Dokumentumcsoportok automatikus kulcsszavazása és témakövetés

Ács Zsombor[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem
Acs.Zsombor@stud.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
rfarkas@inf.u-szeged.hu

Kivonat: A cikkben bemutatunk egy olyan algoritmust, mely a Látens Dirichlet Allokációt felhasználva, természetes nyelvű szöveghalmazt klaszterez, majd ezeket a csoportokat jól kifejező szavakkal felcímkézi. A kifejlesztett módszer alkalmas a dokumentumhalmaz időintervallumokra felosztott részein keletkezett klaszterhalmazok közötti összefüggések, átmenetek, illetve trendek feltárására. Kidolgoztunk egy olyan entrópiasúlyozáson alapuló címkézőt, mely empirikusan is jobb kulcsszavakkal látja el a klasztereket, mint a sztenderd módszerek (term-frekvencia, χ-négyzet statisztika).

1 Bevezetés

Az internet korának egyik jelentős tendenciája az adatok rohamosan növekvő mennyisége, melyek nagy része szöveges. A klaszterezés, illetve a csoportok felcímkézése egyre gyakrabban használt eszközzé válik a piackutatásokhoz, cikkbázisok, fórumok vagy blogok elemzéséhez, ill. a keresőmotorok informáltságának növeléséhez. Segítségével egy átfogó képet kapunk a dokumentumhalmaz szerkezetéről, ami a szöveges adat terjedelmét figyelembe véve, emberi erővel gyakorlatilag elképzelhetetlen lenne.

A klaszterező eljárások által meghatározott csoportok csak egy számítógépes reprezentációt adnak, mely a gyakorlatban az ember számára nem (vagy csak közvetett módon) értelmezhető, hiszen nem tudjuk, milyen dokumentumok és miért kerültek egy csoportba. Ezért szükséges a csoportokat a legrelevánsabb kulcsszavakkal, címkékkel ellátni.

Többféle megközelítés létezik a címkék meghatározására, legeredményesebbnek a differenciális csoportcímkéző eljárások [2] bizonyulnak. Erre a célra alkalmazhatók a vektortérmodell dimenziócsökkentéséhez is használt jellemzőkiválasztó módszerek. A munkánk során ilyen címkéző algoritmusokat vizsgáltunk, kidolgoztuk az egyik algoritmus kiterjesztését.

Végső célkitűzésünk az, hogy trendeket azonosítsunk, a témák időbeli lefolyását nyomon kövessük szöveges dokumentumok alapján. A trendkövetés egy információelméleti módszer, mely során különféle tendenciákat keresünk az elmúlt időszak adataiban, mellyel jóslást tehetünk a jövőre vonatkozóan. Ez egy meglehetősen új módszer, elsősorban a gazdasági életben – azon belüli is a pénzügyi szektorban – alkalmazzák. A kutatás során egy dátumokkal felcímkézett dokumentumgyűjteményt

osztottunk fel meghatározott időintervallumokra, és ezen „részkorpuszokon” automatikusan kialakított klaszterek közötti hasonlóságokat és tendenciákat figyeltünk meg.

2 Kapcsolódó munkák

Az alábbi fejezetben ismertetjük a Látens Dirichlet Allokáció (LDA) „előfutárát”, a Látens Szemantikus Indexelést, illetve magát a LDA-t. Ezen módszerek az eredeti jellemzők kombinálásából új, eddig nem létező jellemzőket generálnak, melyek száma kevesebb, mint az eredeti halmaz elemszáma, ezzel jelentős redukciót elérve. Először létrehozzák az új jellemzőket, majd a dokumentumokat az új reprezentációnak megfelelő alakra alakítják.

2.1 Látens szemantikus indexelés

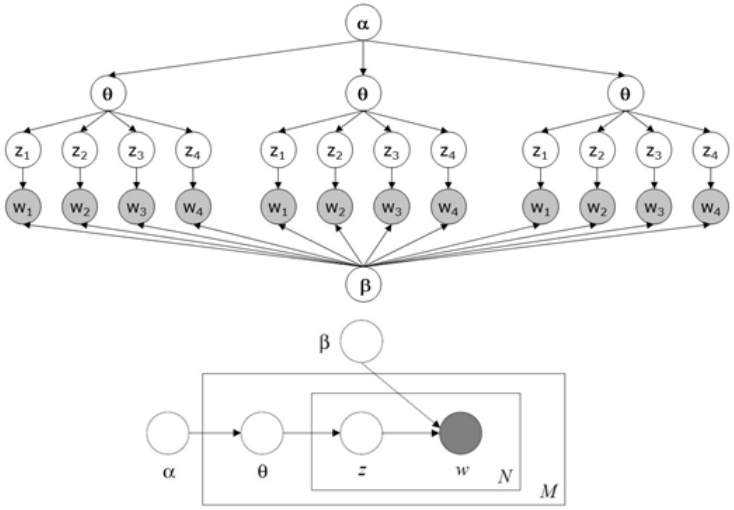
Az egyik legelterjedtebb módszer a látens szemantikus indexelés (LSI), mely egy szinguláris érték-felbontáson alapuló vektortér-transzformáció segítségével az eredeti dokumentumvektorokat kisebb dimenziójú vektorokká alakítja át, melyek meglepően jól jellemzik a korpusz rejtett szemantikai szerkezetét.

Az LSI egyik továbbfejlesztett változata a valószínűségi LSI (pLSI) [2]. A modell lényege, hogy minden dokumentumbeli szó felfogható egy kevert modell mintájának, melynek komponenseit témáknak hívjuk. Így egy dokumentum témák keverékeként értelmezhető.

A generatív valószínűségi modell egy olyan parametrikus modell, amely az egyes változók (paraméterek) olyan értékeit keresi (tanulja), amelyek legnagyobb valószínűséggel generálják a dokumentumkorpuszt. A pLSI hátránya, hogy nem ad generatív valószínűségi modellt a dokumentumok szintjén. Ebből adódóan, nincs természetes mód az előzőleg nem látott dokumentumok priori valószínűségének meghatározására a tanítóhalmazból. További hátrány, hogy a korpusz méretével lineárisan nő az optimalizálandó változók száma ($kV + kM$, ahol k a témák száma, V a szótár, M pedig a dokumentumhalmaz mérete).

2.2 Látens Dirichlet allokáció

A másik kisdimenziós témareprezentációt adó modell, a látens Dirichlet allokáció, többek között a pLSI hátrányait hivatott „kijavítani”. Az LDA is egy generatív valószínűségi modell, amely egy korpusz dokumentumait reprezentálja rögzített számú téma keverékeként. A témákat a szótár felett vett multinomiális valószínűségeloszlásokkal reprezentálja, egy dokumentum pedig ezen eloszlások keveréke (explicit reprezentációt adva) [1]. Így a rejtett multinomiális változók maguk a témák. A modell paramétereinek száma $k + kV$, azaz nem nő a korpusz méretével (ellentétben a pLSI-vel). Az LDA egy háromszintű hierarchikus Bayes-modell, mely a 1. ábrán (fent) látható.



1. ábra. Az LDA Bayes-hálója (fent), illetve gráfikus modellje (lent) [1].

Az LDA gráfikus modelljének reprezentációja az alsó ábrán látható. A téglalapok felfoghatók egymáson lévő lemezeknek, melyek halmazokat ábrázolnak. A külső téglalap a dokumentumok, míg a belső a dokumentumon belüli témák, illetve szavak reprezentációja. α és β hiperparaméterek, az uniform Dirichlet eloszlás paraméterei, előbbi a dokumentumszintű témaeloszlások, utóbbi pedig a szó-téma eloszlás felett. θ_i a témaeloszlást adja meg dokumentumszinten, z_{ij} az i-edik dokumentum j-edik szavához tartozó témát jelenti, w_{ij} pedig az adott szót. Az egyetlen megfigyelhető változó a w_{ij}, a többi rejtett változó.

3 Módszer

A kutatásokhoz rendelkezésre állt az *origo.hu* hírportál *Techbázis* rovatának több mint 10 éves archívuma (1998-2009), mely megközelítőleg 20 000 dokumentumot tartalmaz. A kifejlesztett módszer azonban alkalmas tetszőleges korpusz feldolgozására.

3.1 Előfeldolgozás

A kutatás során reprezentációs modellként a klasszikus vektortérmodellt (VTM) alkalmaztuk.

Előfeldolgozási lépésként a dokumentumhalmazt féléves partíciókra bontottuk, majd elvégeztük a nyers szövegek tokenizálását, lemmatizálását, illetve stopszószűrését [3]. A korpusz szótövezése, illetve a stopszavak eliminálása kulcsfontosságú, általuk jelentősen csökken a VTM dimenziószáma, mely felgyorsítja az LDA

futását, csökkenti a szükséges tárhely méretét, másrészről pedig nagymértékben javít az LDA által generált reprezentáció minőségén, hatékonyságán.

3.1 Kulcsszórangsor

Az LDA modellben is létezik a témákhoz tartozó szavaknak egy relevancia-sorrendje, minden témabeli szó egy súllyal szerepel a témában. Ezen súly meghatározása a dokumentumreprezentálásban általában alkalmazott *term-frekvencia* (TF) súlyozáson alapul – de a dokumentum helyett, a témák szintjén értelmezendő. Így az LDA a következő képlettel számolja a *k*-adik lexikonbeli szó *i*-edik témabeli gyakoriságát:

$$f_{ki} = \frac{n_{ki}}{\sum_{k=1}^{K} n_{ki}}$$

ahol n_{ki}, a *k*-adik lexikonbeli term *i*-edik témabeli előfordulásainak száma (mely a dokumentumok feletti eloszlásból adódik), *K* a lexikon mérete.

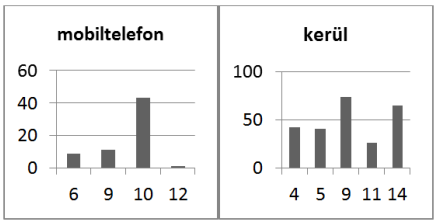
A *χ-négyzet statisztika* olyan klaszterező eljárások utáni címkézésre alkalmas, amelyek egy dokumentumot csak egy csoportba sorolhatnak be. Ezzel szemben az LDA egy valószínűség-eloszlást ad meg dokumentumonként a témák felett.

Lehetséges megoldás lehet, ha egyszerűen vesszük a legnagyobb valószínűséggel rendelkező témát, és azt rendeljük a dokumentumhoz. Így már alkalmazható az eredeti koncepció. Szofisztikáltabb megközelítés, ha megtartjuk a dokumentumok feletti eloszlást, és ennek megfelelően módosítjuk a χ-négyzet statisztika számítási módját:

$$\chi^2(t_i, c_j) = \frac{M \cdot \left(n_{t_i c_j} n_{\overline{t_i c_j}} - n_{t_i \overline{c_j}} n_{\overline{t_i} c_j}\right)^2}{n_{c_j} \cdot n_{\overline{c_j}} \cdot n_{t_i} \cdot n_{\overline{t_i}}}$$

ahol M a dokumentumok számát, t_i az *i*-edik lexikonbeli termet, c_j pedig a *j*-edik témát jelenti. n_{c_j} a dokumentumok *j*-edik klaszterhez tartozásának valószínűségösszege: $\sum_{k=0}^{M} p(d_k|c_j)$. Ebből következik, hogy $n_{\overline{c_j}} = N - n_{c_j}$. n_{t_i} az *i*-edik termet tartalmazó dokumentumok száma, azaz $n_{\overline{t_i}} = N - n_{t_i}$. $n_{t_i c_j}$ az *i*-edik termet tartalmazó dokumentumok *j*-edik klaszterhez tartozásának valószínűségösszege: $\sum_{k=0}^{M} p(d_{k,t_j}|c_j)$. A nem említett mennyiségek a fentiekhez hasonló módon számíthatók ki.

Az általunk kidolgozott kulcsszórangsoroló módszer matematikai alapja az entrópia, mely egy rendszer rendezetlenségi fokát méri. Az alábbi grafikonok a *TechBázis* korpusz 2007/1 félévében levő témaeloszlásokat ábrázolják, a *mobiltelefon*, illetve a *kerül* termekre. Érezhető, hogy a *mobiltelefon* szó sokkal jobban jellemezhetne egy témát, mint a *kerül*.



2. ábra. Termek eloszlása a témák felett.

A címkézés szempontjából releváns szavak entrópiája alacsony, azaz jellemzően egy kimagasló témával rendelkező grafikonnal ábrázolhatók. Csupán entrópiával nem lehet a témákat felcímkézni, hiszen ez az érték egy adott termre minden témában azonos. Az alacsony entrópiájú szavak vélhetően kevés témában vannak jelen nagy számmal, így kihasználhatjuk a termek témabeli gyakoriságát is. Azaz egy alacsony entrópiával rendelkező term azokban a témákban, melyekben nagy előfordulási számmal rendelkezik, hatványozottan magas értéket fog kapni. Lehetséges megoldás, ha vesszük a TF és entrópia egy kombinációját. A *k*-adik lexikonbeli szó *i*-edik témabeli TF-entrópia mértéke a következő:

$$tfent_{ki} = f_{ki} \cdot ((uniform - H(t_k)) + 1)^{\alpha}$$
$$= \frac{n_{ki}}{\sum_{k=1}^{K} n_{ki}} \cdot \left(uniform + \sum_{i=0}^{N} \frac{n_{ki}}{\sum_{j=0}^{N} n_{kj}} \cdot \ln \frac{n_{ki}}{\sum_{j=0}^{N} n_{kj}} + 1 \right)^{\alpha}$$

ahol *K* a szótár mérete, *N* a témák száma, α pedig az entrópia súlyozásának paramétere, mellyel beállíthatjuk az optimális entrópiaarányt. Az *uniform* egy konstans, az előforduló összes entrópia közül a maximumot jelenti. Mivel az alacsony entrópia érték jelent magasabb relevanciát a számunkra, negatív előjellel szerepeltetjük azt. A hozzáadott 1 csupán a [1, uniform+1] értéktartományba való transzformálást eredményezi, hogy az α paraméter hatása minden esetben az entrópiaarány növelését eredményezze.

3.2 Klasztermegfeleltetés

A féléves periódusok klaszterei közötti kapcsolatok feltárására alapvetően három módszert használtunk: halmazmetszeten, vektortávolságon alapuló számításokat, illetve az operációkutatásból ismert hozzárendelési feladatot. A trendek, illetve tendenciák meghatározása egyben a címkéző algoritmusok kiértékelése is volt. A munkahipotézisünk az volt, hogy egy címkéző akkor tekinthető eredményesebbnek, ha az általa felcímkézett klaszterek jobban képezhetők le egy másik, következő félév klaszтereire, azaz erősebb megfeleltetések találhatók a két félév között.

A halmazmetszeten alapuló megközelítés során a témákhoz rendelt szavak címkesúly szerint csökkenő sorrendbe rendezett listájából vettünk egy felső részt (meghatározott százalékot vagy darabszámot), és ezeket a részhalmazokat hasonlítottuk össze.

Vektortávolság esetén a témákat nem csupán a hozzá tartozó szavak halmazaként fogtuk fel, hanem minden term egy súlyértékkel szerepel. Ekkor minden téma felfogható egy vektornak a lexikon szavai által kifeszített vektortérben. Előnye, hogy sokkal jobban reprezentálja az adott témát, hiszen folytonos értékeket használunk. A témavektorok távolságának meghatározását két távolságfüggvény segítségével végeztük: euklideszi, illetve Manhattan-távolság.

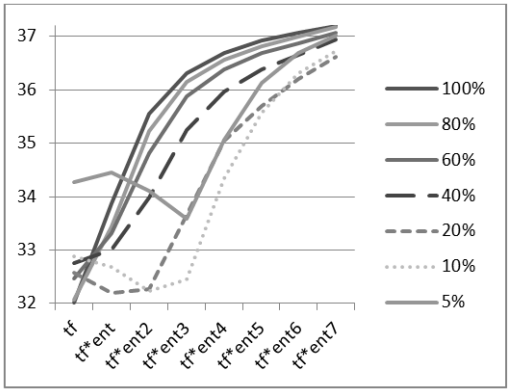
Mindkét esetben a klaszterösszerendelési mátrixon végeztünk maximum (illetve minimum) keresést. A módszer hátránya, hogy egy közel uniform eloszlású sor maximális eleme nem feltétlenül jelent egy valódi átmenetet. Ezen sorok általában a címkéző „gyengeségét" hivatottak jelezni, hiszen közelebb hozzák egymáshoz a klasztereket, így nehezebb a köztük lévő átmenetek felismerése. Az ilyen mátrixok (illetve sorok) „büntetését" a *sorentrópiák* számolása oldotta meg. Ha összegezzük a sorok entrópiáját, kaphatunk egy „jósági értéket" a címkézőre vonatkozóan, melyet a továbbiakban nevezzünk *mátrixentrópiá*nak. Annál jobb a kulcsszórangsor, minél alacsonyabb a mátrixentrópia.

Az entrópiaalapú kiértékelés nem vette figyelembe a metszethalmazok nagyságrendjét, ez motiválta a hozzárendelési feladat alkalmazását. A klaszterek felfoghatók egy gráf csúcsainak, a köztük lévő összehasonlítási mátrix pedig az élek súlyait reprezentálja. Azaz egy teljes páros gráfról beszélünk, ahol mindkét partíció minden csúcsára fennáll, hogy vezet belőle él a másik partíció minden csúcsába. A cél, hogy ebben a gráfban keressünk egy maximális (vektortávolság esetén minimális) összsúlyú teljes párosítást. Így globálisan, a teljes mátrixban találhatunk egy optimális, egy az egyhez összerendelést. A végső jósági érték – azaz a kulcsszórangsorolás minősítése – pedig ezen algoritmus által választott párosítások súlyösszege.

4 Eredmények

Az *entrópiasúlyozás* eredményeit összehasonlítottuk az LDA által adott alap kulcsszó rangsorral (*term-frekvencia*), illetve a hagyományos χ*-négyzet statisztikával*. Emberi kiértékelés alapján a legjobb eredményt produkálta az általunk fejlesztett algoritmus, azaz kifejezőbb címkékkel látta el, mint az alapmódszerek.

A címkéző módszereket automatikus, objektív kiértékelés alá is vetettük, és részben sikerült empirikusan bebizonyítani a saját címkéző jó szereplését. A következő ábra (jobb oldal) a mátrixentrópiák alakulását mutatja különböző címkéző módszerek esetén. Az értékeket a $[0, max - min]$ tartományba transzformáltuk a kifejezőerő növelése érdekében. Az ábra bal oldalán az *entrópiasúlyozás* eredményei láthatóak a növekvő α paraméterének függvényében.



	tf	ent^3	$\chi\,(sz)$	$\chi\,(lt)$
100%	32,022	36,318	38,659	38,881
80%	32,073	36,138	38,657	38,880
60%	32,474	35,881	38,546	38,835
40%	32,755	35,253	38,213	38,594
20%	32,583	33,650	37,989	38,576
10%	32,874	32,446	38,201	38,482
5%	34,265	33,588	38,280	38,438

3. ábra. Mátrixentrópiák alakulása különböző címkéző módszerek esetén.

Megállapítható, hogy kis felső rész vétele esetén *(5%, 10%)* az entrópia kismértékű bevonása (ent^2, ent^3) által csökkent a mátrixentrópia, azaz ekkor könnyebb a klasztermegfeleltetések meghatározása. Ez részben tekinthető egy sikeres bizonyításnak, hiszen a címkézők gyakorlati alkalmazása általában kevés számú címke használatára korlátozódik, és ezen a kisméretű felső részen az entrópiasúlyozás szerepelt a legjobban.

A hozzárendelési feladattal történő kiértékelés során hasonló eredményre jutottunk.

5 További munkák, diszkusszió

A cikkben csupán az egy-egy klasztermegfeleltetéssel foglalkozunk, azaz azzal a feltételezéssel éltünk, hogy a témák halmaza időben lassan változik, egy téma kihalása vagy születése igen ritka. A vizsgált *Techbázis* rovat alapján az esetek nagy részében valóban egy-egy megfeleltetés volt, azonban előfordul a klaszterszétválás, illetve összeolvadás is. A javasolt modellbe bevonhatóak a klaszterek *születésének* és *halálának* esetei is.

A rendszer fő erőssége, hogy segítségével rövid idő alatt határozhatunk meg egy nyers dokumentumhalmazon témákat, melyek egy átfogó, jól használható képet adnak a korpusz tartalmi elemeiről. A trendkövetés nagy előnye, hogy jóslást tehetünk a jövőbeli témákra vonatkozóan, mellyel előre jelezhetők akár bizonyos piacmozgási folyamatok is.

Köszönetnyilvánítás

Jelen kutatást a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

Hivatkozások

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, No. 5 (2003) 993-1022
2. Tikk D. (szerk.): Szövegbányászat. TypoTEX, Budapest (2007)
3. Zsibrita J., Vincze V., Farkas R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 275–283