Revista Română de Interacţiune Om-Calculator 3 (2) 2010, 81-96

O metodă pentru evaluarea şi îmbunătăţirea performanţelor sistemelor de extragere a topicelor

Claudiu Muşat[1], Marian-Andrei Rizoiu[2], Ştefan Trauşan-Matu[1,3]

[1] Universitatea Politehnica Bucureşti
Bd. Splaiul Independenţei, nr. 313 060032, Bucureşti
E-mail: claudiu.musat@cs.pub.ro

[2] Laboratoire ERIC, Université Lyon 2
5, Avenue P. Mendès France 69676 Bron
E-mail: Marian-Andrei.Rizoiu@univ-lyon2.fr
[3] Institutul de Cercetări pentru Inteligenţa Artificială
Calea 13 Septembrie nr.13, Bucureşti
E-mail: trausan@cs.pub.ro

Rezumat. Modelarea topicelor este o direcţie de cercetare în plină ascensiune care necesită noi metode de evaluare şi interpretare a rezultatelor. În această lucrare propunem o metodă pentru evaluarea şi îmbunătăţirea rezultatelor algoritmilor de extragere de topice bazându-ne pe concepte din WordNet. Propunem o metrică pentru determinarea calităţii unui model de topice plecând de la acoperirea sa şi redundanţa sa internă. Apoi, pentru fiecare topic, la nivel individual, cantitatea de informaţie relevantă pe care o aduce modelului este estimată cu ajutorul proiecţiei sale pe ontologie, extrăgând în acelaşi timp cele mai importante cuvinte şi concepte asociate. Modelul ca întreg este îmbunătăţit eliminând cuvintele dăunătoare din fiecare topic, nelegate de conceptul ce reprezintă topicul per ansamblu. Definim apoi o distanţă intre topice pornind de la suprapunerea proiecţiilor lor pe ontologie şi folosim această distanţă pentru a investiga impactul eliminării topicelor redundante din model. Aglomerarea topicelor similare în clustere de topice care sunt relaţionate cu aceleaşi concepte este o altă metodă de îmbunătăţire încercată ca alternativă la ştergerea topicelor irelevante. Rezultatele arată că evaluarea şi îmbunătăţirea modelelor statistice bazându-ne pe baze statice de cunoştinţe duce la modele mai coerente decât cele iniţiale.

Cuvinte cheie: Modelarea Topicelor, WordNet, Ontologie, LDA, Subarbore, Evaluare, Îmbunătăţire.

1. Introducere

Una din provocările zilelor noastre este informaţia, mai precis cantitatea de informaţie furnizată zilnic pe internet. Aceasta ajută utilizatorii să creeze (folosind platforme de genul blogurilor, chat-urilor etc) şi să partajeze

(folosind rețele sociale) informație. Aceasta se găsește, de obicei, în format nestructurat (cel mai adesea în limbaj natural), opus al formatărilor structurate: tabele, formulare. Dat fiind faptul că importanţa economică a informației nestructurate este din ce în ce mai mare (reclame pe rețele sociale, suite de aplicații gen Facebook etc.), unul dintre subiectele de cercetare des abordate în ultimul deceniu este Extragerea Automată de Informaţii din textele în Limbaj Natural.

Una dintre pistele explorate de cercetători este Extragerea de Topice. Dată fiind o colecție de texte în limbaj natural (deseori provenite din internet, ex: bloguri, articole de ziar) se dorește extragerea automată a subiectelor abordate în textele colecției (de exemplu politică, economie, sport). Această posibilitate ar fi de un enorm ajutor pentru toate domeniile unde se lucrează cu cantități mari de text: inventarierea articolelor, filtrarea mailurilor nedorite etc.

Subiectele care apar în textele unei colecții sunt numite în literatura de specialitate topice. Făcând o paralelă cu conceptele, Rizoiu şi Velcin (2010) ne arată că nu există o definiţie larg acceptată a topicelor. În timp ce unii cercetători consideră un topic ca fiind o simplă regrupare a unor texte ce împărtăşesc aceeaşi tematică, alții văd topicele ca pe nişte abstractizări ale grupurilor de texte, fie materializate lingvistic (un cuvânt, o expresie, o propoziție), fie sub forma unui set de perechi cuvânt-probabilitate. În secțiunile următoare vom considera topicul ca o distribuție de probabilități, unde fiecare valoare numerică reprezintă probabilitatea ca un cuvânt să aparțină unui topic.

Dat fiind faptul că distribuțiile de probabilități sunt ilizibile pentru un utilizator uman, au fost propuse sisteme care asociază expresii relevante fiecărui topic pentru a-l rezuma (Rizoiu, Velcin şi Chauchat, 2010; Geraci et al., 2006; Osinski şi Weiss, 2004).

Cu toate că literatura ne oferă multe exemple de sisteme de extragere a topicelor, unele dintre ele fiind prezentate în secţiunea următoare, problema este departe de a fi rezolvată. Rezultatele sunt deseori afectate de zgomot, proprietăți speciale ale cuvintelor (cum ar fi polisemia), dimensiunea setului de texte etc. O altă problematică sensibilă este evaluarea calității topicelor.

Creşterea în popularitate a modelelor de topice a dus şi la generalizarea presupunerii că acestea conțin în mod automat informații utile şi purtătoare de semnificație. Însă, deoarece extragerea topicelor este un proces nesupervizat, a demonstra aceste afirmații este un proces dificil. Acest

obstacol este amplificat şi de faptul că nu există un standard general acceptat cu care să fie comparate modelele studiate. Totodată, deşi s-a depus un volum considerabil de muncă în găsirea unor algoritmi din ce în ce mai performanți şi a unor modele îmbunătățite, comparativ foarte puțin efort a fost dedicat stabilirii unor metode obiective de evaluare a acestora.

În această lucrare propunem o medodă de evaluare complementară celor deja existente, bazată pe caracteristicile intrinseci ale modelelor de topice analizate, şi nu pe documentele din care acestea au rezultat. Folosim apoi această nouă metrică pentru a demonstra posibilitatea îmbunătățirii acestor parametri cu ajutorul datelor dintr-o ontologie binecunoscută, WordNet (Feldbaum, 1998). Prezentăm două modalități de îmbunătățire ale caracteristicilor unui model de topice – în primă instanță acționând asupra topicului la nivel individual şi apoi asupra întregului model.

Expunerea continuă cu descrierea cadrului şi a abordărilor relevante din domeniile conexe în secțiunea 2, acestea fiind urmate de prezentarea metodei propuse în secțiunea 3, rezultatele experimentelor şi concluziile fiind expuse în părțile 4 şi 5.

2. Abordări relevante

Aşa cum am arătat în secţiunea anterioară, sistemele de extragere de topice au utilizări multiple. De aceea, când vine vorba de clasificarea lor, autorii folosesc criterii diferite. De exemplu, în (Rizoiu şi Velcin, 2010) algoritmii sunt împărţiți în funcție de partițiile pe care le crează. Pornind de la observația că un text poate să se refere la mai multe tematici în acelaşi timp (de exemplu efectele economice ale unei decizii politice), autorii argumentează că textele trebuie să aibă posibilitatea să facă parte din mai multe grupuri în acelaşi timp.

Un alt criteriu, pe care îl vom folosi şi noi în continuare, este capacitatea algoritmilor de a lucra direct cu text (de exemplu, LDA (Blei et al., 2003)) sau dacă este nevoie de o tratare inițială. Cei din a doua categorie sunt algoritmi de clustering, adaptați pentru regruparea textelor folosind Modelul Spaţiului Vectorial (Salton, Wong, şi Yang, 1975).

2.1. Algoritmi bazați pe clustering

Clusteringul (clusterizarea), sau regruparea automată nesupervizată, este o tehnică prin care un set de indivizi este partiționat în mulțimi, bazat pe similitudinea trăsăturilor (variabilelor) cu ajutorul cărora indivizii sunt descrişi. Deşi aceştia au fost creați pentru a prelucra date numerice, pot fi adaptaţi pentru text prin traducerea prealabilă a acestuia în Modelul Spațiului Vectorial, introdus de Salton, Wong, şi Yang (1975). În principiu, fiecare text devine un vector într-un spațiu n-dimensional. Fiecărui cuvânt din colecția de texte îi corespunde o dimensiune, iar valorile numerice corespunzătoare fiecărui document pentru fiecare cuvânt sunt determinate folosind o schema de ponderare a cuvintelor (prezenţă/absenţă, frecvența cuvintelor în text, TF-IDF – frecvența termenilor înmulțită cu inversul frecvenței documentelor în care apar termenii).

Rezultatul algoritmilor trecuţi în revistă mai jos este de obicei un "centroid". Acesta este un vector n-dimensional, descris în acelaşi spaţiu ca şi restul documentelor. El reprezintă centrul grupului de texte şi poate fi privit ca o abstractizare a acestora. Deşi centroizii nu sunt distribuții de probabilități, într-o anumită măsură putem să îi considerăm similari acestora şi să aplicăm măsurile de evaluare propuse în secțiunile următoare.

KMeans (Macqueen, 1967) este unul dintre cei mai cunoscuți algoritmi de clustering. A fost îndelung folosit şi numeroase articole i-au demonstrat acuratețea. Algoritmul optimizează iterativ un criteriu obiectiv, de obicei o funcţie de tipul erorii pătratice. În cazul clusteringului textual, distanța bazată pe cosinus este folosită pentru a calcula similitudinea dintre texte.

Bisecting KMeans (Steinbach, Karypis, et Kumar, 2000) este o variantă ierarhică a KMeans, care s-a demonstrat a fi mai precisă pentru clusteringul textual. Este bazat pe un algoritm descendent care divide la fiecare pas documente în 2 sub-clustere. De exemplu, la primul pas întreaga colecţie este divizata în două clustere. La pasul următor, unul dintre grupuri este selectat şi împărțit iarăşi în două subgrupuri. Per total obţinem trei clustere. Procesul se repetă până când o condiție de oprire este îndeplinită (de exemplu un număr fixat de clustere).

Fuzzy KMeans (Dunn, 1973) este o adaptare a algoritmului KMeans folosind logica fuzzy. În clusteringul de tip fuzzy, fiecare document aparţine fiecărui grup într-o anumită măsură, în loc să aparţină integral unui singur grup. De asemenea, documentele care sunt mai departe de centrul grupului aparțin acestuia într-o măsură mai mică decât cele centrale. În Fuzzy

Kmeans, documentele contribuie la calculul centroidului corespunzător în funcţie de gradul lor de apartenență la grupul respectiv.

Latent Semantic Indexing (Berry et al., 1995) este un algoritm statistic de extragere de topice, bazat pe descompunerea în valori singulare. Ideea principală a algoritmului este să descompună matricea cuvânt/document într-un produs de trei matrici: $A = USV^T$. U şi V sunt matrici ortogonale ce conţin valorile singulare la dreapta şi la stânga, iar S este o matrice diagonala având valorile singulare ale lui A ordonate descrescător. Dacă reținem doar primele k valori singulare şi le eliminăm pe restul, împreună cu coloanele şi rândurile corespunzătoare din U, respectiv V, obţinem produsul $A_k = USV^T$, o k-aproximare a lui A. Coloanele din U crează o bază ortogonală pentru spațiul documentelor, deci fiecare document poate fi scris ca o sumă ponderată:

$$d_i = \alpha_1 e_1 + \alpha_2 e_2 + ... + \alpha_k e_k$$

Astfel, elementele e_l , $l \in \{1..k\}$ ale bazei pot fi considerate ca nişte centroizi, iar α_i probabilitatea ca documentul d_i să aparţină clusterului j.

Overlapping KMeans (OKM) (Cleuziou, 2007) este o extensie recentă a bine-cunoscutului KMeans. Funcţionează într-un mod similar, încercând să minimizeze o funcție obiectivă. Principala diferenţa este în OKM, un document poate fi asociat mai multor clustere. Dacă în KMeans, un individ era asociat centroidului cel mai apropiat, măsurat cu distanţa cosinusului, OKM calculează o imagine a centroidului, adăugând documente la cluster până când distanţa dintre el şi imagine este minimă. Această imagine este centrul de greutate documentelor din cluster.

Desigur, în afară de algoritmii prezentaţi mai sus, exista multi alții. Unii dintre ei au chiar şi capacitatea de a asocia o expresie pertinentă pe post de nume, de exemplu AGAPE (Velcin şi Ganascia, 2007). Majoritatea acestora prezintă la ieşire câte un centroid pentru fiecare clasă, ceea ce dă posibilitatea de a utiliza metoda propusa în secţiunea „Sistemul propus” pentru a evalua calitatea topicelor extrase de fiecare din aceste sisteme.

2.2. Algoritmi bazaţi pe distribuţia de probabilităţi

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) este un algoritm probabilist generativ creat special pentru a extrage topice direct din text. LDA consideră documentele ca o colecţie de cuvinte şi modelează fiecare

cuvânt dintr-un document ca un eşantion dintr-un amestec de modele: fiecare componentă a amestecului poate fi văzută ca un "topic". Deci, fiecare cuvânt este generat de un singur topic, dar cuvinte diferite dintr-un document sunt, în general, generate de topice diferite. Fiecare document este reprezentat ca o listă de proporţii diferite pe fiecare componentă a amestecului, deci redus la o distribuţie de probabilități pe un set fixat de topice.

LDA este asemănător cu "probabilistic Latent Semantic Analysis" (pLSA), cu excepţia faptului că în LDA distribuţia de topice se presupune a fi o distribuţie Dirichlet. Acest fapt este foarte important din cauză că permite rezolvarea principalelor limitări ale pLSA: supra-specializarea şi imposibilitatea de a face inferenţe adevărate pe documente noi (Blei et al., 2003). Mai precis, LDA este bazat pe un proces ierarhic şi generativ. Hiper-parametrii α şi β sunt bazele celor 2 distribuții Dirichlet. Prima distribuție Dirichlet controlează modelul de amestecuri pentru fiecare din cele $|D|$ documente. A doua distribuţie controlează generarea amestecului de cuvinte pentru fiecare dintre cele K topice. Deci, fiecare topic este o distribuţie a celor W cuvinte din vocabular.

Procesul generativ este următorul: pentru fiecare cuvânt $w_{d,i}$ din vocabular, generează un topic z dependent de amestecul θ asociat documentului d şi apoi generează un cuvânt din topicul z. Demn de notat este faptul că acele cuvinte care nu au o relevanţă specială (articole, prepoziții) vor avea aproximativ aceeaşi probabilitate pentru fiecare clasă sau pot fi puse într-o clasa aparte. Determinarea parametrilor θ şi z, iar uneori a hiper-parametrilor α şi β, se poate dovedi dificilă pentru că posteriorul $p(\theta, z/D, \alpha, \beta, K)$ nu poate fi calculat cu precizie din cauza sumei infinite de la numitor. De aceea sunt folosiţi diferiți algoritmi de aproximare, cum ar fi EM, Monte-Carlo, procese Markov etc.

O abordare probabilistică ca aceasta prezintă o serie de avantaje şi dezavantaje :

Fundamentul teoretic este bine cunoscut în statistica bayesiana şi a fost îndelung folosit. Este creată pentru a crea inferențe pe documente noi, dar apar probleme: care sunt topicele asociate şi în ce proporţii? Care parte a documentului este asociată cu ce topic? În funcţie de probabilitatea $p(d/\Theta)$, poate un document nou să fie descris de amestecul original de topice sau de un topic nou, care nu a mai fost văzut înainte?

LDA nu prezintă la ieşire un centru pentru fiecare cluster, ci o distribuție a documentelor în funcţie de topice. Acest fapt prezintă o serie de dificultăți în asocierea unui nume lizibil (detalii în Introducere). Lucrări recente încearcă să găsească metode de a asocia nume topicelor găsite de LDA folosind n-grame (Wang, McCallum şi Wei, 2007);

Ca şi celelalte modele prezentate, această metoda probabilistică nu rezolvă problema clasică a găsirii optimului global şi a găsirii numărului *K* de topice. Pentru a doua problemă au fost propuse metode inspirate din selecţia de modele (Rodriguez, 2005).

Algoritmul LDA doar a deschis calea, multe alte modele fiind propuse ulterior pentru a corecta diferite hibe ale modelului iniţial: extragerea de arbori de topice (hLDA) (Blei et al., 2004), inducerea unei structuri de corelaţie între topice (Lafferty şi Blei, 2005), modelarea topicelor dinamice în timp (Blei şi Lafferty, 2006), folosirea de n-grame în loc de cuvinte simple pentru a descrie topicele (Wang, McCallum şi Wei, 2007) etc.

2.3. WordNet

Între sistemele de gestionare şi stocare de cunoştințe, ontologiile ocupă un loc important, între acestea una dintre cele mai cunoscute fiind WordNet (Fellbaum, 1998). Creată la Princeton, această ontologie grupează conceptele limbii engleze, reprezentate atât prin cuvinte cât şi prin grupuri de cuvinte, în mulțimi de sinonime, *synsets*. Fiecare astfel de mulțime defineşte un concept lexical, şi poate fi reprezentat prin o colecție de substantive, adjective, adverbe sau verbe. Trebuie menționat faptul că unui cuvânt din vocabular îi pot corespunde mai multe sensuri, acestea fiind omonime între ele.

Între conceptele astfel definite există unul sau mai multe tipuri de relații, între care se evidențiază cele de hiper şi hiponimie, antonimie sau holonimie. Un caz particular este reprezentat de relațiile de hiper şi hiponimie, care, aplicate substantivelor crează un arbore cu rădăcina în cel mai general concept - *entity*.

Ideea de a folosi ierarhia de concepte în modelarea topicelor a mai fost abordată de Graber (2007) care împreună cu inventatorul LDA, Blei, au creat o versiune modificată a acestui algoritm, versiune care are la bază ideea de drum în WordNet. Potrivit acestei teorii, cuvintele ce corespund unor concepte apropiate din ierarhia din WordNet au şanse mai mari să apară împreună intr-un text. Folosind componența topicelor la finalul rulării,

s-a putut realiza discriminarea între diferitele sensuri ale cuvintelor din experiment.

2.4. Evaluarea modelelor de topice

Conform rezultatelor prezentate atât în (Blei, 2003) cît şi în (Laffery şi Blei, 2005), LDA este, alături de CTM, printre cei mai precişi algoritmi generatori de modele probabilistice. Totodată un avantaj important al LDA din punct de vedere al combinării metodei cu date din WordNet este faptul că majoritatea conceptelor din ontologie sunt unigrame. Astfel putem folosi direct relațiile dintre cuvintele prezente în topice şi conceptele lor omonime.

Cele mai multe metrici folosite frecvent în literatura de specialitate sunt metrici predictive. Acestea descriu capacitatea unui model de a procesa un corpus nou de documente de test după ce au fost antrenate în prealabil. Completarea documentelor (Rosen-Zvi et al., 2004) este una dintre cele mai cunoscute metode de evaluare a performanțelor unui model de topice. Se bazează pe faptul că, dat fiind un model şi o primă jumătate a unui document, probabilitatea ca acesta să fie urmat de cea de-a doua jumătate din textul original trebuie să fie mare. Această probabilitate stă la baza funcției de evaluare a modelului.

Wallach (2009) a demonstrat că metodele folosite în prealabil pentru evaluarea topicelor, inclusiv media armonică sau completarea documentelor dau rezultate necorespunzătoare majoritatea timpului. Totodata a fost prima care a realizat un studiu dedicat acestor metode şi a realizat una proprie, extensibilă la modelele mai complexe.

O metodă originală este cea descrisa de Chang et al. (2009), în care aceştia folosesc adnotarea umană pentru a evalua modelele şi totodată pentru a detecta trăsături noi ale acestora. Aceştia injectează printre cele mai importante cuvinte dintr-un topic unele diferite şi prezintă mulțimea rezultată unui evaluator uman. Apoi demonstrează că, dacă topicul inițial era unul coerent, evaluatorul reuşeşte să diferențieze cuvintele adăugate de cele inițiale. Folosind rata de succes a evaluatorului, aceştia determină calitatea topicelor, iar rezultatele sunt similare cu cele obținute în (Laffery şi Blei, 2005).

Cu toate acestea, nici Wallach, nici Chang nu folosesc în metodele lor caracteristicile intrinseci ale distribuțiilor, ci tratează fiecare model în parte ca o unitate distinctă. În unele aplicații însă, cum este adnotarea topicelor,

este necesară o viziune mai aprofundată, fiecare topic în parte fiind purtător de informație relevantă.

Un corolar al faptului că modelele de topice au fost până acum considerate elemente unitare şi indivizibile este faptul că, din câte ştim, aceasta este prima încercare de a îmbunătăți calitățile modelelor modificându-le componența.

3. Sistemul propus

Aşa cum s-a observat şi în Chang et al. (2009), caracteristicile cele mai importante ale unui topic pot fi sumarizate de cele mai importante cuvinte ale acestuia, în cazul LDA acestea fiind cele a căror probabilitate dat fiind topicul este cea mai mare. Astfel fiecare topic poate fi descris ca o mulțime de cuvinte care îl reprezintă. Notăm cu $W(t_i)$ mulțimea cuvintelor relevante pentru topicul *i*, unde $i \in \{1,..,k\}$, k fiind numărul de topice al modelului.

3.1. Acoperire şi redundanță

Fie *N* numarul de cuvinte diferite prezente în corpusul inițial de texte din care, folosind LDA, sunt extrase topicele. Definim acoperirea modelului ca procentul de cuvinte din total care se regăsesc în reuniunea mulțimilor cuvintelor relevante ale tuturor topicelor.

$$\varphi = card\{\bigcup_{i=1}^{k} W(t_i)\}$$

Un model în care toate cuvintele din vocabular se regăsesc în modelul final nu este însă neapărat unul reuşit. Este necesară contrabalansarea acoperirii cu specificitatea, astfel încat fiecărui topic să îi corespundă cuvinte pe cât posibil diferite de ale celorlalte topice. Definim astfel redundanța topicelor ca numărul de cuvinte care se găsesc în două sau mai multe topice

$$\varrho = \sum_{i=1}^{k} card\{\bigcup_{\substack{j=1 \\ j \neq i}}^{k} (\bigcap W(t_i), W(t_j))\}$$

Reunind cele două concepte, ambele normalizate cu numărul de cuvinte din dicționar *N,* obținem o funcție care relevă caracteristicile topicului dat

$$f(t_i) = \frac{\varphi - \varrho}{N}$$

Menţionăm că atât redundanţa cât şi acoperirea topicelor sunt concepte care în contextul evaluării topicelor nu au mai fost utilizate în prealabil.

3.2. Îmbunătăţirea topicelor

După definirea metricii de calitate a sistemului dorim să aducem îmbunătăţiri modelelor studiate, care apoi să fie analizate conform metricii anterioare. Îmbunătăţirile sunt realizate în două moduri – prin îndepărtarea cuvintelor conceptual diferite sau a topicelor redundante. Eliminarea de topice poate avea loc atât bazându-ne pe toate sensurile asociate cuvintelor conţinute, cat şi doar pe un subset reprezentativ al acestora.

3.2.1. Subarbori WordNet asociaţi topicelor

Sarcina de a găsi cuvintele neînrudite nu este uşor rezolvată folosind o simplă distanţă peste WordNet deoarece fiecare cuvânt dintr-un topic poate corespunde mai multor sensuri din WordNet – uneori de ordinul zecilor. Această complicaţie exclude posibilitatea folosirii unor simpli algoritmi de clustering cum ar fi kMeans.

Pentru a găsi cuvintele neînrudite cu majoritatea celor dintr-un topic, vom defini o structură peste arborele creat de *synset*-urile substantive cu relaţia de hipernimie sau hiponimie din WordNet. Aceasta structură va fi un subarbore al arborelui ce conţine toate *synset*-urile cu proprietatea că acesta conţine toate *synset*-urile asociate cuvintelor din un topic dat.

Generarea acestui subarbore este incrementală – pornind de la mulţimea de concepte direct asociate cuvintelor din topice adăugam acesteia toate conceptele părinte, pana la rădăcina arborelui, *entity*. Mulţimea finală rezultată, împreună cu relaţiile de tip “e un” (*is a*) dintre concepte va reprezenta subarborele ce acoperă în întregime topicul, $S(t_i)$. Fiecare nod c din acest arbore poate fi însă considerat rădăcina unui subarbore $S_c(t_i)$, inclus în $S(t_i)$.

3.2.2. Evaluarea subarborilor

Fiecare subarbore este evaluat în funcţie de specificitatea şi acoperirea sa, cu referire la topicul dat. Definim acoperirea unui subarbore cu rădăcina într-un

nod dat ca numărul de cuvinte din topic care au cel puțin un sens asociat care face parte din structură. Într-un exemplu din lumea animală, dacă un topic ar conține numai trei cuvinte relevante (*cat, dog, idea*) atunci conceptele *cat#n#1*şi dog*#n#1* ar fi ambele radăcinile unui arbore de acoperire 1, pe când părintele lor comun – *animal#n#1*ar avea o acoperire de 2 iar rădăcina celui mai extins arbore ar avea o acoperire maximă, de 3 (egală cu numărul total de cuvinte din topic).

Este evident că acele concepte foarte generale, aflate aproape de rădăcina arborelui cel mai cuprinzător vor avea o acoperire mare în comparație cu cele aflate mai jos în ierarhie. De aceea este necesară şi introducerea unui factor care să măsoare specificitatea fiecărui subarbore în parte. Cu cât un nod este mai aproape de rădăcină, cu atat mai mică este specificitatea sa.

Definim înălțimea unui nod ca distanța sa până la rădăcina arborelui iar adâncimea nodului ca distanța maximă dintre el şi conceptele direct asociate cuvintelor din topic. Distanța în acest caz este măsurată în numărul de tranziții necesare pentru a ajunge parcurgând ramurile arborelui de la un nod la celălalt. Ajungem astfel la o valoare pentru funcția de evaluare a fiecărui nod

$$f = \alpha_d \, d + \alpha_h \, h + \alpha_c \, c$$

unde d,h,c sunt în ordine adâncimea, înălțimea şi acoperirea sa, iar $\alpha_{d,h,c}$ ponderile corespunzătoare, stabilite a priori.

3.2.3. Eliminarea cuvintelor neînrudite

După faza de construcție a subarborelui corespunzător unui topic şi evaluarea individuală a fiecarui nod, cuvintele neînrudite pot fi definite ca acele cuvinte aparținând topicului care au toate sensurile pe ramuri ale subarborelui cu o valoare mică a funcției de evaluare.

Un exemplu ar fi cel al topicului ipotetic anterior (*cat, dog, idea)*, în care cel mai depărtat cuvinte de celelalte este *idea*. Cum $f(animal\#n\#1)$ este valoarea maximă a funcției iar conceptele corespunzătoare cuvântului *idea* nu aparțin subarborelui lui *animal#n#1*, acesta va fi considerat cel mai îndepărtat din model.

3.2.4. Eliminarea şi contopirea topicelor redundante

Pentru o pereche de topice t_i, t_j determinăm similaritatea acestora pornind de la valoarea funcției de evaluare pentru toate nodurile comune $f_i(n), f_j(n)$ ale subarborilor asociați, reținând minimul dintre cele două valori şi însumând

$$S = \sum_{N} \min\left(f_i(n), f_j(n)\right)$$

Două topice care au similaritatea mai mare decât un prag prestabilit pot fi considerate redundante. Pentru grupurile de topice redundante prezentăm două opțiuni – suprimarea topicelor inițiale redundante şi menținerea unui singur reprezentant bazandu-ne pe toate sensurile cuvintelor din topice sau doar pe cele asociate nodurilor din subarbore care au acoperire minim 2.

4. Rezultate experimentale

În oricare dintre cele trei metode prezentate, rezultatele au fost evaluate folosind variația acoperirii şi redundanței modelului rezultat față de cel inițial. În toate testele atât acoperirea cât şi redundanța au variat în acelaşi sens, dar cu viteze diferite, ceeea ce a condus la o diferență care este exact valoarea căutată. Dacă redundanța scade cu o viteză semnificativ mai mare decât acoperirea modelul final poate fi considerat superior.

Prag	Acoperire	Redundanță	Variația acoperirii (%)	Variația redundanței %)	Câştig (%)
0	1531	3238	0.00	0.00	0.00
0.1	1496	3127	2.29	3.43	1.14
0.2	1473	3018	3.79	6.79	3.01
0.3	1410	2785	7.90	13.99	6.09
0.4	1311	2403	14.37	25.79	11.42
0.5	1179	1921	22.99	40.67	17.68
0.6	1066	1508	30.37	53.43	23.06
0.7	**993**	**1244**	**35.14**	**61.58**	**26.44**
0.8	916	1086	40.17	66.46	26.29
0.9	847	981	44.68	69.70	25.03
1	814	917	46.83	71.68	24.85

Tabel 1 – Câştigul obținut prin eliminarea cuvintelor neînrudite

Corpusul folosit cuprinde 21106 articole publicate între 2007 şi 2010 în secțiunea *Finance* a cotidianului *The Daily Telegraph*, iar vocabularul cuprinde 44484 cuvinte distincte. Suita Mallet (http://mallet.cs.umass.edu), dezvoltată de Universitatea Massachusetts a fost folosită pentru determinarea topicelor cu algoritmul LDA cu 100 topice, din fiecare reținând cele mai importante 50 cuvinte cheie.

Tabelele 1, 2 şi 3 prezintă câştigurile de performanță obținute în cele trei experimente, iar cazul în care este obținut câştigul maximal este prezentat cu font mărit şi îngroşat.

μ	Topice sterse	Acoperire	Redundanță	Variația acoperirii (%)	Variația redundanței (%)	Câştig (%)
2		1531	3238	0.00	0.00	0.00
1.9	0	1531	3238	0.00	0.00	0.00
1.8	5	1497	3033	2.22	6.33	4.11
1.7	11	1453	2786	5.09	13.96	8.86
1.6	23	1363	2300	10.97	28.97	18.00
1.5	36	1252	1797	18.22	44.50	26.28
1.4	**48**	**1119**	**1356**	**26.91**	**58.12**	**31.21**
1.3	72	760	573	50.36	82.30	31.94
1.2	87	421	195	72.50	93.98	21.48
1.1	94	225	62	85.30	98.09	12.78
1	98	39	10	97.45	99.69	2.24

Tabel 2 – Câştigul obținut prin eliminarea topicelor redundante

În primul experiment fiecare topic este tratat separat de celelalte iar cuvintele sale cele mai puțin înrudite cu celelalte eliminate. Pentru a distinge între nodurile subarborelui asociat care au o funcție de evaluare suficient de mare şi cele care trebuie eliminate am folosit un prag variabil, ca în tabelul 1, prezentat mai jos. Cu prețul pierderii a 35% din acoperirea modelului, redundanța acestuia scade cu peste 60%.

Un al doilea test conține câştigul obținut prin ştergerea topicelor redundante din model.Un topic este considerat redundant dat fiind alt topic procesat anterior

dacă similaritatea lor depăşeşte cu un factor μ similaritatea medie a modelului. Rezultatele sunt semnificativ mai bune decât în primul experiment, o scădere de aproape 60% a redundanţei fiind atinsă cu o pierdere de numai 26% a acoperirii.

Metoda de mai sus poate fi îmbunătăţită prin includerea în formula similarităţii între topice numai a nodurilor care nu sunt direct asociate unor cuvinte din topice. Limităm astfel riscul ca, dacă două topice conţin aceleaşi cuvinte, dar care au alte sensuri, să fie considerate pe nedrept redundante. Luăm astfel în considerare numai acele noduri care au o acoperire egală sau mai mare ca doi. Astfel fiecare cuvânt din topic va fi reprezentat numai de sensurile sale cele mai probabile. Rezultele sunt prezentate în tabelul 3 unde se observă o creştere a caştigului de peste 2.5% faţă de metoda precedentă.

μ	Topice sterse	Acoperire	Redundanţă	Variaţia acoperirii (%)	Variaţia redundanţei (%)	Câştig (%)
2		1531	3238	0	0	0
1.9	0	1531	3238	0	0	0
1.8	0	1531	3238	0	0	0
1.7	0	1531	3238	0	0	0
1.6	0	1531	3238	0	0	0
1.5	2	1523	3151	0.52	2.69	2.16
1.4	11	1458	2778	4.77	14.21	9.44
1.3	22	1372	2334	10.39	27.92	17.53
1.2	**50**	**1111**	**1250**	**27.43**	**61.40**	**33.96**
1.1	79	614	371	59.90	88.54	28.65
1	92	287	93	81.25	97.13	15.87

Tabel 3 – Câştigul obţinut prin eliminarea topicelor redundante bazat doar pe noduri semnificative

5. Concluzii

În această lucrare am prezentat o metodă de evaluare a calităţilor unui model de topice complementar celor deja existente. Folosind această metrică am propus trei modalităţi de a îmbunătăţi trăsăturile modelelor analizate.

Rezultatele obținute – cu ameliorarea de peste o treime a trasăturilor inițiale ne îndreptățesc să sperăm că aceste metode vor putea fi folosite ca post-procesare pentru algoritmii de extragere de topice în viitor. Important este faptul că ameliorarea se poate face la nivel de topic individual, fapt care poate permite o mai bună etichetare a topicului în cauză, ceea ce lărgeşte semnificativ aria de aplicabilitate a modelului în sine.

Mulțumiri

Adresăm mulțumiri prof. Julien Velcin, Université Lumière Lyon2, pentru ajutorul şi implicarea sa în realizarea acestei metode.

Rezultatele prezentate în acest articol au fost obținute cu sprijinul Ministerului Muncii, Familiei şi Protecției Sociale prin Programul Operațional Sectorial Dezvoltarea Resurselor Umane 2007-2013, Contract nr. POSDRU/6/1.5/S/19.

Referințe

Berry, M. W., Dumais, S., O'Brien, G., Berry, M. W., Dumais, S. T., and Gavin. Using linear algebra for intelligent information retrieval. SIAM Review, 1995.

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information processing systems, 2004.

Blei, D. and Lafferty, J. Dynamic topic models. In Proceedings of the 23rd international conference on machine learning, 2006

Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J. Latent dirichlet allocation. Journal of Machine Learning Research, 3 , 2003.

Chang, J., Jordan B.G., Gerrish S., Wang C. and Blei D. Reading Tea Leaves: How Humans Interpret Topic Models. *NIPS.* 2009

Cleuziou, G. Okm : une extension des k-moyennes pour la recherche de classes recouvrantes. In M. Noirhomme-Fraiture & G. Venturini (Eds.), Egc, Cépaduès-Editions,2007.

Dunn, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, 1973.

Felbaum, C.: WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. (1998)

Geraci, F., Pellegrini, M., Maggini, M. and Sebastiani, F. Cluster generation and cluster labelling for web snippets: A fast and accurate hierarchical solution. 2006.

Graber, J., Blei, D., Zhu, X., A Topic Model for Word Sense Disambiguation. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.

Lafferty, J. D. and Blei, D. M., Correlated topic models. In NIPS. 2005.

Macqueen, J. B. Some methods of classification and analysis of multivariate observations. In Proceedings of the fifth berkeley symposium on mathematical statistics and probability,1967.

Osinski, S. and Weiss, D. Conceptual clustering using lingo algorithm: Evaluation on open directory project data. 2004.

Rizoiu, M-A. and Velcin, J. Topic Extraction for Ontology Learning. Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances. IGI Global. 2010.

Rizoiu, M.-A., Velcin, J. and Chauchat, J.-H. Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes. 10éme conférence Extraction et Gestion des Connaissances (EGC 2010), Hammamet, Tunisie, 2010.

Rodriguez, C. The ABC of Model Selection: AIC, BIC and the New CIC. Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 2007.

Rosen-Zvi, M., Gri_ths, T., Steyvers, M., & Smyth, P. The author-topic model for authors and documents. Conference on Uncertainty in Artificial Intelligence 2004.

Salton, G., Wong, A., and Yang, C. S. A vector space model for automatic indexing. Commun. ACM, 1975

Steinbach, M., Karypis, G., and Kumar, V. A comparison of document clustering techniques, 2000

Velcin, J., and Ganascia, J.-G. Topic extraction with agape. In Adma, 2007.

Wallach H, Murray I, Salakhutdinov R. and Mimno D. Evaluation Methods for Topic Models. ICML, 2009, Montreal, Quebec

Wang, X., McCallum, A., and Wei, X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Proceedings of the 7th ieee international conference on data mining, 2007.

Wang, X., McCallum, A., and Wei, X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In Proceedings of the 7th ieee international conference on data mining, 2007