Összehasonlító klaszterjellemzés külső, szöveges források bevonásával

Kruzslicz Ferenc
PhD, a Pécsi Tudományegyetem docense
E-mail: kruzslic@ktk.pte.hu

Kovács Balázs,
a Pécsi Tudományegyetem tanársegédje
E-mail: kovacsb@ktk.pte.hu

Hornyák Miklós,
a Pécsi Tudományegyetem tanársegédje
E-mail: hornyakm@ktk.pte.hu

Klaszterezési módszereket használó kutatások során nagyon fontos, hogy a kapott klaszteknek lényegre törő elnevezést találjunk. Különösen fontos ez olyankor, amikor a klaszterezés nemcsak egy köztes módszer, hanem ez képezi az elemzés végeredményét. Ilyenkor a klaszter mögött meghúzódó fogalom maga az, amit hasznosítani szeretnénk. A klaszter azonban valójában a vizsgálódó elméjében jön létre azáltal, hogy azt szavakkal definiálni vagy legalábbis körülírni tudja. A szerzők módszere ezt a verbális klaszterjellemzési folyamatot kívánja megkönnyíteni és részben automatizálni. A klaszterek elemeihez a klaszterképző adatok között nem szereplő külső szöveges adatbázist csatolnak. A klaszterjellemzést a külső adatok szövegbányászati elemzésével végzik. Az előállított szófelhő milyensége a klaszterezés indokoltságának jellemzésére is felhasználható.

TÁRGYSZÓ:
Klaszterminőség.
Külső index.
Összehasonlító szófelhő.

DOI: 10.20311/stat2016.11-12.hu1123

A klaszterezési technikák – melyek segítségével az adatobjektumokat (egyedeket) tulajdonságaik hasonlósága alapján előre nem definiált (al)csoportokba soroljuk – már jóval a big data trendek előtt a figyelem középpontjába kerültek. Noha maga a fogalom a kulturális antropológia területén már nagyon korán megjelent (*Driver–Kroeber* [1932]), a feladat számítási igénye miatt a rutinszerű használatáról csak az 1960-as évektől kezdve beszélhetünk. A számítástechnika fejlődésének köszönhetően az alapvető sokváltozós matematikai statisztika algoritmusai már minden standard programcsomagban elérhetők. A klaszterezési módszerek alkalmazhatóságának azonban továbbra is jelentős korlátja maradt, hogy csak viszonylag kis elemszámú adathalmazon adnak belátható időn belül eredményt. Ráadásul az eredmény minőségének megítélése sem egyszerű, és függ a klaszterezés típusától is. A big data 5V-ként (volume – méret; variety – változatosság, sokoldalúság; velocity – gyorsaság; veracity – igazságérték, megbízhatóság; value – érték, fontosság) emlegetett jellemzői közül tehát az adatmennyiség és ezek keletkezési sebessége jelenti a legnagyobb kihívást a klaszterezés számára. Míg a nem numerikus adatok kezelése és a bizonytalan információk feldolgozása ma már nem akadály az ilyen algoritmusok számára. A klaszterezési feladattípusok letisztulásával nagyjából egy időben jelent meg a *Statisztikai Szemlében* az első két, klaszterelemzési módszerekkel foglalkozó cikk *Csicsman József* [1979] és *Futó Péter* [1979]. Ezek több szempontból is előremutatónak bizonyultak az évek során. Egyrészt rávilágítottak, hogy a klaszterezési feladatok a (hiper)gráfok kvázi komponenseinek megkeresésével ekvivalensek. Másrészt bemutatták, hogy módszerük alkalmas szöveges adatok klaszterezésére is. Az általuk definiált információtudományi klaszterelemzés valójában a szövegbányászatban használt dokumentumklaszterezési problémának feleltethető meg.

Az idő nemcsak a terminológiát változtatta meg, hanem a klaszterezés célját is. Akkoriban az elsődleges alkalmazási területek a tipizálás, a modellillesztés, a csoportokra alapozott becslés és a hipotézistesztelés voltak. Részben a big data által támasztott igényekre adott válaszként mára mindez kiegészült a nagy adathalmazok mintázatainak felderítésével, az adathalmaz tömörítésével és hipotézisek generálásával. Információ-visszakereső rendszerek esetében például hírek klaszterezésével nagy mennyiségű szöveges adat sűríthető össze szemantikailag úgy, hogy a felhasználói keresésekre adott válaszok gyorsabbak és pontosabbak lesznek. A címben szereplő módszerünk is alapvetően a dokumentumklaszterezési, azon belül pedig a címkézési technikákra épül.

A nagyméretű adathalmazokból mintákat kinyerő adatbányászati módszertanokban a klaszterezés elsősorban nem mint a modellezés végeredménye, hanem

mint adat-előkészítési és adattisztítási segédeszköz használatos. Ezek a kezdeti lépések kulcsfontosságúak az elemzés kimenetelére nézve, ezért a klaszterelemzés alkalmazása több lépésben, iteratív módon történik. Ahhoz, hogy egyre jobb adatminőséget lehessen elérni, össze kell tudni hasonlítani az egyes klaszterezési szerkezeteket egymással, hogy megtaláljuk a feladathoz illeszkedő legalkalmasabb változatot. A hazai szakirodalom első, klaszter kiértékelést is érintő cikkét *Füstös László* és szerzőtársai a *Szigmá*ban jelentették meg (*Füstös–Meszéna–S.-né Mosolygó* [1977]).

A klaszterérvényességet alapvetően kétféle módon állapíthatjuk meg. A kiértékelés (validálás) során, ha csak ugyanazokat az adatokat használhatjuk, mint amiket a klaszterezéskor is, akkor felügyelet nélküli érvényességvizsgálatról beszélünk. Viszont, ha lehetséges a kapott klaszterezést egy külső, már ismert szerkezethez (például létező osztályokhoz) hasonlítani, akkor az illeszkedés mértékét már felügyelt módszernek tekintjük. A bemutatni kívánt klaszterkiértékelési módszer e kettő között helyezkedik el. A kiértékelés során a klaszterezéshez nem használt, külső, szöveges információkat is fel fogunk használni, de ezek nem rendelkeznek semmiféle felügyelt módszernek tekinthető szerkezettel. A big data világában a klaszterezendő objektumokhoz általában nem nehéz szöveges jellemzőket kapcsolni aszerint, hogy azok kitől származnak, vagy mire vonatkoznak.

A klaszterezés megbízhatóságát nemcsak mutatószámokkal (indexekkel) lehet jellemezni, hanem egyéb módszerekkel is. Például, hogy egy tesztalgoritmus végrehajtható-e a klasztereken, vagy található-e olyan ábrázolási mód, amelyről leolvasható a klaszterek létezése. Bár végső soron minden alternatív klaszterkiértékelés mutatóvá konvertálható, az érvényesség egy számmá tömörítésével sok addicionális információtól elesünk. Ráadásul azt is nehéz megítélni, hogy egy így kapott mutatószámhoz mennyire megfelelő klaszterezés tartozik. Legfeljebb egy adott adathalmazon ugyanolyan típusú módszerrel kapott két klaszterezés közötti relatív választásban lehetnek segítségünkre. Részint ezért is tartja magát az a nézet, hogy a klaszterezés inkább művészet, mint tudomány. Az általunk javasolt klaszterjellemzési megoldás sem számokban fejezi ki a klaszterek minőségét, hanem egy új ábrázolási technikával, amiről lényegében az állapítható meg, hogy mennyire lehet pontosan szövegesen körülírni és megkülönböztetni az egyes klasztereket. Ez pedig nem más, mint az infografika elterjedésével népszerűvé vált szófelhődiagram egy speciális, *Drew Conway* ötlete alapján kialakított változata (*Conway* [2011]).

A tanulmány további részeiben először az eddigi klasztercímkézési megoldásokat, klaszterérvényességi módszereket és szófelhőváltozatokat tekintjük át, majd az ezek kombinációjaként kialakított szövegjellemzés-alapú összehasonlító módszerünket ismertetjük. A módszer használatát egy internetes adatforrást használó példán keresztül mutatjuk be.

1. Dokumentumklaszterezés

A dokumentumklaszterezés nem más, mint a hagyományos klaszterezési technikák alkalmazása szöveges állományokra. Ez anélkül is megoldható, hogy speciális, szöveges adatokon működő algoritmusokat fejlesztenénk ki, amennyiben a dokumentumokhoz sikerül megfelelő numerikus reprezentációt találnunk. Ezek közül a legegyszerűbb és általunk is alkalmazott módszer a vektortér-reprezentáció, melynek során egy **d** dokumentumot a lehetséges szavak terében értelmezett $\mathbf{d} = (x_1, x_2, \ldots, x_m)$ vektorként adunk meg. A vizsgált dokumentumhalmaz összességét a szakirodalom korpusznak nevezi. Egy dokumentumvektor x_i eleme a korpusz szavaiból alkotott szótár *i.* szavára vonatkozó mutató. A leggyakrabban használt szózsák modell (bag of words) esetén ezt a mutatót az adott szó gyakorisági értékének (term frequency) választjuk meg. Igaz, hogy ez információveszteséggel jár, hiszen elveszítjük a szósorrendet, de egy ilyen vektortéren azután már minden szokásos klaszterezési művelet elvégezhető. A számítások eredményének értelmezése adhat némi nehézséget, például átlagszámításkor kapott nem egész számot is tartalmazó dokumentumvektorok esetén. Ilyenkor vagy a nominális és ordinális adattípusokra is alkalmazható klaszterezési megoldásokat használjuk csak, vagy a végeredményként kapott nem valós dokumentumokat reprezentáló vektorokhoz rendelünk közelítő dokumentumokat. Prototípus-alapú klaszterezés esetén a szóhalmazok mediánjának meghatározására kifejlesztett módszer (*Kruzslicz* [1999]) a klaszterezési algoritmusokba közvetlenül is beépíthető. A többféle klasztermegközelítés közül a továbbiakban kizárólag a particionáló módszerekkel foglalkozunk, melyek az objektumokat diszjunkt csoportokba sorolják. Ezzel nem zárjuk ki az egyéb (fuzzy, hierarchikus) módszerek alkalmazhatóságát sem, csak feltételezzük, hogy az ilyen algoritmusok végeredményét mindig particionálássá konvertáljuk.

Egy korpusz dokumentumait gyakran nem eredeti formájukban konvertáljuk vektorokká, hanem különféle előkészítő transzformációs lépést végzünk rajtuk. Szövegelemekre bontás (tokenelés) során a szavakat és az írásjeleket választjuk szét egymástól. Szótövezés esetén a szavak különböző alakú előfordulásait helyettesítjük egy közös alakkal. Ez nyelvfüggetlen módon történhet egyszerű, végződéslevágó algoritmussal vagy nyelvdetektálás után a megfelelő szótőkereső szabályok alkalmazásával. A szintaktikai megoldásokhoz képest léteznek szemantikai próbálkozások is, ahol a helyettesítésnél a rokon értelmű (akár a korpuszban nem szereplő) szavakat is figyelembe veszszük. A stopszavazás során a dokumentumokból töröljük azokat a töltelékszavakat, amelyek az adott nyelv leggyakoribb, de legkevesebb jelentéstartalommal rendelkező szavai közé tartoznak, mint például a névelők. Ezek az átalakítások is információ veszteséggel járnak, és fő céljuk a vektortér dimenziószámának csökkentése. Általánosan elfogadott vélekedés, hogy mindez csak a modellezés számítási idő- és tárigényének

csökkentése miatt szükséges. Ha elegendően sok dokumentum áll rendelkezésre, akkor a szövegbányászati modellek magukban automatikusan is elvégeznék ezt helyettünk, csak egyelőre sokkal több idő alatt és kevésbé pontosan. Az adatok előkészítésébe a nyelvészeti tudáson kívül egyéb szakértői ismertek is bevonhatók, ha a korpusz egy jól behatárolható fogalomkörhöz tartozik. A *Statisztikai Szemle* cikkeinek korpuszán például maga a „statisztika" szó valószínűleg stopszónak minősülne, és a „Gauss-eloszlás" pedig egybe nem írt, állandósult összetétel lenne.

A szózsákmodellen kívül sok egyéb dokumentumreprezentációs modell is ismert: a szósorrend részleges megőrzésére például a szavak egymásutániságát leíró, állapotátmeneti gráf mátrixa használható, vagy a szózsákmodell vektorát kiegészíthetjük a szavak szófaji megjelölésével, szemantikai elemeket pedig a mondatszerkezeti fák segítségével vihetünk a modellünkbe. Az egyes dokumentumrészek nyelvének megállapítása után külső lexikai források (például Wikipédia, WordNet) is becsatolhatók az előkészítési folyamatokba. Ezek közül azonban jelenleg még kevésnek van standard szoftvercsomagokban elérhető támogatása. Ennek legfőbb oka az, hogy az összetettebb adatszerkezetek egyelőre nagymértékben megnövelik az algoritmusok számítási igényét.

A szózsák-reprezentáció nemcsak a klaszterezéskor teszi lehetővé a hagyományos algoritmusok alkalmazását, hanem az érvényességvizsgálatok során is támaszkodhatunk a jól ismert módszerekre. Abból kiindulva, hogy a dokumentumok akkor hasonlók egymáshoz, ha leginkább ugyanazon szavakat és nagyjából ugyanolyan arányban tartalmaznak, a klaszterezéshez használt hasonlóságot a koszinusz-távolságmértékből származtathatjuk. Két dokumentum ($\mathbf{d_1}$ és $\mathbf{d_2}$) *koszinusztávolságát* a következő képlettel lehet meghatározni, ahol a számlálóban a két vektor skaláris szorzata szerepel, a $\|\mathbf{d}\|$ pedig a vektor hosszát jelöli:

$$\cos\left(\mathbf{d}_1,\ \mathbf{d}_2\right) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\left(\|\mathbf{d}_1\| \cdot \|\mathbf{d}_2\|\right)}.$$

A klaszterfüggvény ismeretében már tetszőleges kohéziós (klaszter kompaktság) és szeparációs (klaszterek közötti elkülönülés) index vagy ilyenek kombinációja használható a klaszterérvényesség megállapításához. Távolság jellegű klaszterfüggvénynél a kisebb kohéziós és nagyobb szeparációs érték tartozik a jobb klaszterezéshez. A klaszterezés felügyelt, felügyelet nélküli és relatív indexeiről bővebb áttekintést találunk például *Legány–Juhász–Babos* [2006] és *Rendón–Abundez–Arizmendi* [2011] ezzel a témával foglalkozó cikkeiben.

2. Dokumentumcímkézés

Szöveges adatállomány címkézése alatt azt a T hozzárendelést értjük, amely a **d** dokumentumhoz egy olyan, dokumentumonként változó elemszámú $T(\mathbf{d}) = \{t_1, t_2, \ldots, t_k\}$ szóhalmazt (címkéket) rendel hozzá, ami a legjobban jellemzi az adott dokumentum tartalmát, és leginkább megkülönbözteti a többi dokumentumtól. Klaszterezés esetén az előbbi definícióban mindössze annyit kell módosítani, hogy **d** dokumentum helyett hasonló dokumentumok $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\}$ halmazához kell címkéket rendelni.

A címkehalmaz szavaira semmiféle megkötés nincs, és kapcsolatot sem feltételezünk közöttük. Amennyiben a címkeadat-szerkezet halmaz helyett fa, akkor hierarchikus címkézésről (kategóriarendszerről), ha pedig irányított gráf, akkor taxonómiáról beszélünk. Hierarchikus címkézés speciális esetében az egyes klaszterek automatikus elnevezésekor tekintettel kell lenni a címkekategóriák szülő-gyermek kapcsolataira is. *Mao et al.* [2012] a helyes címkézés kialakításához a testvér-testvér és a szülő-gyermek címkékre vonatkozóan két-két kritériumot határozott meg. A testvér kapcsolatokra vonatkozó első és negyedik kritérium lényegében a szóhalmazokra előbbiekben megfogalmazott elvárások átfogalmazása:

> – Az első elv szerint a klasztereket külön-külön vizsgálva, azok címkéinek reprezentatívnak és fontosnak kell lenniük az adott klaszterre nézve.
>
> – A negyedik elv értelmében egy testvérkategória címkéje annál jobb, minél kevesebb másik klaszterben fordul elő gyakori kifejezésként.

A címkehierarchia helyességének definiálásához azonban további két kritériumra is szükség van:

> – A második elv szerint a szülőkategóriák címkéjének a gyermekkategóriákhoz tartozó klaszterek mindegyikében gyakori kifejezésnek kell lennie.
>
> – A harmadik elv szerint pedig a szülőkategória címkéje általánosabb legyen a gyermekkategóriák címkéinél.

Ontológiák[1] használatakor további elnevezéshelyességi elvek bevezetésére van szükség.

Egy egész klasztert jól leíró, és a többi klasztertől megkülönböztető címkézés során kihasználhatjuk a klaszterezés szerkezetét, sőt akár a klasztert előállító algoritmus rész-

[1] Ontológia: megegyezésen alapuló fogalmi rendszer formális, egyértelmű leírása (*Gruber* [1993]).

eredményeit is, mint ahogy azt a nemnegatív mátrixfaktorizációra és a látens Dirichlet-allokációra *Mei–Shen–Zhai* [2007] tették. A korai címkéző módszerek például a hierarchikus klaszterezés-összevonási (vagy szétvágási) fájában található információkat használták fel az egymás alá- és fölérendelt klasztereket megkülönböztető kifejezések megkeresésére. Ennél egyszerűbb megoldás az, ha a klaszterben levő dokumentumok szövegeinek egyesítésével egy új dokumentumot állítunk elő, és az ehhez rendelt címkét tekintjük a klaszter címkéjének. Ezt már csak azért is megtehetjük, mert feltételezzük, hogy minden klaszterbe eleve hasonló dokumentumok kerültek, így az egyesített dokumentum tekinthető a klaszter középpontjának (centroid). De nem muszáj mindjárt a dokumentumokat összesíteni. Az is járható út, ha klaszterezés előtt minden dokumentumot felcímkézünk, és a végén pedig az azonos klaszterbe került dokumentumok címkehalmazát valami módon (például a címkék uniója vagy metszeteként) aggregáljuk. A dokumentumcímkézés is egyfajta tömörítési eljárás, melynek során a legjellemzőbb kifejezések *Manning–Raghavan–Schütze* [2009] szerint alapvetően kétféleképpen állíthatók elő: önleíró (cluster-internal) vagy összehasonlító (cluster-differential) módszerrel. Az önleíró címkézés során csak a klaszterben levő dokumentumokat használhatjuk a helyes címkék megkereséséhez. Ennek meg van az a veszélye, hogy a klaszteren belüli gyakori kifejezések nem biztos, hogy csak az adott klaszterre jellemzők. Önleíró címkézés előtt tehát mindenképpen ajánlott a dokumentumokat stopszavazni nemcsak a nyelv egészére nézve általános szavakkal, hanem a témához tartozókkal is. Az összehasonlító-módszer során már felhasználhatjuk a többi klaszter tartalmát is, így ez az eljárás a hatékonyabb ez előzőnél. A címkék kiválasztása ezen belül is több irányelv szerint történhet. Használhatunk kulcsszó- vagy entitáskinyerő (személy, helyszín stb.) algoritmusokat. A szövegkivonatoló és a lényeget kiemelő módszerek eredménye is átalakítható címkehalmazzá. Ügyeljünk azonban arra, hogy amennyiben a lehetséges címkék halmaza előre adott és ismert, úgy nem klaszterezési, hanem osztályozási feladattal állunk szemben. Tehát a címkeként használható kulcsszavak körének meghatározása is a dokumentumcímkézés része, ami legegyszerűbb módon a korpusz szavainak súlyozásával tehető meg.

Hogy egy ilyen szó mennyire jellemző a dokumentumra (vagy azok egy csoportjára), alapvetően kétféle mutató kombinálásával mérhető: a d_{ij} szógyakoriság azt mutatja meg, hogy az *i.* dokumentumban a korpusz *j.* szava hányszor található meg; míg az f_j dokumentumgyakoriság (document frequency) azt adja meg, hogy a *j.* szó hány különböző dokumentumban található meg. A szógyakoriság tehát önleíró, míg a dokumentumgyakoriság összehasonlító mutató. Egy adott szó annál inkább jellemzője egy dokumentumnak, minél többször fordul benne elő, azaz nagyobb a szógyakorisága; és minél jobban jellemző csak az adott dokumentumra, azaz kicsi a dokumentumgyakorisága. Ez utóbbi mutató irányának megfordításával és az így kapott két érték szorzatából képezzük a *tf-idf* (term frequency-inverse document frequency – szó- és inverz dokumentumgyakoriság) mutatót, melynek képlete:

$$tf\text{-}idf\left(t_j, \mathbf{d}_i\right) = d_{ij} \cdot \log\left(n / f_j\right),$$

ahol n a dokumentumok száma.

Ennek ismeretében a címkézésre leginkább a magas *tf-idf* értékkel rendelkező szavak lesznek alkalmasak. Sokszor azonban a legjobb címkék a dokumentumban meg sincsenek említve, ami főleg hierarchikus kategóriarendszerek esetében fordul elő. Például tudományos közleményeknél a magasabb szintű témaspecifikus megjelölések a publikációt megjelentető szakfolyóirat témaköri leírásában találhatók csak meg. Magában a cikkben ennél pontosabb részterületi szakkifejezések fordulnak csupán elő. Ezért címkézéshez is érdemes külső big data forrásokat (például kategorizált adatbázisokat: *Carmel–Roitman–Zwerdling* [2009], Wikipédia és *Kashireddy–Gauch–Billah* [2013], CiteSeerX) bevonni.

A klaszterérvényesség meghatározásához hasonlóan a címkézés helyességének mérésére is kétféle megközelítés lehetséges. Felügyelt indexek esetén feltételezzük, hogy a dokumentumoknak ismert a valódi címkéje, így csak az azzal való egybeesés mértékét kell meghatározni (*Treeratpituk–Callan* [2006]). Az ilyen pontossági, tisztasági, felidézés, precizitás, normalizált kölcsönös információ, rand index és hasonló mutatók szinte minden adatbányászati témájú könyvben megtalálhatók (*Aggarwal–Zhai* [2012], *Tikk* [2007]). Ha a kérdéses dokumentumoknak nincsenek mérvadó címkéjük, amihez az eredményeket hasonlítani lehetne, akkor ismét külső kategorizált szövegforrások és szövegbányászati módszerek lehetnek a segítségünkre azzal, ha a dokumentumaink idegen nyelvű vagy hasonló tartalmú megfelelői már címkézve vannak. Ezeknek azonban többnyire csak az algoritmusok fejlesztésénél és tesztelésénél van jelentősége. A valóságban inkább felügyelet nélküli címkeérvényességi módszerekre lenne szükség, de ez külső erőforrások bevonása nélkül nagyon nehezen automatizálható. A kutatók a címkézés ellenőrzését legtöbbször emberi szakértő közreműködéssel, manuálisan oldják meg (*Pantel–Ravichandran* [2004], *Maqbool–Babri* [2006], *Geraci et al.* [2006]), ami természetesen nehezen tehető objektív mértékké.

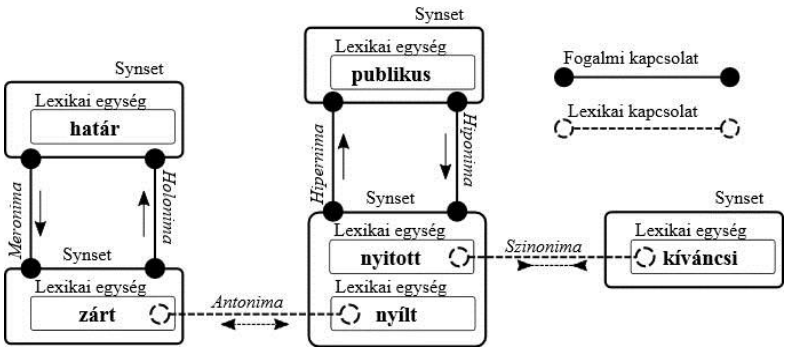
3. Szöveges adatbázisok

Az eddigiek során láthattuk, hogy címkézéskor egy adott korpusz dokumentumai mellett érdemes külső szöveggyűjteményeket, illetve egyéb nyelvi adatbázisokat is felhasználni. Az továbbiakban ezek közül a két leggyakrabban használt adatforrást mutatjuk be, valamint a kapcsolódó kutatások néhány érdekes megoldását és eredményét ismertetjük.

3.1. WordNet

Az egyik ilyen legnagyobb és legszélesebb körben alkalmazott angol nyelvű lexikai adatbázis a Princeton egyetem kutatói által készített WordNet (wordnet.princeton.edu), amely mintegy 150 ezer nyelvtani elemet tartalmaz, az 1. ábrán látható lexikai egységekbe (synset) szervezve. A lexikai egységek révén a WordNet fastruktúra formájában képes megadni egy szó alá-, mellé- vagy fölérendelt kapcsolatait. Az adatbázisban található hasonló és ellentétes jelentésű szavak összefüggéseit szinonima, illetve antonima mellérendelő kapcsolatnak tekintjük, míg a hierarchikus összefüggések közül az általánosító és specializáló (hipernima és hiponima), valamint a része (meronima és holonima) kapcsolatok a legfontosabbak.

1. ábra. Példa lexikai egység kapcsolatokra a WordNet-adatbázisban



Forrás: Saját szerkesztés.

A lexikai egységek kapcsolatai révén a WordNet-rendszer a szemantikai hasonlóság kereséséhez is használható, hiszen két fogalom annál közelebb áll egymáshoz, minél inkább hasonló a kapcsolatrendszerének struktúrája, és minél több a közös lexikai egység bennük. Az egyik legelső ilyen hasonlósági mérték *Wu* és *Palmer* [1994] nevéhez fűződik, és a WordNet-re alapuló, *wup* nevű implementációjára néhány példát is adunk:

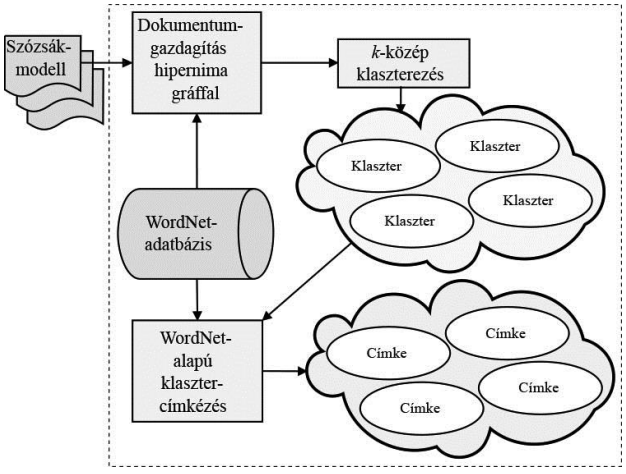
```
wup(statistics, mathematics)  = 0,900,
wup(statistics, economics)    = 0,762,
wup(statistics, distribution) = 0,444,
wup(statistics, probability)  = 0,267.
```

A szemantikai hasonlóság szerinti klaszterezésről és az ilyen mértékek fejlődéséről *Shenoy–Shet–Acharya* [2012] cikkében olvashatunk bővebben, ahol a kutatók arra is felhívják a figyelmet, hogy a lexikai egység kapcsolatainak felhasználásával kapott szóhalmazok túlságosan heterogének lehetnek, amit az algoritmusoknak ke-

zelni kell. A WordNet adatszerkezete nemcsak újfajta klaszterfüggvények megalkotására alkalmas, hanem az adatok előkészítése során is felhasználható. *Gharib–Fouad–Aref* [2010] a szótövezés minőségének javítására használták a WordNet-adatbázist. A hagyományos szótövező algoritmusokkal kapott szótöveket a WordNet szerinti hipernimájukkal helyettesítették. Ezzel egyszerre sikerült a modell dimenzióját csökkenteniük és a hagyományos klaszterező algoritmusok hatékonyságát növelniük a szöveges adatokon. Mivel a WordNet-tel a feldolgozandó korpuszhoz plusz külső fogalmi rendszer kapcsolható, ezért címkézési problémákhoz, különösen önleíró címkézés esetén is hasznos segédeszköz, hiszen a kapcsolati struktúra megnöveli a szövegek belső információtartalmát.

A *Bouras–Tsogkas* [2012] szerzőpáros által kifejlesztett, a 2. ábrán látható *W*-kmeans algoritmus a WordNet-adatbázist egyszerre használja külső forrásként a klaszterezés és a címkézés javítására is. A klaszterezést nem az eredeti szózsák modellen végzik el, hanem a leggyakrabban előforduló szavak 20 százalékára meghatározott, a szóból kiinduló WordNet hipernima gráfokat felhasználva. Például: *statistics → (datum, data_point) → information → (cognition, knowledge, noesis) → psychological_feature → (abstraction, abstract_entity) → entity.* Az így kapott fákat egyesítik, majd a benne szereplő szavakat súlyozzák a gyakoriságuk és a fabeli pozíciójuk alapján. Az egyesített fa legnagyobb súlyú, új szavait hozzáadják a dokumentumhoz. A szemantikai tartalommal így gazdagított szövegreprezentációkon azután hagyományos dokumentum *k*-közép klaszterezést végeznek. Címkézéskor ugyanezt az eljárást követik, csak a dokumentumok helyett a klaszter szavainak 10 százalékra, és az így egyesített hipernima gráf öt legnagyobb súlyú szavát rendelik hozzá a klaszterhez címkeként.

2. ábra. A W-kmeans algoritmus vázlata



Forrás: *Bouras–Tsogkas* [2012] alapján saját szerkesztés.

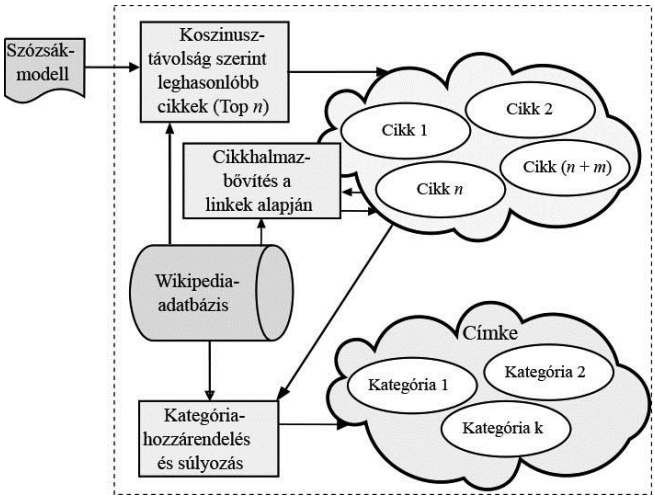
3.2. Wikipédia

A 2001-ben *Jimmy Wales* által útjára indított Wikipédia (wikipedia.org) nevű közösségi, bárki által szerkeszthető enciklopédiát valószínűleg többen ismerik és használják, mint a WordNet-et. A Wikipédia mára a világ egyik legnagyobb szövegkorpuszává nőtte ki magát a maga több mint 5 millió angol nyelvű szócikkével. Mivel egy-egy szócikk az összesen 241-féle nyelvből több „fordításban" is megtalálható, a Wikipédia kimondottan alkalmas fordítástechnológiai megoldásokhoz és nyelvek közötti (cross-lingual) szövegbányászati módszerek fejlesztéséhez. A Wikipédia további fontos tulajdonsága, hogy tartalma kategorizálva van, így külső adatforrásként a klaszterek címkézésében is a segítségünkre lehet. Fontos, hogy a Wikipédia kategóriarendszere elsősorban indexelési célokat szolgál, és így nem szigorúan faszerkezetű. Azaz egy alkategóriának több főkategóriája is lehet. A Wikipédia tartalmak kulcsszavai ontológiába is rendezhetők, ahol az egyes dokumentumok címszavai a hiperlinkeken keresztül kapcsolódnak egymáshoz, sőt akár egyéb külső, formális ontológiákhoz is (például DBpedia, Semantic MediaWiki). Mint ontológia, kellően széles körű, többnyire magas színvonalú, gyorsan alkalmazkodik az újdonságokhoz, és az ember számára is könnyen értelmezhető.

Syed–Finin–Joshi [2008] a Wikipédia alapján olyan gráfot építettek fel, melynek csúcsai az egyes bejegyzések címei (fogalmak) voltak, és a közöttük húzódó irányított élek a bejegyzések közötti linkeknek feleltek meg. Majd ezen a gráfon definiáltak szétterjedéses aktivációs (spreading activation) módszerre alapuló címkézési heurisztikákat. Ezek a módszerek mind bizonyos számú aktív csúcspontból indulnak ki, és az élek mentén újabb kapcsolódó csúcsokat aktiválnak az egyes iterációk során. Az aktiváció terjedését különböző korlátozó feltételekkel (például távolsági, fokszámlimit) lehet szabályozni. Az első módszernél megkeresték az adott dokumentumhoz koszinuszmérték szerint legjobban hasonlító Wikipédia-cikkeket, és vették az ezekhez tartozó kategóriákat. A kategóriákat előfordulásuk száma vagy az összesített koszinuszhasonlóságuk alapján súlyozták, és így választották ki a vizsgált dokumentumhoz leginkább illeszkedő néhány kategória megnevezését címkeként. A második módszerük annyiban különbözött az elsőtől, hogy az ott kapott kategóriákból kiindulva, a Wikipédia-kategóriák gráfjának éleit követve néhányszor alkalmazták a terjedéses aktiválást. Végül az így kibővített kategóriahalmazon végezték el a szokásos súlyozás alapú címkekiválasztást. A 3. ábrán látható harmadik módszernél már felhasználták a bejegyzések linkjeiben tárolt információkat is. Miután meghatározták a dokumentumhoz legjobban hasonlító Wikipédia-cikkeket, a bennük található linkek szerint néhányszor alkalmazva a terjedéses módszert, újabbakat vettek hozzá a legjobban hasonlító cikkek halmazához. Úgy korlátozták a terjedést az irreleváns helyre mutató linkeken, hogy a Wikipédia-cikkek kapcsolati hálójából törölték azokat az éleket, amelyek koszinuszhasonlósága egy adott korlát alá (0,4) esett. A

címkéket az előző módszer szerint, de már a kibővített cikkhalmazt használva állapították meg.

3. ábra. A szétterjedéses aktiváció algoritmusának vázlata



Forrás: *Syed–Finin–Joshi* [2008] alapján saját szerkesztés.

4. Kategóriarendszer-építés

A Wikipédia kapcsán említettük, hogy egy kategorizált cikkadatbázisról van szó, ahol a kategóriákat a szerzők és a szerkesztők kézzel alakították ki és az új eseményeknek, személyeknek megfelelően frissítik. A szétterjedéses aktiváció módszere olyan más adatbázissal is megvalósítható, ahol a dokumentumok egymással össze vannak kapcsolva, és kategóriákba vannak sorolva. Ha a dokumentumok között nincsenek linkek, akkor ez az út már nem járható, de a kategorizált korpuszokat ettől még nagyon jól lehet használni a címkézőalgoritmusok minőségének megítélésére. A vizsgált dokumentumhoz ugyanis először valamilyen módszerrel címké(ke)t rendelünk, majd megkeressük az adatbázisban a dokumentumhoz leginkább hasonló szövegeket, és a címké(ke)t összevetjük az így kapott kategóriákkal. Lényegében ilyen adatbázisnak tekinthető minden olyan webportál, aminek van menüszerkezete. Tipikusan ilyenek a hírportálok különböző (belföld, sport, gazdaság stb.) rovatai. A Reuters hírügynökség kutatási céllal 2000-ben közzé is adott egy 10 788 cikkből álló (1,3 millió szót tartalmazó) korpuszt (http://about.reuters.com/researchandstandards/corpus/), melynek dokumentumai 90 kategóriába vannak besorolva.

Az internetes keresők első generációjának számító webkatalógusok is a kategorizálásra épültek. Ezek közül a legismertebb a Yahoo! (yahoo.com), melynek Yahoo Directory nevű rendszerét szintén manuális úton állították elő. A szolgáltatás ugyan 2014-ben megszűnt, de az akkori állapota (1,8 millió tétel) kutatási célokra továbbra is felhasználható. Helyét a weboldalakat manuálisan kategorizáló Open Directory Project (dmoz.org) vette át, amelyik több mint 4 millió tételével az egyik legnagyobb általános célú webes katalógusrendszerré nőtte ki magát. Bizonyos témaköröknek vagy szakterületeknek saját katalógusrendszerük van. Ilyen például az ACM CCS (Computing Classification System – Számítógépes Osztályozási Rendszer) (http://www.acm.org/about/class/), hatszintű ontológiája, melynek legújabb 2012-es változata 2 113 osztályba sorolja a számítástechnikához és információfeldolgozáshoz kötődő tudományos közleményeket. A klaszteranalízis ebben a hierarchiarendszerben több helyen is megtalálható:

Számítástechnikai módszerek
∟ Gépi tanulás
∟ Tanulási paradigmák
∟ Felügyelet nélküli tanulás
∟ Klaszteranalízis

Matematikai számítástudomány
∟ Valószínűség-számítás és statisztika
∟ Statisztikai paradigmák
∟ Klaszteranalízis

A kategóriarendszerek manuális elkészítéséhez és aktualizálásához jelentős erőforrások szükségesek, ezért komoly igény mutatkozik e tevékenység automatizálására is. Taxonómiák előállítása nem más, mint adott hasonlósági függvény mellett megtalálni egy dokumentumhalmaz legjobb hierarchikus klaszterezését, majd meghatározni a kapott klaszterek legjobb címkézését. Kategóriarendszerek építésének másik jellegzetessége, hogy közben nem támaszkodhatunk egyéb, külső adatforrásokra. Tudományos közlemények ontológiákba sorolásakor ennek eleve kisebb jelentősége van, mert a kategóriacímkék kiválasztásánál a szerzőktől elvárt, hogy pontosan meghatározzák, hogy az írásmű mely területtel foglalkozik, és mi a cikk hozzájárulása a szakirodalomhoz.

Kashireddy–Gauch–Billah [2013] a CiteSeerX (citeseerx.ist.psu.edu) korpusz tudományos cikkeinek automatikus besorolásához fejlesztettek egy módszert, amely alkalmas arra, hogy a bővülő ontológia új kategóriáit automatikus címkékkel lássa el. A munkájuk kezdetén a CiteSeerX korpusz 2 millió dokumentumának mindössze csak 2,6 százaléka volt felcímkézve a CCS ontológia legfeljebb 3. kategóriái szerint. A korpusz többi részét k-NN osztályozási módszerrel sorolták be a létező kategóriákba. Ám ezzel az egyes kategóriák elemszáma több tízezerre nőtt, ami a felhasználók számára böngészésre alkalmatlan. A harmadik szintű kategóriák továbbbontásával viszont pontosabb ontológiát tudtak készíteni és egyben javítani is a cikkek tematikus visszakereshetőségét. A nagyméretű kategóriák publikációit particionáló módszerrel k darab klaszterre bontották, és egy automatikus címkézési algoritmussal

adtak nekik nevet. A megfelelő címke megtalálásához minden klaszterből véletlenszerűen kiválasztottak száz-száz dokumentumot, és meghatározták a bennük szereplő szavak szófajait (POS-tag). Ezek közül csak a főneveket tartották meg potenciális klasztercímkékként. A címkejelöltek rangsorolására a *tf-idf* mutatón kívül két saját mutatót (*delta-tf* és *tf-stdev*) is kipróbáltak. A *delta-tf* az adott klaszteren belüli szógyakoriság és a többi klaszterben megfigyelhető átlagos szógyakoriság különbségeként áll elő. A *tf-stdev* súlyozás lényege, hogy a kifejezés kategóriák közötti szórását veszi figyelembe a súly megállapításakor úgy, hogy a mutató a klaszteren belüli szógyakoriság és a szórás szorzataként számítható ki. A módszer hatékonyságát az ismert kategóriájú dokumentumokon vizsgálva azt tapasztalták, hogy a *tf-stdev* súlyozás adta a legjobb eredményeket. A címkézés minőségének megítélésekor a különböző rangsorok első három-három kifejezésére nézték meg, hogy azok valamelyike megegyezett-e az emberi címkézés által adott kategóriacímmel.

5. Címkézés minőségének mérése

A kategorizált korpuszok azért jelentősek a címkézés szempontjából, mert segítségükkel megvizsgálhatjuk a címkéző algoritmusok eredményességének mértékét. A helyes címkék ismeretében megállapíthatók az egyes címkézések találati arányai. Mint ahogyan azt *Kashireddy–Gauch–Billah* [2013] munkájának bemutatásánál is láttuk, a címkézési feladat nehézsége miatt a teljesen pontos találat nagyon ritka. Ha például egy kézi címke nem szerepel a dokumentumokban, akkor annak megfelelőségét külső adatforrások bevonása nélkül lehetetlen megítélni. Először tehát definiálnunk kell, hogy mit értünk címketalálaton. Ha a definícióban használjuk a *k* paramétert, akkor az azt jelenti, hogy a találatot a súlyozott címkelista hány első elemére vizsgáljuk meg. Ezután már minden olyan mutató kiszámítható, amely osztályozási feladatok teljesítménymérésére használatos, de az egyértelműség miatt a mutató neve mögé odaírjuk az *@K* megkülönböztetést is.

Amennyiben a találatok számát az automatikus címkéző által javasolt címkék számához viszonyítjuk, akkor a precizitás (Precision@K) mutatót kapjuk; ha pedig az emberi címkéző által javasolt címkék számához, akkor a felidézés (Recall@K) mutatót. E két mutató harmonikus közepe pedig az F_1-mérték. A *tf-stdev* súlyozással kapott címkék esetén „az első három közül legalább egy” találati definícióval *Kashireddy–Gauch–Billah* [2013] a következő eredményeket kapták a CiteSeerX korpuszon: precizitás@3 = 0,47; felidézés@3 = 0,56; azaz F_1@3 = 0,55.

Carmel–Roitman–Zwerdling [2009] az ODP (open directory project – nyílt, kategorizált webhivatkozás-gyűjtemény) (http://www.dmoz.org/), illetve a 20NG (20

Newsgroups – 20 hírcsoport húszezer híre) (http://qwone.com/~jason/20Newsgroups/) korpuszokból véletlenszerűen mintavételezett dokumentumokat címkéztek automatikusan újra. A címkézésük jóságának méréséhez pedig a WordNet-et is felhasználták. Náluk egy címke akkor minősült találatnak, ha az megegyezett a kategória vagy a WordNet szerinti lexikai egységének valamelyik (szinonim) szavával vagy egy ilyen szó valamely ragozott alakjával. Ennek megfelelően az általuk használt *Match@K* mutató azt méri, hogy a klaszterek hány százalékában ad találatot a rangsorolt címkejelölt lista top-*k* eleme közül legalább az egyik.

A külső adatforrás bevonása nélkül a szerzők azt figyelték meg, hogy nagyjából az esetek 15 százalékában nincs jelen a kategóriához tartozó dokumentumok szövegében a helyes kategóriacímke. Tehát csupán a dokumentumok szókincse alapján nem lehet tovább javítani az automatikus címkézés minőségét. Ugyanakkor a hosszú toplisták a gyakorlati alkalmazás szempontjából nem kedvezők, azt célszerű alacsonyan $k = 5$ környékén tartani. Ennél alacsonyabb k értékekre viszont a *Match@K* mutató 0,7, valamint 0,5 alá esik vissza a két korpuszon. Végül a dokumentumok maradék 85 százalékát megvizsgálva azt tapasztalták, hogy a jellemzőkiválasztó-módszerek nagyon ritkán tartották az emberi címkéket megfelelőnek. Ami azt mutatja, hogy sok esetben az ember által adott címkék a statisztikai eszköztár szerint nem relevánsak.

A címkézés javítása érdekében a szerzők külső adatforrásként bevonták a Wikipédiát is a folyamatba. A klaszterekben található legfontosabb szavakból keresőkifejezést állítottak össze, majd az ez alapján talált Wikipédia-oldalak metaadatait elemezték. A címükben és kategóriájukban található kifejezéseket hozzávették a szövegben található kifejezésekhez, ezt követően került sor az összes kifejezés súlyozására és rangsorolására. Ezzel a módszerrel a *Match@K* mutató értékét sikerült a k paraméter értékétől függően 10–40 százalékkal javítani. A $k = 5$ esetén a *Match@5* mutató értéke mindkét korpuszon 0,85 fölé emelkedett.

A szerzők által használt másik mutató a reciprokrangátlag (*Mean Reciprocal Rank@K*) volt. Egy klaszter címkejelölt listája top-*k* elemének reciprokrangja a legelső találatot adó címke sorszámának reciproka, illetve nulla, ha egyikük sem ad találatot. Ha több elem is találatot adna, akkor csak a legelsőt vesszük figyelembe. Az *MRR@K* mutató nem más, mint klaszterek reciprok rangjainak az átlaga.

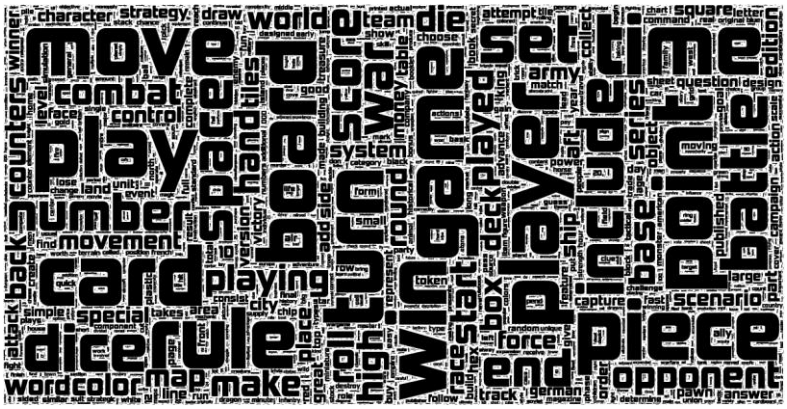
Mao et al. [2012] a Yahoo! Answers (https://answers.yahoo.com/) és a Wikipédia korpusz néhány kiválasztott főkategóriájába tartozó dokumentumokon végeztek hierarchikus címkézést. A címketalálatok között megkülönböztettek pontos és részleges találatot aszerint, hogy a dokumentumhoz rendelt címke az aktuális és szülőkategóriának együttesen is megfelel vagy csak ezek valamelyikének. A megfelelésnél ők is elfogadták a szinonimaszintű egyezéseket is. Mindkét módszerrel mindkét korpuszra kiszámolták az említett mutatók értékeit $k = 1, 3, 5$ mellett. Azt figyelték meg, hogy míg k növelésével a *Match@K* értéke növekszik, addig a *Precision@K*

értéke csökken. Összehasonlításképpen a legjobb pontos találati *Match@5* eredményük minkét korpuszon 0,448 volt, míg a legjobb részleges találati eredmény 0,867 és 0,673 lett, rendre a Yahoo! és a Wikipédia korpuszon.

6. Szófelhődiagramok

A szófelhő, amelyre a 4. ábrán látunk példát, olyan ábrázolási módszer, ahol a diagramon az adatpontokat a címkéjükkel tüntetjük fel, és a hozzá tartozó értékek alapján határozzuk meg a megjelenítendő szöveg pozícióját, irányát, betűtípusát, méretét, színét és egyéb tulajdonságait.

4. ábra. A BGG társasjáték-adatbázis ismertetőinek áttekintő szófelhője*



* Board Game Geek – táblajátékos közösség.
Forrás: Saját szerkesztés.

A legelső ismert ilyen ábrázolási forma *Jodelet–Milgram* [1976] munkája, akik még kézzel rajzolták Párizs térképére a főbb turistalátványosságok neveit úgy, hogy a szöveg nagysága jelentette az adott hely népszerűségét. Az első automatikus szófelhőelőállító, Zeitgeist nevű program 1997-ben jelent meg, ami egy weboldalon használt keresőkifejezésekből alkotott a weboldalba beágyazható diagramot. A Perl szkript szerzője *Jim Flanagan* a HTML (hypertext markup language – hiperszöveges jelölőnyelv) azon képességét használta ki, hogy nagyon könnyű benne a szavak betűméretét és színét beállítani aszerint, hogy azok milyen gyakran fordultak elő. A szófelhő használata csak a 2000-es évektől terjedt el igazán, amikor megjelentek az első online szófelhőgeneráló-webszolgáltatások, melyek közül legismertebb a Wordle (wordle.net)

rendszer. A kezdeti szófelhős vizualizációk megosztották a kutatókat, hiszen csak egy szimpla gyakorisági táblázat egyszerű ábrázolását látták benne, amiről elég nehezen és pontatlanul olvashatók le az információk. Ezért több kutatás is foglalkozott azzal, hogy miként lehet a szófelhő-ábrázolást hatékonyabbá tenni.

Bateman–Gutwin–Nacenta [2008] arra keresték a választ, hogy a betűk típusa, mérete, színe, intenzitása, a szöveg hossza, iránya és pozíciója, valamint a diagramterület mérete és felbontása hogyan befolyásolja az olvashatóságot. *Lohmann–Ziegler–Tetzlaff* [2009] azt vizsgálták meg, hogy a szavak sorrendje, elrendezése és csoportosítása miként segíti a szófelhő értelmezését, illetve milyen a felhasználók szemmozgása a diagram olvasása közben, és mit tudnak leginkább felidézni róla. A szófelhő nemcsak mint látványelem, hanem mint felhasználói navigációs eszköz is az infografika egyik alapelemévé vált, ez annak köszönhető, hogy eléggé jól támogatja a következő négy tevékenységet:

> *Keresés*: valami (vagy valami hiányának) célirányos megtalálása, alternatív találatok gyors beazonosítása.
>
> *Böngészés*: előzetes cél nélkül felfedezés, a felhasználó érdeklődésének felkeltése egy vagy több területen.
>
> *Impresszió*: olvasás során általános benyomás szerzése a jellemzett, mögöttes tartalomról, releváns fogalmakról.
>
> *Mintafelismerés*: bizonyos fogalmak együttállásából a jellemzett fogalomra vonatkozó mélyebb kapcsolatok megsejtése és felismerése.

Népszerűségének növekedésével egyre több olyan publikáció jelent meg, ami új funkciókkal (például alakzatkitöltés, szövegelforgatás) gazdagította a technikát. A diagramon rendelkezésre álló hely optimális kitöltésének algoritmusa is egyre bonyolultabbá vált. *Burch et al.* [2013] olyan elrendező algoritmust fejlesztettek ki, amelyik az azonos tőből származó kifejezések egymásba illesztésével állítja elő az ún. prefixszófelhőt. *Jänicke et al.* [2015] a szófelhő előnyeivel javították fel a tortadiagram hátrányait és alkották meg a tortaszófelhőt (TagPie). Ez egyben a párhuzamos szófelhőkre (Parallel Tag Cloud) is jó példa, ahol egy diagramon egyszerre több korpusz adatait jelenítjük meg. *Collins–Viégas–Wattenberg* [2009] cikkükben már, mint interaktív felhasználói felületet tesztelték a párhuzamos szófelhők különféle megvalósítási lehetőségeinek használhatóságát. *Carpendale et al.* [2010] a szöveges adatok időbeli változásának elemzésére dolgozta ki az értékgörbék (Sparkline) mintájára a trend felismerésben jól használható szöveges értékfelhő (SparkCloud) változatot. A *Diakopoulos et al.* [2015] által fejlesztett összehasonlító szófelhő (Compare Cloud) egy adott kulcsszóval együtt gyakran előforduló szavakat tudja két szövegkorpuszból megjeleníteni úgy, hogy megállapítható legyen, melyik előfordulás melyik korpuszra jellemző leginkább.

5. ábra. A Herman–Rédey [2005] és Pintér–Rappai [2005] tankönyvek egybevetése

Forrás: Saját szerkesztés.

A címkejelöltek listája, amely szavakból és a hozzájuk rendelt súlyokból áll, tökéletesen megfelel a szófelhőkészítés feltételeinek. Az így kapott szófelhő pedig bővebb jellemzését adja az adott klaszternek, mintha csak a lista első *k* elemét sorolnánk fel. Az összehasonlító szófelhők segítségével páronként, a párhuzamos szófelhőkkel pedig együttesen jeleníthetjük meg a klaszterezés címkejelölt-halmazainak egymáshoz viszonyított szerkezetét. A továbbiakban a Conway-féle összehasonlító szófelhő 5. ábrán látható továbbfejlesztett változatát fogjuk használni. A diagram középső részében azok a szavak vannak feltüntetve, melyek mindkét dokumentumban megtalálhatók. A szavak nagysága (elfoglalt területe) az együttes említések számának megfelelően határozható meg. Mivel egy szó kiterjedése kétdimenziós, ezért a betűmérete a gyakoriság négyzetgyökével arányos. Amennyiben az összehasonlítandó két korpusz több dokumentumból áll, akkor a gyakoriságnál kifinomultabb súlyozási módszer (például *tf-idf*) is használható. Az itt elhelyezkedő szavak vízszintes pozíciója azt mutatja, hogy az adott szó mennyivel többször fordul elő az egyik korpuszban, mint a másikban. Az ábra két szélén azok a kifejezések vannak hasonló méretezéssel, de ömlesztve, amelyek csak az adott korpuszra jellemzők. A szavak színezése csak az elkülönítést és a könnyebb olvashatóságot segíti. A diagram közepének és két oldalának betűmérete egymással nem összemérhető. Mivel a közös szavak sokkal gyakoribbak, mint az egyediek, ezért a két szélső szófelhő szavait fel kell nagyítani, hogy olvashatók legyenek. Szintén az áttekinthetőség érdekében az ábrán feltüntetendő szavak számát érdemes korlátozni. Az ábra készítése előtt a dokumentumokon végrehajtottuk a szokásos adat-előkészítési (tokenelés, szótövezés, stopszavazás) lépéseket, és az így kapott szóhalmaz közös és az egyedi szavaiból is csak a leggyakoribbakat tüntettük fel.

Az összehasonlító szófelhő készítésének ismeretében a diagramról azt is leolvashatjuk, hogy a két tankönyv szóhasználata az eltérő szerzők ellenére nem különbözik jelentősen. A könyvek egymásra épülnek, és a második kötet aktívan használja az elsőben található fogalmakat. Az ábraszéli egyedi szavak pedig arról tanúskodnak, hogy a két tananyag egymástól jól elhatárolt.

7. Klaszterkiértékelés összehasonlító címkézéssel

Vegyük észre, hogy ha az 5. ábrán az összehasonlítás tárgya nem két dokumentum, hanem két dokumentumklaszter lenne, akkor azok elkülönülésének mértékét az egyedi szavak minősége és a közös szavak két szélre koncentrált „pillangóalakja" alapján lehetne megítélni. A klaszterek kohézióját pedig a nagyméretű és kisméretű szavak aránya és elhelyezkedése fejezné ki. Azaz vizuális megerősítést kaphatunk arról, hogy a két klaszterbe tartozó objektumok mennyire csoportosulnak egy-egy fogalom köré. A klaszterezés minőségének megállapítását segítő módszerünk pontosan erre az észrevételre épül.

A címkézéssel történő klaszterkiértékelés menete tehát a következő:

1. Klaszterezzük az objektumainkat tetszőlegesen megválasztott módszerrel.

2. Amennyiben az objektumaink nem szövegek, úgy keressünk és rendeljünk hozzájuk egy-egy alkalmas kapcsolódó dokumentumot.

3. Állítsuk elő az egyes klaszterekhez rendelt dokumentumhalmazok címkejelöltjeinek súlyozott listáját.

4. Klaszterpáronként készítsünk a címkék szavaiból egy-egy összehasonlító szófelhőt.

5. Az összehasonlító szófelhők segítségével alkossunk véleményt az egyes klaszterek körülhatároltságáról és definiáltságáról.

A módszer működését egy gyakorlati példán keresztül is szeretnénk bemutatni. Ehhez olyan adatbázist kellett keresnünk, amely objektumainak többféle attribútuma is van, valamint könnyen és egyértelműen lehet dokumentumot hozzájuk rendelni. A választás nemcsak azért esett a BGG (http://boardgamegeek.com/) adatbázisában tárolt társasjátékokra, mert ezek többféle változatos szempont szerint is csoportosíthatók, hanem azért is, mert többnyire mindenki rendelkezik elegendő háttérismerettel erről a szakterületről, így az eredményeket értékelni tudja. Az adatbázisból véletlenszerűen választottunk ki 10 ezer játékot és az ellenőrizhetőség kedvéért nem klaszteresztük őket, hanem előre definiált lekérdezések alapján bontottuk őket csoportokra. A módszer tesztelése során arra vagyunk kíváncsiak, hogy az egyes lekérdezések

mögött meghúzódó előfeltételezéseink visszaigazolódnak-e az összehasonlító szófelhőkben.

6. ábra. Társasjátékcsoportok összehasonlító diagramjai

a) Kategória szerinti bontás

Kártyajáték Kockajáték

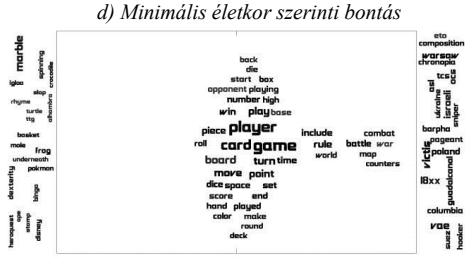
b) A játék mechanizmusa szerinti bontás

Licitálás Mintafelismerés

c) Résztvevők száma szerinti bontás

Maximum 2 fő Minimum 3 fő

d) Minimális életkor szerinti bontás

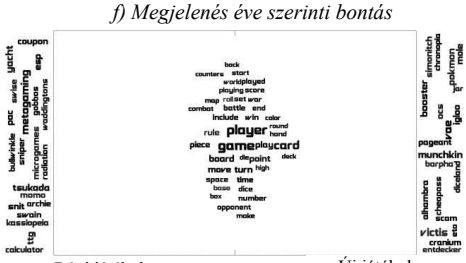


Gyermekjátékok Felnőttjátékok

e) Játékidő szerinti bontás

Rövid Hosszú

f) Megjelenés éve szerinti bontás



Régi játékok Új játékok

Megjegyzés. A diagramok forrásául szolgáló gyakorisági táblázatok az online Mellékletben érhetők el (www.ksh.hu/statszemle).

Forrás: Saját szerkesztés.

A lekérdezésekhez rendelkezésre álló főbb attribútumok a következők: megnevezés, megjelenés éve, játékosok minimális és maximális száma, a játékidő és az ajánlott korhatár alsó és felső korlátja, a játék tervezője és kiadója, a játék típusának és mechanikájának kategóriája, kell-e hozzá nyelvismeret, a játék élvezeti pontértéke és a ranglistán elfoglalt helyezése. A játékokhoz szöveges dokumentumként a rövid ismertetőjüket (description) rendeltük hozzá.

A standard adat-előkészítés utáni korpuszról készült szófelhő nagyméretű, gyakori szavai az adott tématerület speciális, nem túl meglepő (játék, játékos, tábla, bábu, kártya, lép stb.) stopszavait adják meg. A 6. ábrán többféle szempont szerint képeztünk halmazpárokat a játékokból, és a két halmaz leggyakoribb 20-20 közös és egyedi szavai alapján elkészítettük az összehasonlító szófelhőiket. A diagramokon meghagytuk a témaspecifikus stopszavakat is, mert e szavak elhelyezkedéséről a halmazok méretarányai is megállapíthatók. A minden játékra jellemző szavak elhagyásával amúgy tisztább képet kaphatunk a halmazok különbözőségeiről.

a) *Kategória szerinti* bontás esetén azt várjuk, hogy a játékokhoz használt kategóriarendszer találó és odaillő nevekkel jelöli meg a bele tartozó elemeket. A kiválasztott két kategória a kártya- (card) és a kockajátékok (dice) kulcsszava valóban az ábra bal, illetve a jobb oldalára került. Mivel ezek nagyméretűek is, ezért a két dokumentumhalmaz jól szeparáltnak és koherensnek tekinthető. Érdemes megfigyelni, hogy az ábráról azt is megtudhatjuk, hogy az adott tárgyakkal mit szokás csinálni: kártyát húzunk (draw) és kockát dobunk (roll). A gyakori témaközös szavak (game, player) tengelye az ábra bal oldalára tolódott, ami azt jelenti, hogy a halmazok mérete jelentősen eltérő. Valóban, az adatbázisból kiválasztott mintában sokkal több kártyajáték (1644 darab) található, mint kockajáték (595 darab). Az egyedi szavak a kívülállóknak nem sokat mondanak, de leginkább az adott kategóriába tartozó játékcsaládok nevei találhatók itt. Ezek inkább azonosító jellegű kifejezések mintsem általánosítók, ezért a két halmaz további alábontása nem indokolt.

b) A *játékmechanizmus szerinti* bontásnál hasonló figyelhető meg, csak a halmazok elemszáma kiegyenlítettebb: 405 licitálásra és 318 darab mintafelismerésre épülő játék tartozik ide. A másik különbség, hogy a halmaz jellemzéséhez itt már több szó kapcsolódik, hiszen maga a mechanizmus neve nem, vagy nagyon ritkán szerepel a leírásokban. A licitálós játékok jellemző vonásai a pénzhasználat, a vásárlás, az ajánlat és az aukció; míg a mintafelismeréshez szavakat, színeket, lapocskákat, irányokat kell megtalálni. Az itt szereplő egyedi szavak általánosabb jelentésűek, mint az előbb, és ezért alkalmasabbak is az adott mechanizmus jellegzetességeit feltárni. Ilyenkor nagyobb adathalmazok esetén felmerülhet a csoportok alábontása is. A mechanizmus szerinti hasonlóság tehát jobb minőségű bontást ad, mint a kategória szerinti.

c) A *játékosok száma szerinti* bontásnál is megállapíthatjuk, hogy jóval több többszemélyes játék van az adatbázisból vett mintában és a valóságban is. A maximum kétfős játékok egyedi szavai ismét azonosító jellegűek, míg a minimum háromfős játékok egyedi szava inkább általános jelentésűek. A kétféle csoport terminológiájában nagy különbség nincs, mindössze a kétszemélyes játékok között több a háborús (war, battle, combat) tematikájú, és a játékostársat is sokkal inkább hívják ellenfélnek (opponent). Több játékos esetén viszont definiálni kell a sorrendet (round) és a kezdőjátékost (start) is. A két halmaz ennek ellenére a játékosok számának ismerete nélkül nem különböztethető meg markánsan egymástól.

d) A *minimális életkor szerinti* bontás elsőre meglepő módon nagyon hasonlít a játékosok száma szerinti bontáshoz. A 10 éves korhatár megválasztásától azt vártuk volna, hogy a gyerekjátékok még nem igényelnek komolyabb ismereteket és az írni, olvasni és számolni tudás sem feltétel. Ehhez képest a szófelhő más jellemzést adott: a gyerekjátékok egyedi szavai arra utalnak, hogy ezeket a játék tematikája (állatok, rajzfigurák) különbözteti meg elsősorban a felnőtt játékoktól. A közös szavakból pedig azt olvashatjuk le, hogy a gyerekjátékokban a dobás (roll) balra tolódása miatt picit több a szerencsefaktor. A háborús tematika távol áll a gyerekjátékoktól, és főleg a felnőttek játékaira jellemző. A felnőttek egyedi szavai ismét általánosak és arra utalnak, hogy ezekhez már szükség van földrajzi és történelmi ismeretekre is. A két halmaz megkülönböztethetősége itt sem erős, hiszen egy játék külső designjának megváltoztatásával könnyen átkerülhetne az egyik csoportból a másikba.

e) A *játékidő szerinti* bontásnál az egyedi szavak ismét azonosító jellegűek, így nem tudunk általánosabb jellemzést adni arról, hogy mitől igényel egy játék 60 percnél több időt vagy kevesebbet. A közös szavak is csak azt mutatják, hogy a háborús, térképes játékok általában hosszabb ideig tartanak. Az előző két diagrammal összevetve azt is megállapíthatjuk, hogy a játékosok száma, a minimális életkor és a játékidő szerinti bontások nem is önmagukban érdekesek, hanem együtt. Együtt ugyanis azt sejtetik, hogy a felnőtt játékok inkább hosszabbak, a gyerekeknek szánt játékok rövidebbek, és a gyerekek társasjátékai általában többszemélyesek, míg a párharc többnyire a felnőttekre jellemző. Vagyis mindhárom esetben valamiféle olyan csoportot akartunk leírni, aminek mélyebben gyökerező jellemzői vannak.

f) A *játék megjelenésének éve szerinti* bontással azt akartuk megvizsgálni, hogy van-e valami felismerhető trend a játékok fejlődésében. Látható, hogy az 1990-es évet választva határvonalnak az adathalmazt nagyjából két egyenlő részre vágtuk. Azt kaptuk, hogy az egyedi szavak főleg játékok neveit azonosítják, a terminológia pedig szinte teljesen közös a régi és az új megjelenésű játékok között. Ez a fajta kettébontás tehát teljesen indokolatlan, a játék örök.

8. Összefoglalás

Tanulmányunkban egy vizuális klaszterkiértékelési módszert és annak előzményeit mutattuk be. A klaszterek címkejelöltjeinek páros összehasonlítása minden esetben elvégezhető, ha a klaszterezett objektumokhoz szöveges dokumentum, vagy ilyenek halmaza rendelhető hozzá. A módszer újdonsága egyrészt abban rejlik, hogy eddig nem vizsgált módon ad a kezünkbe, külső szöveges adatforrás felhasználásával egy olyan félig felügyelt eljárást, melynek segítségével fogalmat alkothatunk az egyes klaszterek jelentéséről és a klaszterezés minőségéről. Másrészt az elemzéshez használt összehasonlító szófelhő szerkezete is egy korábbi megoldás külalakbeli és

algoritmikus továbbfejlesztése. Minden apró előrelépésre szükség lehet, mert a klaszterkiértékelés mindmáig nem túl kidolgozott és kevésbé használt része a klaszteranalízisnek, holott minden ilyen elemzés kötelező része kellene, hogy legyen. Az általunk javasolt módszer közel sem mondható egzaktnak, de egyfajta előremozdulást jelent a teljesen szubjektív klasztermegítéléstől és -jellemzéstől. Alkalmazásának szubjektív elemei miatt az elsajátítása sem egyszerű, de a bemutatott példák igazolják, hogy valóban hasznos segédeszköz.

A felvázolt technika eléggé általános ahhoz, hogy igény szerint testre szabható és továbbfejleszthető legyen. Legelőször érdemes megvizsgálni, hogy az egyszerű gyakorisági mutató helyett milyen kifinomultabb súlyozási módszerek adnak jobb ábrákat. Egy másik lehetőség a módszer javítására az alkalmazási példában is említett témaspecifikus stopszavak szűrése. Jelen állapotában az összehasonlító diagramnak csak a vízszintes kiterjedéséhez és a betűmérethez rendeltünk adattartalmat. További vizsgálatok folynak a diagram adatgazdagításának érdekében, a színek, a függőleges kiterjedés, az irányultság, a kontraszt és egyéb paraméterek jelentéssel való megtöltésére. Az összehasonlító szófelhő interaktív változata hasznos segítséget nyújthat a manuális címkézés során is a legjobb jelöltek kiválasztásánál. A legnagyobb kihívást a többszörös összehasonlítás megvalósítása jelenti, hogy egyszerre ne csak két klasztert lehessen vizsgálni, hanem hármat vagy többet is.

A módszer kritikus pontja az, hogy a jövő rendszereiben mennyire könnyen lehet majd a strukturált adatokhoz minőségi, nem strukturált adatokat kapcsolni. Ezért a big data trendektől azt reméljük, hogy az adatok egyre kevésbé lesznek strukturált formában tárolva, és elérhetővé válnak azok a technológiák, melyekkel mindebből könnyen kinyerhetők az elemzéshez szükséges strukturált adatok.

Irodalom

AGGARWAL, C. C. – ZHAI, C. X. [2012]: *Mining Text Data*. Springer. Cham. http://dx.doi.org/10.1007/978-1-4614-3223-4

BATEMAN, S. – GUTWIN, C. – NACENTA, M. [2008]: *Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections.* Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia. ACM New York. New York. pp. 193–202. http://dx.doi.org/10.1145/1379092.1379130

BOURAS, C. – TSOGKAS, V. [2012]: A clustering technique for news articles using WordNet. *Knowledge-Based Systems.* Vol. 36. pp. 115–128. http://dx.doi.org/10.1016/j.knosys.2012.06.015

BURCH, M. – LOHMANN, S. – POMPE, D. – WEISKOPF D. [2013]: *Prefix Tag Clouds*. 17th International Conference on Information Visualisation. IEEE Computer Society. Washington, D.C. pp. 45–50. http://dx.doi.org/10.1109/IV.2013.5

CARMEL, D. – ROITMAN, H. – ZWERDLING, N. [2009]: *Enhancing Cluster Labeling Using Wikipedia*. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development

in Information Retrieval. ACM New York. New York. pp. 139–146. http://dx.doi.org/10.1145/1571941.1571967

CARPENDALE, S. – KARLSON, A. K. – LEE, B. – RICHE, N. H. [2010]: SparkClouds: Visualizing trends in tag clouds. *Visualization and Computer Graphics*. Vol. 16. No. 6. pp. 1182–1189. http://dx.doi.org/10.1109/TVCG.2010.194

COLLINS, C. – VIÉGAS, F. B. – WATTENBERG, M. [2009]: *Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora*. Proceedings of the Symposium on Visual Analytics Science and Technology. IEEE. New York. pp. 91–98. http://dx.doi.org/10.1109/VAST.2009.5333443

CONWAY, D. [2011]: *Building a Better Word Cloud*. http://drewconway.com/zia/2013/3/26/building-a-better-word-cloud

CSICSMAN J. [1979]: A klaszter-elemzés módszerei és alkalmazási lehetőségei a statisztikában. *Statisztikai Szemle*. 57. évf. 2. sz. 137–145. old.

DIAKOPOULOS, N. – ELGESEM, D. – SALWAY, A. – ZHANG, A. – HOFLAND, K. [2015]: *Compare Clouds: Visualizing Text Corpora to Compare Media Frames*. IUI Workshop on Visual Text Analytics. CiteSeerX 10 M. 29 March. Atlanta.

DRIVER, H. E. – KROEBER, A. L. [1932]: *Quantitative Expression of Cultural Relationships*. University of California Press. Berkeley.

FUTÓ P. [1979]: Hipergráf modellen alapuló klaszter-elemzés és alkalmazása. *Statisztikai Szemle*. 57. évf. 2. sz. 130–136. old.

FÜSTÖS L. – MESZÉNA GY. – S.-NÉ MOSOLYGÓ N. [1977]: Cluster analízis: fogalmak és módszerek. *Szigma.* X. évf. 3. sz. 111–148. old.

GERACI, F. – PELLEGRINI, M. – MAGGINI, M. – SEBASTIANI, F. [2006]: Cluster generation and cluster labelling for web snippets. *Lecture Notes in Computer Science*. Vol. 4209. pp. 25–36. http://dx.doi.org/10.1007/11880561_3

GHARIB, T. F. – FOUAD, M. M. – AREF, M. M. [2010]: Fuzzy document clustering approach using WordNet lexical categories. In: *Elleithy, K.* (ed.): *Advanced Techniques in Computing Sciences and Software Engineering*. Springer. Cham. http://dx.doi.org/10.1007/978-90-481-3660-5_31

GRUBER, T. R. [1993]: A translation approach to portable ontology specifications. *Knowledge Acquisition.* Vol. 5. Issue 2. pp. 199–220. http://dx.doi.org/10.1006/knac.1993.1008

HERMAN S. – RÉDEY K. [2005]: *Statisztika I.* Pécsi Tudományegyetem. Pécs. http://exam.ktk.pte.hu:81/BSc/tananyag/galfa14a.pdf

JÄNICKE, S. – BLUMENSTEIN, J. – RÜCKER, M. – ZECKZER, D. – SCHEUERMANN, G. [2015]: *Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies*. Digital Humanities Quarterly. University of Leipzig. Leipzig. http://www.informatik.uni-leipzig.de/~stjaenicke/TagPies.pdf

JODELET, D. – MILGRAM, S. [1976]: Psychological maps of Paris. In: *Proshansky, H. – Ittelson, W. H. – Rivlin, L. G.* (eds.): *Environmental Psychology: People and Their Physical Settings*. Rinehart & Winston. New York. pp. 104–124.

KASHIREDDY, S. D. – GAUCH, S. – BILLAH, S. M. [2013]: *Automatic Class Labeling for CiteSeerX*. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. IEEE, WIC, ACM. New York. pp. 241–245. http://dx.doi.org/10.1109/WI-IAT.2013.35

KRUZSLICZ, F. [1999]: Improved greedy algorithm for computing approximate median strings. *Acta Cybernetica*. Vol. 14. Issue 2. pp. 331–339.

LEGÁNY, CS. – JUHÁSZ, S. – BABOS, A. [2006]: *Cluster Validity Measurement Techniques*. Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. WSEAS. Madrid. pp. 388–393.

LOHMANN, S. – ZIEGLER, J. – TETZLAFF, L. [2009]: *Comparison of Tag Cloud Layouts: Task-related Performance and Visual Exploration*. Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I. Springer-Verlag. Berlin, Heidelberg. pp. 392–404. http://dx.doi.org/10.1007/978-3-642-03655-2_43

MANNING, C. D. – RAGHAVAN, P. – SCHÜTZE, H. [2008]: *Introduction to Information Retrieval*. Cambridge University Press. Cambridge. http://dx.doi.org/10.1017/CBO9780511809071

MAO, X. – MING, Z. – ZHA, Z. – CHUA, T. – YAN, H. – LI, X. [2012]: *Automatic Labeling Hierarchical Topics.* Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM New York. New York. pp. 2383–2386. http://dx.doi.org/10.1145/2396761.2398646

MAQBOOL, O. – BABRI, H. A. [2006]: Automated software clustering: An insight using cluster labels. *The Journal of Systems and Software*. Vol. 79. Issue 11. pp. 1632–1648. http://dx.doi.org/10.1016/j.jss.2006.03.013

MEI, Q. – SHEN, X. – ZHAI, C. [2007]: *Automatic Labeling of Multinomial Topic Models*. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM New York. New York. pp. 490–499. http://dx.doi.org/ 10.1145/1281192.1281246

PANTEL, P. – RAVICHANDRAN, D. [2004]: *Automatically Labeling Semantic Classes. HLT-NAACL 2004: Main Proceedings*. Association for Computational Linguistics. Boston. pp. 321–328.

PINTÉR J. – RAPPAI G. [2005]: *Statisztika II*. A következtetéses statisztika alapjai. Pécsi Tudományegyetem. Pécs. http://exam.ktk.pte.hu:81/BSc/tananyag/galfa15a.pdf

RENDÓN, E. – ABUNDEZ, I. – ARIZMENDI, A. – QUIROZ, E. M. [2011]: Internal versus external cluster validation indexes. *International Journal of Computers and Communications*. Vol. 5. Issue 1. pp. 27–34.

SHENOY, M. K. – SHET, K. C. – ACHARYA, D. U. [2012]: A new similarity measure for taxonomy based on edge counting. *International Journal of Web & Semantic Technology*. Vol. 3. Issue 4. pp. 23–30. http://dx.doi.org/10.5121/ijwest.2012.3403

SYED, Z. S. – FININ, T. – JOSHI, A. [2008]: *Wikipedia as an Ontology for Describing Documents.* Proceedings of the Second International Conference on Weblogs and Social Media. AAAI Press. Palo Alto. pp. 136–144. http://ebiquity.umbc.edu/paper/html/id/383/Wikipedia-as-an-Ontology-for-Describing-Documents

TIKK D. [2007]: *Szövegbányászat*. TypoTex Kiadó. Budapest.

TREERATPITUK, P. – CALLAN, J. [2006]: *Automatically Labeling Hierarchical Clusters*. Proceedings of the 2006 International Conference on Digital Government Research. Digital Government Society of North America. San Diego. pp. 167–176. http://dx.doi.org/10.1145/ 1146598.1146650

WU, Z. – PALMER, M. [1994]: *Verb Semantics and Lexical Selection*. Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics. Association for Computational Linguistics. Stroudsburg. pp. 133–138. http://dx.doi.org/10.3115/ 981732.981751

Summary

Giving straightforward names to the divided result groups of clustering data is very important to make a research useable. This is especially important when clustering is the real outcome of the analysis and not just a tool used for data preparation. In this case, the underlying concept of the cluster itself makes the result meaningful and useful. However, a cluster comes into being only in the investigator's mind as one can define or describe it with words. Our method introduced in this paper aims to facilitate and partly automate this verbal characterization process. The objects of clustering are accompanied by an external text database that adds new, previously unused features to the dataset. Clusters are described by labels produced with text mining analytics. The validity of clustering can be characterized by the shape of the final word cloud.