

O metodă pentru evaluarea și îmbunătățirea performanțelor sistemelor de extragere a topicelor

Claudiu Mușat¹, Marian-Andrei Rizoiu², Ștefan Traușan-Matu³

¹ Universitatea Politehnică București
Bd. Splaiul Independenței, nr. 313 060032, București
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 Universitatea Politehnică București
Bd. Splaiul Independenței, nr. 313 060032, București

Institutul de Cercetări pentru Inteligența Artificială
Calea 13 Septembrie nr.13, București
E-mail: trausan@cs.pub.ro

Rezumat. Modelarea topicelor este o direcție de cercetare în plină ascensiune care necesită noi metode de evaluare și interpretare a rezultatelor. În această lucrare propunem o metodă pentru evaluarea și îmbunătățirea rezultatelor algoritmilor de extragere de topic bazându-ne pe concepte din WordNet. Propunem o metrică pentru determinarea calității unui model de topic plecând de la acoperirea sa și redundanța sa internă. Apoi, pentru fiecare topic, la nivel individual, cantitatea de informație relevantă pe care o aduce modelului este estimată cu ajutorul proiecției sale pe ontologie, extrăgând în același timp cele mai importante cuvinte și concepte asociate. Modelul ca întreg este îmbunătățit eliminând cuvintele dăunătoare din fiecare topic, nelegate de conceptul ce reprezintă topicul per ansamblu. Definim apoi o distanță între topic pornind de la suprapunerea proiecțiilor lor pe ontologie și folosim această distanță pentru a investiga impactul eliminării topicelor redundante din model. Aglomerarea topicelor similare în clustere de topic care sunt relaționate cu aceleași concepte este o altă metodă de îmbunătățire încercată ca alternativă la ștergerea topicelor irelevante. Rezultatele arată că evaluarea și îmbunătățirea modelelor statistice bazându-ne pe baze statice de cunoștințe duce la modele mai coerente decât cele inițiale.

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

1. Introducere

Una din provocările zilelor noastre este informația, mai precis cantitatea de informație furnizată zilnic pe internet. Aceasta ajută utilizatorii să creeze (folosind platforme de genul blogurilor, chat-urilor etc) și 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 importanța 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 Informații din textele în Limbaj Natural.

Una dintre piste 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 și Velcin (2010) ne arată că nu există o definiție larg acceptată a topicelor. În timp ce unii cercetători consideră un topic ca fiind o simplă regrupare a unor texte ce împărtășesc aceeași tematică, alții văd topicile ca pe niște 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 și Chauchat, 2010; Geraci et al., 2006; Osinski și Weiss, 2004).

Cu toate că literatura ne oferă multe exemple de sisteme de extragere a topicelor, unele dintre ele fiind prezentate în secțiunea 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 problemă sensibilă este evaluarea calității topicelor.

Creșterea în popularitate a modelelor de topice a dus și la generalizarea presupunerii că acestea conțin în mod automat informații utile și purtătoare de semnificație. Însă, deoarece extragerea topicelor este un proces nesupervizat, a demonstra aceste afirmații este un proces dificil. Acest obstacol este amplificat și de faptul că nu există un standard general acceptat cu care să fie comparate modelele studiate. Totodată, deși s-a depus un volum considerabil de muncă în găsirea unor algoritmi din ce în ce mai performanți și a unor modele îmbunătățite, comparativ foarte puțin efort a fost dedicat stabilirii unor metode obiective de evaluare a acestora.

În această lucrare propunem o metodă de evaluare complementară celor deja existente, bazată pe caracteristicile intrinseci ale modelelor de topice analizate, și 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 și apoi asupra întregului model.

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

2. Abordări relevante

Așa cum am arătat în secțiunea 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 și Velcin, 2010) algoritmi sunt împărțiți în funcție de partițiile pe care le crează. Pornind de la observația că un text poate să se refere la mai multe tematici în același 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 același timp.

Un alt criteriu, pe care îl vom folosi și 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 Spațiului Vectorial (Salton, Wong, și 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 descriși. Deși aceștia au fost creați pentru a prelucra date numerice, pot fi adaptați pentru text prin traducerea prealabilă a acestuia în Modelul Spațiului Vectorial, introdus de Salton, Wong, și 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 (prezență/absență, frecvența cuvintelor în text, TF-IDF – frecvența termenilor înmulțită cu inversul frecvenței documentelor în care apar termenii).

Rezultatul algoritmilor trecuți în revistă mai jos este de obicei un “centroid”. Acesta este un vector n-dimensional, descris în același spațiu ca și restul documentelor. El reprezintă centrul grupului de texte și poate fi privit ca o abstractizare a acestora. Deși centroizii nu sunt distribuții de probabilități, într-o anumită măsură putem să îi considerăm similari acestora și 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 și numeroase articole i-au demonstrat acuratețea. Algoritmul optimizează iterativ un criteriu obiectiv, de obicei o funcție de tipul erorii pătratică. 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 colecție este divizată în două clustere. La pasul următor, unul dintre grupuri este selectat și împărțit iarăși în două subgrupuri. Per total obținem 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 aparține fiecărui grup într-o anumită măsură, în loc să aparțină 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 funcție de gradul lor de apartenență la grupul respectiv.

Latent Semantic Indexing (Berry et al., 1995) este un algoritm statistic de extragere de topic, 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 și V sunt matrici ortogonale ce conțin valorile singulare la dreapta și la stânga, iar S este o matrice diagonală având valorile singulare ale lui A ordonate descrescător. Dacă reținem doar primele k valori singulare și le eliminăm pe restul, împreună cu coloanele și rândurile corespunzătoare din U , respectiv V , obținem 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:

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

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

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

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

2.2. Algoritmi bazați pe distribuția de probabilități

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) este un algoritm probabilist generativ creat special pentru a extrage topic direct din text. LDA consideră documentele ca o colecție de cuvinte și modelează fiecare cuvânt dintr-un document ca un eșantion 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 proporții diferite pe fiecare componentă a amestecului, deci redus la o distribuție de probabilități pe un set fixat de topice.

LDA este asemănător cu “probabilistic Latent Semantic Analysis” (pLSA), cu excepția faptului că în LDA distribuția de topice se presupune a fi o distribuție Dirichlet. Acest fapt este foarte important din cauză că permite rezolvarea principalelor limitări ale pLSA: supra-specializarea și imposibilitatea de a face inferențe adevărate pe documente noi (Blei et al., 2003). Mai precis, LDA este bazat pe un proces ierarhic și generativ. Hiper-parametrii α și β sunt bazele celor 2 distribuții Dirichlet. Prima distribuție Dirichlet controlează modelul de amestecuri pentru fiecare din cele $|D|$ documente. A doua distribuție controlează generarea amestecului de cuvinte pentru fiecare dintre cele K topice. Deci, fiecare topic este o distribuție 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 și apoi generează un cuvânt din topicul z . Demn de notat este faptul că acele cuvinte care nu au o relevanță specială (articole, prepoziții) vor avea aproximativ aceeași probabilitate pentru fiecare clasă sau pot fi puse într-o clasă aparte. Determinarea parametrilor θ și z , iar uneori a hiper-parametrilor α și 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 folosiți diferiți algoritmi de aproximare, cum ar fi EM, Monte-Carlo, procese Markov etc.

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

Fundamentul teoretic este bine cunoscut în statistica bayesiană și a fost îndelung folosit. Este creată pentru a crea inferențe pe documente noi, dar apar probleme: care sunt topicele asociate și în ce proporții? Care parte a documentului este asociată cu ce topic? În funcție 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 ieșire un centru pentru fiecare cluster, ci o distribuție a documentelor în funcție 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 și Wei, 2007);

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

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

2.3. WordNet

Între sistemele de gestionare și stocare de cunoștințe, 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 și prin grupuri de cuvinte, în mulțimi de sinonime, *synsets*. Fiecare astfel de mulțime definește un concept lexical, și 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 și hiponimie, antonimie sau holonimie. Un caz particular este reprezentat de relațiile de hiper și 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 șanse mai mari să apară împreună într-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 topic

Conform rezultatelor prezentate atât în (Blei, 2003) cât și în (Laffery și Blei, 2005), LDA este, alături de CTM, printre cei mai preciși 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 topic și conceptele lor omonime.

Cele mai multe metrice folosite frecvent în literatura de specialitate sunt metrice 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 topic. Se bazează pe faptul că, dat fiind un model și 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. Totodată a fost prima care a realizat un studiu dedicat acestor metode și a realizat una proprie, extensibilă la modelele mai complexe.

O metodă originală este cea descrisă de Chang et al. (2009), în care aceștia folosesc adnotarea umană pentru a evalua modelele și totodată pentru a detecta trăsături noi ale acestora. Aceștia injectează printre cele mai importante cuvinte dintr-un topic unele diferite și prezintă mulțimea rezultată unui evaluator uman. Apoi demonstrează că, dacă topicul inițial era unul coerent, evaluatorul reușește să diferențieze cuvintele adăugate de cele inițiale. Folosind rata de succes a evaluatorului, aceștia determină calitatea topicelor, iar rezultatele sunt similare cu cele obținute în (Laffery și 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 topic au fost până acum considerate elemente unitare și indivizibile este faptul că, din câte știm,

aceasta este prima încercare de a îmbunătăți calitățile modelelor modificându-le componența.

3. Sistemul propus

Așa cum s-a observat și 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 topicuri al modelului.

3.1. Acoperire și redundanță

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

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

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

$$q = \sum_{i=1}^k \text{card}\left\{\bigcup_{j=1, j \neq i}^k (W(t_i) \cap W(t_j))\right\}$$

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

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

Menționăm că atât redundanța cât și acoperirea topicurilor sunt concepte care în contextul evaluării topicurilor nu au mai fost utilizate în prealabil.

3.2. Îmbunătățirea topicelor

După definirea metricii de calitate a sistemului dorim să aducem îmbunătățiri modelelor studiate, care apoi să fie analizate conform metricii anterioare. Îmbunătățirile 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 conținute, cât și doar pe un subset reprezentativ al acestora.

3.2.1. Subarbori WordNet asociați topicelor

Sarcina de a găsi cuvintele neînrudite nu este ușor rezolvată folosind o simplă distanță peste WordNet deoarece fiecare cuvânt dintr-un topic poate corespunde mai multor sensuri din WordNet – uneori de ordinul zecilor. Această complicație 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 relația de hipernimie sau hiponimie din WordNet. Aceasta structură va fi un subarbor al arborelui ce conține toate *synset*-urile cu proprietatea că acesta conține toate *synset*-urile asociate cuvintelor din un topic dat.

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

3.2.2. Evaluarea subarborilor

Fiecare subarbor este evaluat în funcție de specificitatea și acoperirea sa, cu referire la topicul dat. Definim acoperirea unui subarbor 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###1* și *dog###1* ar fi ambele rădăcinile unui arbore de acoperire 1, pe când părintele lor comun – *animal###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 și introducerea unui factor care să măsoare specificitatea fiecărui subarbore în parte. Cu cât un nod este mai aproape de rădăcină, cu atât mai mică este specificitatea sa.

Definim înălțimea unui nod ca distanța sa până la rădăcina arborelui iar adâncimea nodului ca distanța maximă dintre el și 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 și 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 și evaluarea individuală a fiecărui 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 $f(\text{animal}\#n\#1)$ 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 și 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 și însumând

$$S = \sum_n \min(f_i(n), f_j(n))$$

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 și 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 și redundanței modelului rezultat față de cel inițial. În toate testele atât acoperirea cât și redundanța au variat în același sens, dar cu viteze diferite, ceea 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.

Corpusul folosit cuprinde 21106 articole publicate între 2007 și 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 și 3 prezintă câștigurile de performanță obținute în cele trei experimente, iar cazul în care este obținut câștigul maximal este prezentat cu font mărit și îngroșat.

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 și 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%.

Prag	Acoperire	Redundanță	Variația acoperirii (%)	Variația redundanței (%)	Câștig (%)
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 – Câștigul obținut prin eliminarea cuvintelor neînrudite

Un al doilea test conține câștigul obținut prin ștergerea topicelor redundante din model. Un topic este considerat redundant dat fiind alt topic procesat anterior dacă similaritatea lor depășește cu un factor μ similaritatea medie a modelului. Rezultatele sunt semnificativ mai bune decât în primul experiment, o scădere de aproape 60% a redundanței fiind atinsă cu o pierdere de numai 26% a acoperirii.

μ	Topice sterse	Acoperire	Redundanță	Variația acoperirii (%)	Variația redundanței (%)	Câștig (%)
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 – Câștigul obținut prin eliminarea topicelor redundante

Metoda de mai sus poate fi îmbunătățită prin includerea în formula similarității între topice numai a nodurilor care nu sunt direct asociate unor cuvinte din topice. Limităm astfel riscul ca, dacă două topice conțin aceleași 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 creștere a câștigului de peste 2.5% față de metoda precedentă.

μ	Topice sterse	Acoperire	Redundanță	Variația acoperirii (%)	Variația redundanței (%)	Câștig (%)
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 – Câștigul obținut prin eliminarea topicelor redundante bazat doar pe noduri semnificative

5. Concluzii

În această lucrare am prezentat o metodă de evaluare a calităților unui model de topice complementar celor deja existente. Folosind această metrică am propus trei modalități de a îmbunătăți trăsăturile modelelor analizate. Rezultatele obținute – cu ameliorarea de peste o treime a trăsăturilor inițiale ne îndreptățesc să sperăm că aceste metode vor putea fi folosite ca post-procesare pentru algoritmi 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 lărgeste semnificativ aria de aplicabilitate a modelului în sine.

Mulțumiri

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

Rezultatele prezentate în acest articol au fost obținute cu sprijinul Ministerului Muncii, Familiei și 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. 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. 10^{ème} 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 ieeee 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 ieeee international conference on data mining, 2007.

An Intra and Inter-Topic Evaluation and Cleansing Method

Claudiu Mușat¹, Marian-Andrei Rizoiu², Ștefan Traușan-Matu³

¹ Universitatea Politehnica București
Bd. Splaiul Independenței, nr. 313 060032, București
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 Universitatea Politehnica București
Bd. Splaiul Independenței, nr. 313 060032, București

Institutul de Cercetări pentru Inteligența Artificială
Calea 13 Septembrie nr.13, București
E-mail: trausan@cs.pub.ro

Abstract. Topic modeling is a growing research field and novel ways of interpreting and evaluating results are necessary. We propose a method for evaluating and improving the performance of topic models generating algorithms relying on WordNet data. We first propose a measure for determining a topic model's fitness factoring in its broadness and redundancy. Then, for each individual topic, the amount of relevant information it provides, along with its most important words and related concepts are determined by defining a cohesion function based on the topic's projection on WordNet concepts. The model as a whole is improved by eliminating each topic's outliers with respect to the ontology projection. We define a inter topic ontology based distance and we further use it to investigate the impact of removing redundant topics from a model with regard to the overlap between topics' ontological projections. Clustering similar topics into conceptually cohesive groups is tried as an alternative to pruning less relevant topics. Results show that evaluating and improving statistical models with WordNet is a promising research track that leads to more coherent topic models.

Key words: Topic Modeling, WordNet, Ontology, LDA, Subtree, Evaluation, Improvement.