben is lehetséges megfelelő hasonlósági számítás.

Az első problémán úgy tudunk segíteni, hogy nemcsak
az azonos pozícióban álló "1-es" elemeket, hanem az
azonos pozícióban álló zérus elemeket is összeszámlál-
juk. Ez a művelet úgy hajtható végre, hogy képezzük a
vektorok komplementereit^x, s az így kapott komplementer
vektorokkal ugyanúgy végezzük el a számolást, mint az
eredeti vektorokkal. A képlet így alakul végül:

$$h_{\underline{a}, \underline{b}} = \sum_{i=1}^n a_i b_i + \sum_{i=1}^n \bar{a}_i \bar{b}_i$$

ahol \bar{a} és \bar{b} a komplementer vektorok.

Az előbbi példa ezzel a képlettel:

$$\underline{a} = /0100010110/ \quad \bar{a} = /1011101001/$$

$$\underline{b} = /0100000101/ \quad \bar{b} = /1011111010/$$

$$\underline{a_i} \cdot \underline{b_i} = /0100000100/ \quad \bar{a_i} \cdot \bar{b_i} = /1011101000/$$

$$h_{\underline{a}, \underline{b}} = 2 + 5 = 7$$

^x Egy vektor komplementere az a vektor, amelynek pozíciójába 0 helyett 1, illetve 1 helyett 0 kerül.

A teljesség kedvéért meg kell említeni, hogy "euklideszi távolságon" az alábbiit kell érteni. Legyen a és b vektor, akkor euklideszi távolságukon a

$$d/a, b/ = \sqrt{\sum_{i=1}^n /a_i - b_i/^2}$$

érték értendő, és $0 \leq d/a, b/ \leq \sqrt{n}$, bináris esetben a Hamming távolságot adja. /A képlet verbálisan: a komponensek különbségeinek négyzetösszegéből vont négyzetgyök./

G.N.Zsitkov összefoglaló tanulmányában a "távolság" kiszámítására hat különböző formulát ismertet, érzékelte, hogy ez a kérdés is összetett problémát takar.

3. Hasonlósági függvények.

A fentebb ismertetett távolsági mérőszámok valamilyen módon tükrözik a dokumentumok közti hasonlóságot, tartalmi rokonságot. Nagy hátrányuk azonban, hogy az abszolút számban kapott mértékszámokkal nem tudunk egzaktan bánni, mivel nem tudjuk pontosan, mit is jelent pl. a

8, vagy a 34,... mértékű távolság. A mértékeket tehát egy olyan intervallumban kívánatos megkapni,- mondjuk 0 és 1 közti intervallumban -, ahol a maximális hasonlóság /azonosság/ értékéhez - 1-hez - tudjuk a hasonlósági értékeket hozzávetni. A teljes különbözőséget /semmiben sem hasonlók/ kifejező 0 és maximális hasonlóságot /azonosság/ kifejező 1 érték között bevezethetünk meghatározott küszöbértéket is, amely fölött hasonlóságot mutató dokumentumok egy csoportba sorolhatók. Ez a küszöbérték /cut-off level/ az ismert rendszerekben általában 0,7 körül mozog. Ha alacsonyabb, akkor a hasonlósági csoportba már kisebb mértékben hasonló dokumentumok is belekerülnek, míg ha magasabb, akkor a hasonlóság az egy csoportba kerülő dokumentumok között szorosabb. Ilyen módon lehet szabályozni, hogy kevesebb, de nagyobb és átfogóbb, vagy több, de magasabb homogenitást mutató dokumentumcsoportot kapjunk. Az így kialakult objektum - dokumentum csoportokat klaszternek /cluster/ hívjuk. A szó eredeti jelentése: halom, csoport - egy rakás valamiből. Az elnevezés utal arra, hogy nem osztályról van szó, mint a hagyományos logikán nyugvó osztályozás esetében, hiszen itt már nincs szó logikai értelemben vett osztályba sorolásról. De a halmaz szótól is meg kell különböztetni. A halmaz szónak ugyanis más matematikai jelentése van /bár az egy klaszterbe sorolt

- vagy ide került - dokumentumok összessége bizonyos esetekben halmazoknak tekinthető és halmazokként kezelhető/.

A képletek megértéséhez előzetesen értelmezni kell a műveleteket. D.S o e r g e l nyomán e műveleteknek az alábbi jelentést fogunk tulajdonítani. /17/

Ha adva van két vektor, \underline{a}_i és \underline{b}_i akkor ezek közös részén, metszetén, szorzatán az a \underline{c} vektor értendő, amely

a/ az azonos pozícióban álló komponensek közül a kisebbiket tartalmazza:

$$\underline{a}_i \cap \underline{b}_i = \underline{c}$$

$$\text{ahol } \underline{c} = \min/a_i, b_i/$$

$$\underline{a}_i = /111000/$$

$$\underline{b}_i = /101010/$$

$$\underline{c} = /101000/$$

$$\text{b/ } \underline{a}_i \cap \underline{b}_i = \underline{c}$$

ahol $\underline{c} = a_i b_i$ tehát a két vektor komponensenkénti szorzata.

Példa:

$$\underline{a}_i = /111000/$$

$$\underline{b}_i = /101010/$$

$$\underline{c} = /101000/$$

/Bináris vektorról lévén szó, a két eredmény azonos/.

Legyen két nem bináris vektor

$$\underline{d} = /12003/ \text{ és } \underline{e} = /20012/$$

Metszetüket Soergel alapján az alábbi két módon lehet értelmezni:

$$\begin{array}{l} \text{a/} \quad \underline{d} = /12003/ \\ \quad \underline{e} = /20012/ \\ \hline \underline{e} = \min/\underline{d}, \underline{e}/ = /10012/, \text{ az azonos pozícióban álló} \\ \text{komponensek közül a kisebbiket választjuk.} \end{array}$$

$$\begin{array}{l} \text{b/} \quad \underline{d} = /12003/ \\ \quad \underline{e} = /20012/ \\ \hline \underline{e} = \underline{d} \cdot \underline{e} = /20006/, \text{ komponensenkénti szorzat.} \end{array}$$

Ezt a példát azért volt szükséges megadni, mert a metszet, szorzat két interpretációja csak nem bináris vektorok esetében különbözik.

Két vektor összegén, unióján, egyesítésén az a \underline{c} vektor értendő, amely

$$\text{a/ } \underline{a}_i \cup \underline{b}_i = \underline{c}$$

ahol $\underline{c}_i = \max/\underline{a}_i, \underline{b}_i/$, azaz \underline{c} a két vektor komponens közül mindig a nagyobbikból áll.

$$\begin{array}{r} \underline{a}_i = /111000/ \\ \underline{b}_i = /101010/ \\ \hline \underline{c} = /111010/ \end{array}$$

$$b/ \underline{a}_i \cup \underline{b}_i = \underline{c}$$

$$\text{ahol } c = \underline{a}_i + \underline{b}_i$$

$$\begin{array}{r} \underline{a}_i = /111000/ \\ \underline{b}_i = /101010/ \\ \hline \underline{c} = 212010 \end{array}$$

Más szóval \underline{c} vektor az azonos pozícióban álló komponensek összegéből áll.

Az előbbi két nem bináris vektorral elvégezve ugyan-
ezen műveleteket, azt kapjuk, hogy

$$\begin{array}{r} \underline{d} = /12003/ \\ \underline{e} = /20012/ \\ \hline \underline{c} = \max/\underline{d}, \underline{e}/ = /22013/ \end{array}$$

illetve

$$\begin{array}{r} \underline{d} = /12003/ \\ \underline{e} = /20012/ \\ \hline \underline{c} = /32015/, \text{ komponensenkénti összeg.} \end{array}$$

Végül az $\underline{a} = /111000/$ vektor komplementerén az
 $\bar{\underline{a}} = /000111/$ vektor értendő.

Bináris esetekre az egyesítésre és szorzatra általában csak az a/ alatt tárgyalt formulák használatosak.

Milyen jelentést tulajdoníthatunk a fenti képleteknek? A metszet azokat az ismérveket eredményezi, amelyeket mind a két dokumentum megkapott, tehát a közös tulajdonságokat. /Ha a vektor nem bináris, akkor az ismérveket súlyozottan adtuk a dokumentumoknak, azaz, hogy kevésbé, vagy erőteljesebben jellemző-e. A metszetben megjelenik ez a tulajdonság, a/ esetben - jeleztük, hogy ezt használjuk a továbbiakban - a kisebbik súllyal./ A b/ alatti esetek vektoralgebrai műveletek.

A két vektor egyesítése /összeg/ olyan vektort eredményez, amely kifejezi azt, hogy a két dokumentum bármelyike milyen ismérveket kapott, függetlenül attól, hogy mindkettő megkapta-e, vagy csak az egyik. Kettejük összes tulajdonsága jelenik meg c vektorban.

Ezek előrebocsátásával a dokumentumvektorok hasonlóságát kifejező különböző függvényeket az alábbiakban foglaljuk össze. A közölt áttekintés M.Fritschetől származik./7/

A képletekben az N a komponensek összegezésének jele.

Legyen a és b két összehasonlításra szánt, egységesen n komponensből álló bináris vektor.

Hasonlósági függvények	Intervallum
1. $h/\underline{a}, \underline{b}/ = N/\underline{a} \cap \underline{b}/$	0,n
2. $h/\underline{a}, \underline{b}/ = \frac{1}{n} N/\underline{a} \cap \underline{b}/$	0,1
3. $h/\underline{a}, \underline{b}/ = \frac{1}{n} [N/\underline{a} \cap \underline{b}/ + N/\bar{\underline{a}} \cap \bar{\underline{b}}/]$	0,1
4. $h/\underline{a}, \underline{b}/ = \frac{1}{n} [nN/\underline{a} \cap \underline{b}/ + nN/\bar{\underline{a}} \cap \bar{\underline{b}}/]$	0,n
5. $h/\underline{a}, \underline{b}/ = \frac{N/\underline{a} \cap \underline{b}/}{N/\underline{a}/ + N/\underline{b}/ - N/\underline{a} \cap \underline{b}/}$ bináris esetben $\frac{N/\underline{a} \cap \underline{b}/}{N/\underline{a} \cup \underline{b}/}$	0,1
6. $h/\underline{a}, \underline{b}/ = \frac{N/\underline{a} \cap \underline{b}/}{N/\underline{a}/ + N/\underline{b}/}$	0,1
7. ÁTFED $h/\underline{a}, \underline{b}/ = \frac{N/\underline{a} \cap \underline{b}/}{\min N/\underline{a}/, N/\underline{b}/}$	0,1
8. ASZIM $h/\underline{a}, \underline{b}/ = \frac{N/\underline{a} \cap \underline{b}/}{N/\underline{a}/}$	0,1
9. cos $h/\underline{a}, \underline{b}/ = \frac{N/\underline{a} \cap \underline{b}/}{\sqrt{N/\underline{a} \cap \underline{a}/ \cdot N/\underline{b} \cap \underline{b}/}}$	0,1
10. $h/\underline{a}, \underline{b}/ = \frac{nN/\underline{a} \cap \underline{b}/ - N/\underline{a}/ \cdot N/\underline{b}/}{\sqrt{nN/\underline{a} \cap \underline{a}/ - N/\underline{a}/^2 \cdot nN/\underline{b} \cap \underline{b}/ - N/\underline{b}/^2}} - 1,1$	1,1

Látható, hogy e képletek legtöbbszörének alapgondolata az, hogy az összehasonlítandó vektorok /tulajdonságok/ közös részét fejezik ki az összes tulajdonság hányadában. Lefordítva az osztályozás problémakörére a fenti-eket, a képletek úgy fejezik ki a dokumentumok hasonlóságát, hogy a két dokumentum közös jellemzőit osztják a két dokumentumnak kiadott valamennyi jellemző számával. Ezt az alaphelyzetet finomítják azzal, hogy közös "nemjellemző" deszkriptorokat is figyelembe vesznek, hogy egyik vagy másik esetben nagyobb jelentőséget tulajdonítanak egyik vagy másik tényezőnek. Néhány elterjedt, széles körben alkalmazott függvényt vizsgáljuk meg közelebbről.

Az 5. képletet Tanimoto formulának is hívják, amely valóban azt testesíti meg, amit fentebb a formulákról mondtunk. A képlet számlálójában két vektor közös része áll, nevezőjében összege. A példaként közölt információs mátrixban az 1. és 6. számú dokumentum hasonlósága Tanimoto alapján /képviselje őket a és b/:

$$\begin{array}{rcl} \underline{a} & = & /0100000001/ \\ \underline{b} & = & /0100000101/ \\ \hline \underline{a} \cap \underline{b} & = & /0100000001/ \\ N/\underline{a} \cap \underline{b}/ & = & 2 \\ \underline{a} \cup \underline{b} & = & /0100000101/ \\ N/\underline{a} \cup \underline{b}/ & = & 3 \end{array}$$

Ennek alapján

$$h/\underline{a}, \underline{b}/ = \frac{2}{3}$$

Más szóval: három kiadott ismerv, kulcsszó közül kettő közös. Tanimoto módosított formulája /6.számú képlet/ alapján

$$N/\underline{a}/=2, N/\underline{b}/=3, N/\underline{a}/ + N/\underline{b}/=5, -$$

a hasonlósági mérték

$$h/\underline{a}, \underline{b}/ = \frac{2}{5}$$

Abban különbözik az előzőtől, hogy itt minden kiadott kulcsszó annyiszor számít, ahányszor előfordul.

Vektorműveletekkel leírva Tanimoto módosított képletét azt kapjuk, hogy

$$h/\underline{a}, \underline{b}/ = \frac{\sum_{i=1}^n a_i b_i}{\sum_{i=1}^n a_i + \sum_{i=1}^n b_i}$$

Igen elterjedt az ún. cosinus módszer klaszterek meghatározására /9.sz.formula./ A képlettel a két vektor által bezárt szög cosinusát számítjuk ki. Ha a két vektor hajlásszöge 0° , cosinusuk 1, s ha merőlegesek egymásra 90° -os hajlásszögeük/, cosinusuk 0.

Vektorműveletekkel kifejezve a szóban forgó összefüggést

$$h/\underline{a}, \underline{b}/ = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2}}$$

Az előző példán illusztrálva az elmondottakat, legyen

$$\begin{aligned} \underline{a} &= /0100000001/ \\ \underline{b} &= /0100000101/ \\ \hline \underline{a}_i \cdot \underline{b}_i &= 0100000001 \\ \sum_{i=1}^{10} a_i b_i &= 2 \end{aligned}$$

$$\sum_{i=1}^{10} /a_i/^2 = /1^2 + 1^2/ = 2 \quad \sum_{i=1}^{10} /b_i/^2 = /1^2 + 1^2 + 1^2/ = 3$$

Tehát

$$h/a, b/ = \frac{\sum_{i=1}^{10} a_i b_i}{\sum_{i=1}^{10} a_i^2 \cdot \sum_{i=1}^{10} b_i^2} = \frac{2}{\sqrt{2 \cdot 3}} = \frac{2}{\sqrt{6}}$$

A két ismerttetett eljárásról a szakirodalom azt tartja, hogy információkeresésnél Tanimoto nagyobb pontossággal jár együtt azonos teljességi mutató esetén, viszont nagyobb a klaszterálási műveletek száma. A cosinus módszernél viszont kisebb a veszteség. Tanimoto módszer alapján több klasztert kapunk, a klaszterek kizáróak

/diszjunktak, nem tartalmazznak átfedést/, a cosinus módszernél kevesebb, de elmosódóbb klaszterek keletkeznek, amelyekbe kisebb hasonlóság alapján is beke-
rülhetnek dokumentumok.

A táblázat 7. sz. képlete átfedésses klaszterek elő-
állítására alkalmas, míg a 8.sz. képlet az un. aszim-
metrikus hasonlóság számítására. Utóbbinak az automa-
tikus tezaurusz építésben van jelentősége, amikor a
szóban forgó vektorok deszkriptorok vagy kulcsszavak
tulajdonságainak sorozatából állnak, tehát egy ismerv-
ismerv mátrixból származnak, vagy - ami ugyanaz - egy
kulcsszó-kulcsszó, deszkriptor-deszkriptor mátrixból.
Az aszimmetrikus hasonlósági függvény segítségével
osztályozási kifejezések között "fölötte" "alatta"
kapcsolat, tehát hierarchikus viszony határozható meg.

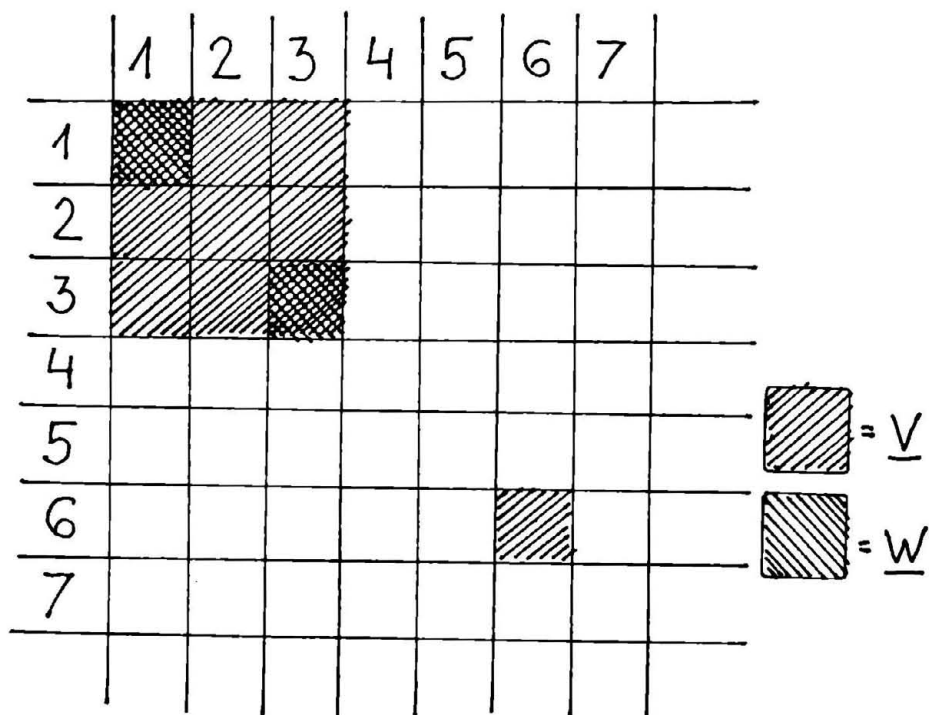
Legyen két ismervvektor $\underline{v} = /1110010/$ és $\underline{w} = /1010001/$.
Az aszimmetrikus hasonlósági függvény kiszámított érté-
kei:

$$h_{\underline{v}, \underline{w}} = \frac{N/\underline{v} \cap \underline{w}/}{N\underline{v}} = \frac{2}{4} \quad \text{és}$$

$$h_{\underline{w}, \underline{v}} = \frac{N/\underline{v} \cap \underline{w} /}{N\underline{w}} = \frac{2}{3}$$

Mivel $h_{\underline{w}, \underline{v}}$ értéke nagyobb, mint $h_{\underline{v}, \underline{w}}$, ezért \underline{w} átfo-
góbb, generikusabb kulcsszó, mint \underline{v} által képviselt

ismérv. Ez az eredmény egybevág azzal, amit a formális logika tanít a fogalmak tartalmának és terjedelmének viszonyáról. w kulcsszava kevesebb tartalmi jeggyel rendelkezik, mint v. Természetesen nem véletlen, hogy csak "alatta" vagy szűkebb, és "fölötte" vagy tágabb viszony határozható meg köztük, s nem a szigorúbb alá-fölérendelés, mivel mindkét kulcsszónak létezik olyan tulajdonsága is, amely a másiknak nincs. Ábrázolva a két kulcsszó által képviselt fogalom terjedelmét, pontosan az alábbi ábrát kapjuk. /Az ábrán függőlegesen is és vízszintesen is a két komponens látható, azaz a fogalom tartalmi jegyei. A bevonalkázott területek jelentik a fogalmak terjedelmét./



5. ÁBRA

A képletet éppen fordítva alkalmazzuk, ha az ismervvektorok egy dokumentum-ismerv matrixból származnak. Itt a komponensek /mint fentebb láttuk/ azt jelölik, hogy mennyi dokumentum és mely dokumentumok kapták a vizsgált ismérveket. Generikus, tágabb ismervhez több dokumentum tartozik, mint a specifikusakhoz. /V.ö. a kulcsszavak gyakorisági intervallumáról mondottakkal, amely szintén megerősíti ezt a tételt./ Ennek folytán éppen a generikus ismerv vektorában találunk több komponenst.

A számítás tehát azt eredményezi, hogy ha

$$h_{v,w} > h_{w,v}$$

akkor v a specifikusabb kulcsszó, az előző esetnek fordítottja.

4. Klaszterek kialakítása

Ha a dokumentumok között páronként meghatároztuk a hasonlóság mértékét, akkor ezt felírhatjuk táblázat formájában. Egy dokumentum-dokumentum mátrixot nyerünk, amelynek soraiban és oszlopaiban is ugyanazok a dokumentumok vannak, a sorok és oszlopok találkozásánál pedig a hasonlósági együttható található.

Példánkban szerepeljen 5 dokumentum és legyenek a hasonlósági értékek az alábbiak

	1.	2.	3.	4.	5.
1.	-	2/3	1/5	0	2/3
2.		-	1/6	1/5	2/4
3.			-	2/5	2/5
4.				-	0
5.					-

Ha bevezetünk egy küszöbértéket - legyen ez 0,4 -, akkor a mátrix egyszerűsíthető oly módon, hogy csak azt kell vizsgálni, elér-e a hasonlósági együttható legalább ezt a határt. Ha ~~eléri~~, vagy meg is haladja, akkor csak azt kell jelölni, hogy a küszöbérték fölötti hasonlóság fennáll, vagy nem áll fenn. Így az alábbi mátrix nyerhető:

	1.	2.	3.	4.	5.
1.		1			1
2.					1
3.				1	1
4.					
5.					

A mátrix többi eleme zérus. A valóságban természetesen igen nagy mátrixok adódnak, amelyek annyi sorból, ill. oszlopból állnak, ahány dokumentum szerepel a feldolgozásban.

A klaszterek meghatározása a mátrixok segítségével megy végbe. A probléma világos: ennek az ún. hasonlósági mátrixnak kell meghatározni azokat a részeit, amelyek az összetartozó dokumentumok csoportjait, klasztereit alkotják. A kérdés tehát az, hogyan lehet ezt a mátrixot a kívánt módon részeire szedni.

A probléma a mátrixalgebra ismert problémájához, az ún. faktorizációhoz vezet, a megoldást tehát a faktoranalízis nyújtja. Ennek értelmében a hasonlósági mátrixot szorzat alakban kell előállítani, ahol a szorzat tényezői, faktoraik egyszerűbb felépítésű mátrixok. A szakirodalomban több megoldás is ismeretes klaszterek meghatározására, amelyeket jelen helyen nem ismertetünk.

A kialakult klaszterekről azonban meg kell azt is határozni, miben áll az a hasonlóság, mi benne az a közös, ami összetartja. G. S. Altman szavaival, meg kell határozni a "gravitációs központját". A klaszter jellemzésére kívánatos meghatározni az ún. centroid vektort. A centroidnak is több értelmezése, definíciója lehetséges. Legegyszerűbben úgy definiálható, mint valamilyen "átlag" vektor, s úgy számítjuk ki, hogy összeadjuk valamennyi a klaszterbe tartozó dokumentumvektorok valamennyi komponensét, s ezt elosztjuk a vektorok /klaszterbe került dokumentumok/ számával:

$$c_i = \frac{1}{k} \sum_{p=1}^K a_{ip}$$

ahol c a centroid vektort jelöli, k a klaszterhez tartozó dokumentumvektorok száma. Ennek un. normalizált alakja a $\frac{c_i}{|c_i|}$

hányados, ahol $|c_i|$ a centroid vektor előzőekben kapott alakjának hossza, abszolút értéke

$$|c| = \sqrt{\sum_{i=1}^n c_i^2}$$

Értelmezzük a centroidot úgy, mint a vektorok komponenseinek összegéből képzett hányadosokból álló vektort:

$$c_i = \sum_{p=1}^K \frac{a_{ip}}{|a_i|}$$

Vegyük egy példát. Legyen három vektor,

$$\underline{v} = 1110001$$

$$\underline{w} = 1010110$$

$$\underline{z} = 1100010$$

Ha a centroidot "átlagnak" értelmezzük, akkor legegyszerűbben úgy kapjuk, ha képezzük a komponensenkénti átlagot. A centroid első komponense $3/3$ lesz, mert az első pozícióban álló komponensek összege 3, ezt osztjuk

a vektorok számával. A második komponens $2/3$ lesz, s.i.t. Azaz

$$\underline{c} = /1, 2/3, 2/3, 0, 1/3, 2/3, 1/3/$$

A centroidnak igen nagy fontossága van a gyakorlati klaszterálás során. Mindenekelőtt, ha egy új dokumentum érkezik, s be kell iktatni vektorát, akkor nem szükséges minden dokumentum vektorával összehasonlítani, elég ha a klasztereket képviselő centroidokkal megvégezzük az egybevetést, s így a műveletek száma lényegesen lecsökken. És ez nem lebecsülendő előny a nagy számítási igényű eljárásoknál. Másodszor a centroidnak igen nagy a jelentősége a visszakeresési eljárások során. A keresőprofil /keresőkép/ vektorát nem kell valamennyi dokumentumvektorral egybevetni, hanem elegendő a centroidokkal elvégezni ezt. Azaz, meghatározzuk előbb azokat a klasztereket, amelyekben a releváns dokumentumok lehetnek, majd az így kiválasztott klaszterekben végzzük el az összehasonlítást dokumentumról dokumentumra. Ismét igen jelentős számú lépés takarítható meg.

A centroid igen lényeges tulajdonsága, hogy változik. Ha egy új dokumentum egy klaszterbe besorolásra kerül, akkor a centroidot újra kell számolni, s kicsit elmozdulhat, mint ahogyan egy statisztikai sokaság átlaga is elmozdulhat, ha a sokasághoz új elem kerül. Ennek a ténynek felbecsülhetetlen szerepe van a d i n a m i k u s

k ö n y v t á r m o d e l l j é b e n, azaz az olyan könyvtár esetén, amely képes folyamatosan követni a változásokat és nem megmerevedett, statikus feltérési módszereket alkalmaz. Erre a modellre később néhány mondat erejéig visszatérünk.

A klaszterálási f o l y a m a t jobb megértéséért tisztázni kell azt is, hogyan indul a klaszterálás, s hogy vajon az indulás befolyásolja-e a klaszterek kialakítását. Tisztáztuk már, hogy ez a feldolgozás a dokumentumvektorok összehasonlításából áll, illetve ha már vannak kialakult klaszterek, akkor a probléma egy új dokumentum beiktatásánál az, hogy megtaláljuk azt a klasztert, amelybe besorolható. De hogyan indul a feldolgozási folyamat? Mihez hasonlítjuk az első dokumentum vektorát, vagy az első dokumentumok vektorait? Látni fogjuk, hogy az indulás befolyásolja a klaszterek kialakulását is. Több indulási lehetőség közül lehet választani.

a/ Találomra kiválasztjuk az első dokumentumot. Vektorát úgy kezeljük, mintha egy klaszter centroidja lenne. A másodiknak választott dokumentum vektorát hasonlítjuk hozzá. Ha a hasonlósági együttható a küszöbérték felett van, besoroljuk a klaszterbe, s kiszámítjuk az új centroidot. Ha a hasonlóság a küszöbérték

alatt marad, ennek a dokumentumnak a vektorát egy másik, új klaszter centroidjának tekintjük. Az eljárást ezen az úton folytatjuk.

b/ A dokumentumok közül előzetes vizsgálattal kiválasztjuk azokat, amelyeket "tipikus" tartalmúaknak tekintünk. Vektoraikat a lehetséges klaszterek centroidjainak tekintjük. Majd a dokumentumokat rendre összehasonlítjuk ezekkel az előre meghatározott centroidokkal, s besoroljuk őket a megfelelő klaszterekbe. Közben minden klaszter centroidját újra számítjuk az új tételek beiktatásának megfelelően. Ha egy dokumentum vektora egyik klaszterhez sem mutat hasonlóságot, akkor ezt új klaszter centroidjának tekintjük.

c/ A b/ alatti változat azzal a különbséggel, hogy nem tipikus dokumentumokat választunk, hanem meglévő osztályozási tapasztalataink alapján határozzuk meg tipikus vektorokat, mintha azok tipikus dokumentumok vektorai lennének. Az eljárás a továbbiakban az előzőekhez hasonlóan megy végbe.

d/ Működő információs rendszerek kérdéseit, keresőprofilokat ugyanúgy lehet csoportosítani, klaszterálni, mint a dokumentumokat. A k é r d é s - k l a s z t e r e k objektíven tükrözik a mindenkori felhasználói igényeket, a kérdésklaszterek centroidjai a tipikus

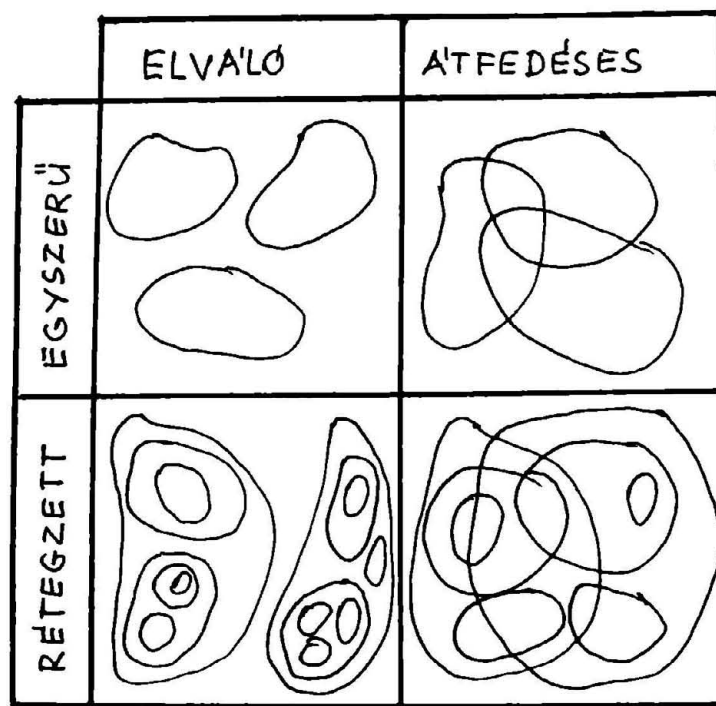
felhasználói kívánságokat. Mivel az új kérdések felmerülésekor a klaszterek és centroidjaik is változnak, a kérdésklaszterek hűen követik az igények változásait, s így naprakész igényszerkezeteket tudunk előállítani. Nos, a kérdésklaszterek centroidjai lehetnek a dokumentumok klaszterbe sorolásának centroidjai, azaz a dokumentumok vektorait az igényeket tükröző vektorokhoz klaszteráljuk. Így egy, a mindenkori igényekhez igazodó csoportosítást kapunk. Ha megfontoljuk, hogy mind az igényklaszterek, mind a d o k u m e n t u m k l a s z t e r e k változnak, új vektor belépésekor már annak sajátosságait is érvényesítik, akkor egy folyton változó, az igényekhez naprakészen igazodó megoldást találtunk. Ez a dinamikus könyvtár lényege. Kidolgozója G.Salton.
/21, 22/

5. Klaszterek típusai

A hasonlóság meghatározására szolgáló függvény és az input paraméterek megválasztása a klaszterek számos fajtáját eredményezi. Ha valamely dokumentum csak egy klaszterbe tartozhat, azaz a klaszterek diszjunkt csoportokat eredményeznek, akkor e l v á l ó, diszkrét diszjunkt klaszterekről van szó. Ha a dokumentumok egyszerre több klaszterhez is tartozhatnak, akkor á t f e d é s e s e k.

Másfelől a klasztereket is lehet klaszterálni, kisebb csoportokat nagyobbakká lehet összefoglalni /alacsonyabb hasonlósági küszöbértékeket bevezetve/. Ekkor van szó r é t e g z e t t, ellenkező esetben pedig e g y s z e r ű klaszterekről.

Az alaptípusokat mutatja a következő ábra.



6. ÁBRA

Hierarchikus a klaszterálás akkor, ha elváló és rétegzett. Ha a klasztereket táblázatos formában számítógéppel kiíratjuk úgy, hogy a függőleges tengelyen a hasonlóság mértéke, a vízszintesen a dokumentumok kerülnek ábrázolásra /a hasonlósági mérték csökkenő sorrendjében/, akkor d e n d o g r a m r ó l beszélünk.

6. Az alkalmazás területei

Történetileg a klaszterálás a nagy rendszertani hagyományokkal rendelkező tudományokban jelent meg, így a biológiában, ahol a taxonómia, a rendszerezés diszciplinája önálló ismeretág. Innen vette át az orvostudomány, majd rohamosan elterjedt más területeken: műszaki tudományokban, lélektanban, szociológiában, - általában azokban az ismeretágakban, ahol nagytömegű adatot, objektumokat kívánatos csoportosítani, vagy rendszerezni. Sikerét annak köszönhetette, hogy az egyéni véleményekkel szemben "objektív" - az idézőjeleket azért kell kitenni, mert ez az objektivitás addig terjed, ameddig a matematikai módszerek objektivitása terjed -, tovább, hogy automatizálható.

/Az input paraméterek megválasztásában "szubjektív" szempontok is érvényesülhetnek./ Az elektronikus számítógépek megjelenése előtt gyakorlati klaszterálás nem volt elképzelhető.

Az információtárolás és visszakeresés nagy rendszerei hamar felfedezték az ebben rejlő lehetőségeket, s a hatvanas években már se szeri, se száma a közleményeknek. A felfedezés joga szakmánkban - úgy hiszem - G.Saltont és munkatársait illeti, akik a SMART "mágikus" rendszerükben mindent automatizálni kívántak, az eddig legintellektuálisabbnak tartott tevékenységeket is.

Ma nincs figyelemre érdemes információs tevékenységet folytató ország, amely ne folytatna kísérleteket automatikus osztályozással. A szocialista országok közül jelentős sikereket könyvelhet el a Szovjetunió, Pozsonyban Marek Ciganik kísérletei figyelemre méltóak. A magyarországi tájékoztatásügy elemibb kérdésekkel foglalkozik.

Ennek az is oka lehet, hogy ez a technika nagy számítógépeket tételez fel, igen nagy a tárolási igénye. De ennél is lényegesebb, hogy e technika bevezetése előtt még számos kérdés vár megoldásra.

A tájékoztatásügy ezt a technikát legalább három területen alkalmazhatja:

- a/ Osztályozási rendszerek kimunkálásában. Ennek során osztályozási kifejezések, tulajdonságok klaszterálása oldhat meg számos kérdést, olyanokat is, amelyeket a dolgozat alig érintett.
- b/ Dokumentumok osztályozásában. Jelen tanulmány lényegében erre az esetre korlátozódott.
- c/ Visszakeresési stratégiák kiépítésében.

Az utóbbi két területen a klaszterálási eredmények nemcsak elérik, hanem számos vonatkozásban meghaladják azt a határfokot, amelyet a legjobb hagyományos megoldásokkal el lehet érni.

Le kell azonban azt is szögezni, hogy a klaszter-
technika még n e m a m a m ó d s z e r e. A jövőé.
Ezt a jövőt azonban a mában kovácsolják, a mai kuta-
tás a holnap technikája.

I r o d a l o m

1. Automated Language Processing. Ed.by H.Borko.
New-York, London, Sidney, 1968.
2. Babiczky Béla: Bevezetés a könyvtári osztályozás elméletébe és gyakorlatába.
Jegyzet.Bp.Tankönyvkiadó, 1975.
3. Borko H. - Bernick M.D.: Automatic Document Classification. Technical memorandum.
Santa Monica /California/, System Dev.Corp. 1967.
4. Brofitt J.D. - Morgavan H.L. - Soden J.V.: On some clustering techniques for information retrieval.
Information storage and retrieval to the National Science Foundation. Scientific report no. ISR-11.
Cornell Univ.,Ithaca /New-York/, 1966.IX.1-15.1.
5. Cherry C.: On human communication.
Cambridge /Massachusetts/, MIT Press, 1959.
6. Ciganik M.: Informacné systémy vo vede, technike a ekonomike.
Martin, Matica Slovenska, 1969.
7. Fritsche M.: Automatic clustering technique in information retrieval. Commision of the European Communities, Joint Nuclear Research Centre - ISPRA Estalishment /Italy/, Scientific Data Processing Centre - CETIS. Luxembourg, 1974.

8. Garfield E.: Social Science Citation Index clusters.
= Current Comments, 1976, no.27.
9. Hell György: A számítógépes szövegelemzés magyarországi eredményei és trendjei. Kézirat.Bp. 1975.
10. Horváth Tibor: Automatikus osztályozás.
= Könyvtári Figyelő, 1978.évf.5.sz. 528-542.l.
11. Horváth Tibor: A bibliográfiák funkciójáról.
= Bibliográfiai tanulmányok. 11-57.l. Bp. Könyvtártudományi és Módszertani Központ, 1978.
12. Jardine N. - Van Risbergen C.J.: The use of hierarchic clustering in information retrieval.
= Inf. Storage and Retr. 1971. no.5. 217-240.l.
13. Klauszer Judit: A magyar főnevek szintézisének kérdéséhez.
= Ált.nyelvészeti tanulmányok 3.köt. 117-129. l. Bp. Akadémiai Kiadó.
14. Kónyi Sándor: A magyar főnevek elemzése.
= Ált.nyelvészeti tanulmányok 3.köt. 131-143.l. Bp.Akadémiai Kiadó.
15. Luhn H.P.: Automatic creation of literature abstracts.
= IBM Journal of Res. Develop., 2.köt. 1958.2.sz.
16. Meadow Ch.T.: The analysis of information systems. New-York, London, Sidney, John Willey, 1967.

17. Needham R.M.:— Sparck Jones K.: Keywords and clumps.
= Journal of Doc. vol.20.1964. no.1. 5-15.1.
18. Nesitoj V.V.: Raszpredelenie klucsevih szlov v tekszte.
= Kibernetika, 1977. 2.sz. 123-131.1.
19. Párnczky Gábor: A statisztikai informatika alapjai.
Bp.Statisztikai Kiadó, 1967. 135-165.1.
20. Salton G.: Automatic information organisation and retrieval.
New-York, St.Louis, San Francisco, stb. 1968.
21. Salton G.: Proposals for a dynamic library.
Cornell Univ., Dep.of Computer Science, Ithaca, N.Y., 1972.
22. Salton G.: Dynamic information and library processing.
Englewood Cliffs, New-York, Prentice Inc. 1975.
23. Salton G.: Search strategy and optimization of retrieval effectiveness.
= Mechanized information storage, retrieval and dissemination, Proc. of the FID/IFIP Joint Conf. Rome, 1967. Amsterdam, North-Holland Publ. 1968.
24. The SMART System. Experiments in automatic document processing. Ed. G. Salton. Englewood Cliffs, Prentice Hall, 1971.

25. Soergel D.: Mathematical analysis of documenta-
tion systems.
= Inf. Storage and Retr.vol. 3. 1967.no.3.
129-173.1.
26. Sparck Johnes K. - Kay M.: Linguistics and infor-
mation science. New-York, London, Acad.Press,1973.
27. Swanson R.W.: On clustering technique in informa-
tion retrieval.
= Journal of ASIS.24. vol. 1973. no.1. 72-73.1.
28. Zsitkov G.N.: O klasszifikacii sztrukturnüh
elementov pri analize szlovnüh izobrazsenii.
= Naucsno-Tehnicoseszkaja Inf.Szer.2.,1970.
no. 10. 14-18. 1.

