AUTOMATIKUS OSZTÁLYOZÁS

HORVÁTH Tibor

INFORMÁCIÓS MODELL

Automatikusnak nevezünk egy osztályozást, ha a) az osztályozási kifejezések a közlemények eredeti szövegéből (természetes nyelvből) automatikusan kerülnek meghatározásra, b) ha az így nyert kifejezésekkel az összes további művelet (csoportosítás, rendezés, kapcsolatok meghatározása stb) automatikusan megy végbe.

Jelen tanulmány a b) problémakörrel kíván foglalkozni. Az első – a) pontban jelzett – kérdés az alkalmazott nyelvészet és informatika határán kialakult kutatási terület, az osztályozási kifejezések, a kulcsszavak meghatározására kialakított eljárások a természetes nyelv statisztikai vizsgálatán nyugszanak. A továbbiakban feltesszük tehát, hogy a kulcsszavak meghatározása valamilyen módon végbement már – pl. gyakorisági vizsgálatokkal, vagy különböző szótárak segítségével – és minden egyes dokumentumhoz a kulcsszavak egy sorozatát rendeltük hozzá.

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

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 összességét, amelyet egy 0 és 1 jelekből álló jelsorozat, ún. vektor reprezentál. Minden oszlop pedig egy jellemző jegy, ismérv 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.

A KLASSZIKUS LOGIKA PROBLÉMÁJA

Az „osztályozás" lényegesen szélesebb körű intellektuális művelet annál, hogy leszűkíthetnénk a dokumentumok osztályozásának problémakörére. J. P i a g e t 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. 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. A tekintetben, hogy melyek a „lényeges" sajátosságok, tudományos iskolák, irányzatok csaptak össze (de ezek a kérdések már kívül estek a logika hatáskörén). Az ezt a fajta logikát érvényesítő tudományos rendszerezések más-más szempontokat érvényesítettek. Pl. D a r w i n rendszerezése az élővilágról a fajok eredetén alapult, míg L i n n é 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 járt 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 v a l a m e n n y i t u l a j d o n s á g á t (mivel e sajátosságok száma végtelen – mint tisztáztuk fentebb – megelégszünk azzal, ha e l e g e n d ő e n n a g y s z á m ú 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.

A TÁVOLSÁG MEGHATÁROZÁSA

Alkalmasnak látszik erre a célra a dokumentumvektorok közti ún. „távolság", illetve ún. „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.

Még egyszer tudatosítanunk kell, hogy a dokumentumvektorok elemei komponensei egy tulajdonsághalmaz (osztályozási kifejezések) elemei, s a tulajdonságok egy sorozata – esetünkben bináris értékeket véve fel, – alkotja a vektort.

A távolság meghatározására több módszer létezik. Az ún. Hamming-féle távolság (R. W. Hamming matematikusról, az információelmélet kiválóságáról elnevezve) az egybe nem eső, a különböző komponensek számát veszi figyelembe. Ha a = 000 és 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 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 közelség azokat a pozíciókat veszi figyelembe, ahol mindkét vektor azonos pozíciójában nem zérus elem áll. Az előzőekkel ellentétben itt éppen az egyező komponensek számát kívánjuk meghatározni, nem a különbség, hanem az azonosság mértékét. Más szóval: azt kívánjuk meghatározni: hányszor 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 hányszor 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 6. dokumentumán.

$\underline{a}$ (2. dok.) = (0 1 0 0 0 1 0 1 1 0)
$\underline{b}$ (6. dok.) = (0 1 0 0 0 0 0 1 0 1)

$\underline{a}_i\underline{b}_i$ = (0 1 0 0 0 0 0 1 0 0) (komponensenkénti szorzat)

$$\sum_{i=1}^{10} \underline{a}_i\underline{b}_i = 2 \text{ (komponensek összege)}$$

Ennek a módszernek két gyengéje van. Az első, hogy két dokumentum abban is hasonló lehet, hogy mely kulcsszavakat n e m kapták meg e g y s z e r r e (ezt a fenti érték nem jelzi). Valami „egyikre sem jellemző" tulajdonság ez. A másik probléma az, hogy a közös tulajdonságok számát kifejező függvényértékét abszolut számban kaptuk meg, ezért bizonytalanul tudunk ítéletet alkotni a hasonlóság mértékéről. Jobban szeretjük az olyan mérőszámokat, amelyeknél ismerjük a felső és alsó határokat, azokat az értékeket, amelyeket a h függvény egyáltalán felvehet. Az első probléma megoldását röviden vázoljuk, a másik probléma pedig azokhoz az eljárásokhez vezet, amelyekkel már gyakorlati rendszerekben 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 összeszámláljuk. Ez a művelet úgy hajtható végre, hogy képezzük a vektorok komplementereit*, 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 $\bar{\underline{a}}$ és $\bar{\underline{b}}$ a komplementer vektorok.

Az előbbi példa ezzel a képlettel:

$\underline{a}$ = (0 1 0 0 0 1 0 1 1 0) $\bar{\underline{a}}$ = (1 0 1 1 1 0 1 0 0 1)
$\underline{b}$ = (0 1 0 0 0 0 0 1 0 1) $\bar{\underline{b}}$ = (1 0 1 1 1 1 1 0 1 0)

$a_i \cdot b_i$ = (0 1 0 0 0 0 0 1 0 0) $\bar{a}_i \cdot \bar{b}_i$ = (1 0 1 1 1 0 1 0 0 0)

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

* 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 alábbit kell érteni. Legyen $\underline{a}$ és $\underline{b}$ két vektor, akkor euklideszi távolságukon a

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

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

G. N. Z s i t k o v összefoglaló tanulmányában a „távolság" kiszámítására hat különböző formulát ismertet. Az olvasónak további részletek megismerésére ezt a tanulmányt javasoljuk.

HASONLÓSÁGI FÜGGVÉNYEK

A fentebb ismertetett távolsági mérőszámok valamilyen módon tükrözik a dokumentumok közti hasonlóságot, tartalmi rokonságot. Nagy hátrányuk azonban, hogy az abszolut 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 k ü s z ö b é r t é k e t is, amely fölött a 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 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.

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

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

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

ahol $c = \min(a_i, a_j)$

$$\underline{a}_i = (111000)$$
$$\underline{a}_j = (101010)$$
$$\underline{c} = (101000)$$

b) $\underline{a}_i \cap \underline{a}_j = \underline{c}$,
ahol $\underline{c} = \underline{a}_i \underline{b}_j$, tehát a két vektor komponensenkénti szorzata.
Példa:

$$\underline{a}_i = (111000)$$
$$\underline{a}_j = (101010)$$
$$\underline{c} = (101000)$$

(Bináris vektorokról lévén szó, a két eredmény azonos)
Két vektor összegén, unióján, egyesítésén az a $\underline{c}$ vektor értendő, amely
a) $\underline{a}_i \cup \underline{a}_j = \underline{c}$
ahol $\underline{c}_i = \max(\underline{a}_i, \underline{b}_j)$, azaz $\underline{c}$ a két vektor komponensei közül mindig a nagyobbikból áll.

$$\underline{a}_i = (111000)$$
$$\underline{a}_j = (101010)$$
$$\underline{c} = (111010)$$

b) $\underline{a}_i \cup \underline{a}_j = \underline{c}$
ahol $\underline{c} = \underline{a}_i + \underline{b}_j$

$$\underline{a}_i = (111000)$$
$$\underline{a}_j = (101010)$$
$$\underline{c} = 212010$$

Más szóval $\underline{c}$ vektor az azonos pozícióban álló komponensek összegéből áll.
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.
Ezek előrebocsátásával a dokumentumvektorok hasonlóságát kifejező különböző függvényeket az alábbiakban foglaljuk össze. A közölt áttekintés M. Fritschetől származik.
A képletekben az N a komponensek összegezésének jele.
Legyen $\underline{a}$ és $\underline{b}$ két összehasonlításra szánt, egységesen n komponensből álló bináris vektor.

Hasonlósági függvények	Intervallum
1. $h(\underline{a}, \underline{b}) = N(a \cap b)$	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

Hasonlósági függvények (folyt.)	Intervallum
4. $h(\underline{a}, \underline{b}) = [\frac{1}{n} nN(\underline{a} \cap \underline{b}) + nN(\overline{\underline{a}} \cap \underline{b})]$	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 $h(\underline{a}, \underline{b}) = \frac{N(a \cap b)}{N(a)}$	0, 1
9. $\cos h(\underline{a}, \underline{b}) = \frac{N(\underline{a} \cap \underline{b})}{\sqrt{N(\underline{a} \cap \underline{a})\, N(\underline{b} \cap \underline{b})}}$	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

Látható, hogy e képletek legtöbbjének alapgondolata az, hogy az összehasonlítandó vektorok (tulajdonságok) közös részét fejezik ki az összes tulajdonság hányadában. Lefordítva az osztályozás problémakörére a fentieket, 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 veszik, 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 mondottunk. 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):

$$\begin{aligned} \underline{a} &= (0100000001) \\ \underline{b} &= (0100000101) \end{aligned}$$

$$\begin{aligned} \underline{a} \cap \underline{b} &= (0100000001) \\ N(\underline{a} \cap \underline{b}) &= 2 \\ \underline{a} \cup \underline{b} &= (0100000101) \\ N(\underline{a} \cup \underline{b}) &= 3 \end{aligned}$$

Ennek alapján

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

Más szóval: három kiadott ismérv, kulcsszó közül kettő közös. Tanimoto módosított formulája (6. számú képlet) alapján

$$N(\underline{a}) = 2, N(\underline{b}) = 3, \quad N(\underline{a}) + N(\underline{b}) = 5, -$$

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. számú 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 hajlásszögűek), cosinusuk 0.

Vektorműveletekkel kifejezve a szóbanforgó 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

$$\underline{a} = (0\,1\,0\,0\,0\,0\,0\,0\,0\,1)$$
$$\underline{b} = (0\,1\,0\,0\,0\,0\,0\,1\,0\,1)$$

$$\underline{a}_i \cdot \underline{b}_i = 0\,1\,0\,0\,0\,0\,0\,0\,0\,1$$

$$\sum_{i=1}^{10} a_i b_i = 2$$

$$\sum_{i=1}^{10} (a_i)^2 = (1^2 + 1^2) = 2 \qquad \sum_{1}^{10} (b_i)^2 = (1^2 + 1^2 + 1^2) = 3$$

Tehát

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

A két ismertetett eljárásról a szakirodalom azt tartja, hogy információcserénél Tanimoto nagyobb pontossággal jár együtt azonos teljességi mutató esetén, viszont nagyobb a klaszterálási műveletek száma. A cosinus módszernél viszont kisebb a veszteség. Tanimoto módszere alapján több klasztert kapunk, a klaszterek kizáróak (díszjunktak, nem tartalmaznak átfedést), a cosinus módszernél kevesebb, de elmosódóbb klaszterek keletkeznek, amelyekbe kisebb hasonlóság alapján is bekerülhetnek dokumentumok.

A táblázat 7. számú képlete átfedéses klaszterek előállítására alkalmas, míg a 8. számú képlet az ún. aszimmetrikus hasonlóság számítására. Utóbbinak az automatikus tezaurusz építésben van jelentősége, amikor a szóbanforgó vektorok deszkriptorok vagy kulcsszavak tulajdonságainak sorozatából állanak, tehát egy ismérv-ismérv 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.

KLASZTEREK KIALAKITÁSA

Ha a dokumentumok között páronként meghatároztuk a hasonlóság mértékét, akkor ezt táblázat formájában felírhatjuk. 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 alábbak

	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, eléri-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. Igy 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, illetve oszlopból állanak, ahány dokumentum szerepel a feldolgozásban.

A klaszterek meghatározása a mátrixoknak 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, faktorai 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. Salton szavaival, meg kell határozni a „gravitációs központját". A klaszter jellemzésére kívánatos meghatározni az ún. centroid vektort. A centroidnek 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 elsoztjuk a vektorok (klaszterbe került dokumentumok) számával:

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

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

hányados, ahol $|\underline{c}_i|$ a centroid vektor előzőekben kapott alakjának hossza, abszolut értéke

$$(|\underline{c}| = \sqrt{\sum_{i=1}^{n} c_i^2}).$$

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

$$\underline{c}_i = \sum_{p=1}^{K} \frac{a_{ip}}{|a_i|}$$

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 megy végbe az egybevetés, 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 végezzük el az összehasonlítást dokumentumról dokumentumra. Ismét igen jelentős számú lépés takarítható meg.

A centroid igen lényeges tulajdonsága, hogy v á l t o z i k . 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 feltárási 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 centoidja 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 kérdés-klaszterek 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. Igy egy, a mindenkori igényekhez igazodó csoportosítást kapunk. Ha megfontoljuk, hogy mind az igényklaszterek, mind a dokumentumklaszterek 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 professzor ennek a kérdésnek egyik legjelesebb szakértője.

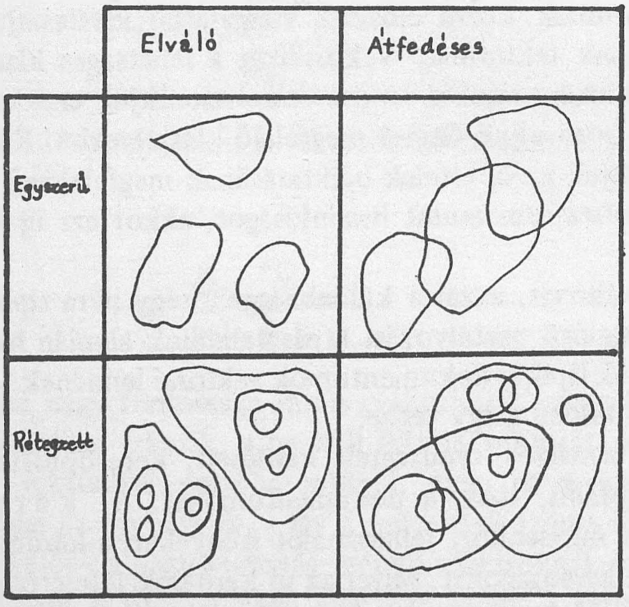
Az elmondottak alapján egy rendkívül rugalmas, magas követelményeknek megfelelő módszer bontakozik ki. Ennek a módszernek számtalan leágazása, igen magas szintű elmélete van, amelynek részleteire itt nincs mód kitérni.

KLASZTEREK TIPUSAI

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 elváló, diszkrét diszjunkt klaszterekről van szó. Ha a dokumentumok egyszerre több klaszterhez is tartozhatnak, akkor átfedésesek.

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 klaszterekről, ellenkező esetben pedig e g y s z e r ű klaszterekről.

Az alaptípusokat mutatja a következő ábra.



15. ábra

H i e r a r c h i k u s a klaszterálás akkor, ha elváló és rétegzett. Ha a klasztereket táblázatos formában számítógéppel kiiratjuk, pl. 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.

AZ ALKALMAZÁS TERÜLETEI

Történetileg a klaszterálás a nagy rendszertani hagyományokkal rendelkező tudományokban jelent meg, így a biológiában, ahol a taxonómia, a rendszerezés diszciplinája önállósult 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 köszönhette, hogy az egyéni véleményekkel szemben „objektív" – az idézőjelet 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. Salton 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 Cigánik kísérletei figyelemre méltóak. A magyarországi tájékoztatásügy elemibb kérdésekkel foglalkozik. Tudomásom szerint az MTA SZTAKI programokat dolgozott ki klaszterálási problémákra, de alkalmazásukra a tájékoztatásügy még nem merült fel.

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, pl. az, hogyan lehet automatikusan meghatározni azokat az ismérveket, amelyekkel a dokumentumokat osztályozzuk. Ez a problémakör az alkalmazott nyelvészet és informatika közös ügye lenne, ha nem szakadt volna meg a hatvanas évek biztató együttműködése a két tudomány művelői között. Pedig 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 jelen cikk nem is 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 hatásfokot, amelyet a legjobb hagyományos megoldásokkal el lehet érni.

Le kell azonban azt is szögezni, hogy a klasztertechnika nem a ma módszere. A jövőé. Ezt a jövőt azonban a mában kovácsolják, a mai kutatás a holnap technikája, módszere. Ami miatt bizonyára az automatikus osztályozás nem kerülhető el a jövőben, azt az a körülmény határozza meg, hogy egyre igénytelenebb osztályozók, egyre igényesebb osztályozási rendszerekkel, egyre gyengébben osztályoznak. Az információk feltártsága iránti kívánalmak nőnek, a szakirodalom gyarapszik, s a társadalom nem képes biztosítani kívánt mennyiségben a legmagasabban kvalifikált osztályozó szakembereket. A jövő számára nemcsak intézményrendszereket kell tervezni, hanem technológiákat is. A mai kísérletek talán tíz esztendő múlva hozhatnak gyakorlatilag hasznosítható eredményt. S miért ne lennének optimisták?

IRODALOM

1. 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.p.
2. CIGÁNIK, M.: Informacné systémy vo vede, technike a ekonomike. Martin, Matica slovenska, 1969. 528.p.
3. FRITSCHE, M.: Automatic clustering techniques in information retrieval. Commission of the European Communities, Joint Nuclear Research Centre – Ispra Establishment (Italy), Scientific Data Processing Centre – CETIS. Luxembourg, 1974.
4. GARFIELD, E.: Social Science Citation Index clusters. = Current Comments, 1976. 27.no.
5. JARDINE, N.–VAN RIJSBERGEN, C. J.: The use of hierarchic clustering in information retrieval. = Inf. Storage and Retr. 1971. 5.no. 217–240.p.
6. NEEDHAM, R. M.–SPARCK, Jones K.: Keywords and clumps. = Journal of Doc. 20.vol. 1964. 1.no. 5–15.p.
7. PÁRNICZKY Gábor: A statisztikai informatika alapjai. Bp. Statisztikai K. 1976. 135–165.p. A korszerű informatika könyvtára 8.)
8. SALTON, G.: Automatic information organization and retrieval. New York–St. Louis–San Francisco stb. 1968. 514.p.
9. SALTON, G.: Dynamic information and library processing. Englewood Cliffs, New York. Prentice Hall Inc. 1975. 523.p.
10. SALTON, G.: Proposals for a dynamic library. Cornell Univ. Dep. of Domputer Science, Ithaca, N.Y. 1972. 62.p.
11. SALTON, G.: Search strategy and the optimization of retrieval effectivness. = Mechanized information storage, retrieval and dissemination. Proc. of the FID/IFIP Joint Conf. Rome 1967. Amsterdam, North-Holland Publ. 1968.
12. The SMART retrieval system. Experiments in automatic document processing. Ed. G. SALTON. Englewood Cliffs, Prentice Hall Inc. 1971. 556.p.
13. SOERGEL, D.: Mathematical analysis of documentation systems. = Inf. Stor. Retr. 3.vol. 1967. 3.no. 129–173.p.
14. SWANSON, R. W.: On clustering technique in information retrieval. = Journal of ASIS. 24.vol. 1973. 1.no. 72–73.p.
15. ZSITKOV, G. N.: O klaszszifikacii sztrukturnüh élementov pri analize szlozsnüh izobrazsenij. = Naucsno-Tehnicseszkaja Inf. Szer.2. 1970. 10.no. 14–18.p.