Dragoş VESPAN

EXTRAGEREA CUNOŞTINŢELOR DIN DOCUMENTELE ELECTRONICE PRIN TEXT MINING

Dragoş VESPAN

EXTRAGEREA CUNOŞTINŢELOR DIN DOCUMENTELE ELECTRONICE PRIN TEXT MINING

PRO UNIVERSITARIA
BUCUREŞTI 2014®

Descrierea CIP a Bibliotecii Naţionale a României
VESPAN, DRAGOŞ MARCEL
Extragerea cunoştinţelor din documentele electronice prin text mining / Dragoş Vespan. - Bucureşti : Pro Universitaria, 2014
Bibliogr.
ISBN 978-606-647-968-4

004

INTRODUCERE

Evoluţia internetului ca mijloc de transmitere a informaţiilor a dus atât la creşterea volumului resurselor de cunoştinţe disponibile on-line cât şi la diversificarea formelor şi formatelor de stocare şi transmitere a acestora: text, date, video, audio. Deşi restricţiile hardware în ceea ce priveşte spaţiul de stocare şi viteza de transmitere a datelor nu mai reprezintă o problemă, textul rămâne cea mai eficientă formă de prezentare a cunoştinţelor pe internet în comparaţie cu diversele formate audio, video sau multimedia.

Webul a fost proiectat ca un spaţiu informaţional cu scopul de a depăşi barierele mijloacelor clasice de comunicaţie şi de a permite maşinilor să ajute utilizatorii să comunice unii cu alţii. Cel mai important obstacol în realizarea acestui deziderat este reprezentat de faptul că informaţiile şi cunoştinţele existente pe internet sunt destinate exclusiv consumului uman. Acest obstacol poate fi depăşit prin utilizarea unor metode şi tehnici de reprezentare a cunoştinţelor existente în documentele text astfel încât acestea să poată fi automat achiziţionate şi procesate de către maşini.

Capitolul „Text mining – caracteristici şi domenii de aplicabilitate” prezintă evoluţia internetului ca mijloc de transmitere a informaţiilor, evidenţiind eterogenitatea şi diversitatea documentelor text existente pe web. Este prezentat stadiul cunoaşterii în text mining, punându-se accentul pe caracterul multidisciplinar, abordările existente şi domeniile de aplicabilitate ale acestui domeniu şi sunt descrise mai multe tipuri de arhitecturi funcţionale ale sistemelor de text mining.

Capitolul „Reprezentarea documentelor”analizează documentul ca element de bază al achiziţiei cunoştinţelorprin text mining. Sunt prezentate mai multe tehnici de reprezentare descriptivă a documentelor, cu accent pe utilizarea tabelelor ca modalitate de organizare a informaţiilor. Se definesc caracteristicile documentelor şise analizeazătehnici de reprezentare a documentelor în spaţiul vectorial.

Capitolul „Clasificarea automată a documentelor” abordează problematica achiziţiei cunoştinţelor prin utilizarea algoritmilor de clasificare. Sunt identificaţi şi analizaţi principalii algoritmi utilizaţi în clasificarea textelor. Se definesc

măsurile de evaluare a eficienţei clasificatorilor de texte şi sunt comparate performanţele algoritmilor de clasificare.

Capitolul „Webul Semantic – Caracteristici şi limbaje" abordează problematica reprezentării cunoştinţelor în contextul Webului Semantic. Sunt comparate modelele de distribuire a informaţiilor pe internet şi se utilizează cadrul de descriere a resurselor pentru a reprezenta datele preluate din documentele web ale ASE. Sunt descrise modele de procese pentru servicii oferite în cadrul ASE şi prezentate modalităţile de dezvoltare a ontologiilor prin text mining.

În capitolul „Reprezentarea cunoştinţelor din documentelor web în sistemul OntoDev", sunt prezentate contribuţiile personale ale doctorandului. Este elaborat un algoritm de reprezentare a documentelor web cu păstrarea informaţiilor legate de structura tabelară a acestora. Se analizează modul în care conceptele existente în cadrul ontologiilor sunt reprezentate în resursele web. Sunt prezentate tehnici deanaliză a logurilor unei organizaţii prin care se identifică modele de comportament ale utilizatorilor de internet.

1. TEXT MINING – CARACTERISTICI ŞI DOMENII DE APLICABILITATE

Evoluţia internetului şi creşterea numărului de utilizatori a impulsionat foarte mult dezvoltarea motoarelor de căutare şi identificarea unor soluţii cât mai eficiente de regăsire a informaţiilor pe internet. La începutul evoluţiei motoarelor de căutare, cea mai potrivită soluţie pentru analizarea paginilor web şi clasificarea lor în categorii, era utilizarea experţilor umani. Creşterea exponenţială a numărului de site-uri web, de la 10.000 la începutul anului 1995, la peste 650.000 la sfârşitul anului 1996(Gray, 1996)ajungându-se la 1 trilion de pagini indexate de Google la 25.08.2008 (Google, 2008), concomitent cu creşterea numărului de utilizatori de la 16 milioane în 2005 la 147 milioane în 1998 şi aproape 1,5 miliarde în iunie 2008(Stats, 2008), a făcut ca această soluţie să devină complet ineficientă. Lansarea motorului de căutare Google, care automatiza procesul de căutare, a fost o adevărată revoluţie, dovadă fiind faptul că astăzi peste 50% din căutările pe internet se fac prin utilizarea acestuia, în august 2007 fiind efectuate 31 de miliarde de căutări prin motorul Google, dintr-un total de 61 miliarde căutări la nivel mondial (Comscore, 2008).

Creşterea rapidă şi continuă a numărului de site-uri dar şi a numărului de utilizatori de internet va impune dezvoltarea motoarelor de căutare dincolo de analiza linkurilor şi a rankurilor de pagină, spre înţelegerea semanticii paginilor analizate. Algoritmii de text mining sunt esenţiali pentru dezvoltarea motoarelor de căutare în această direcţie, pornind de la regăsirea unor pagini care să corespundă unui şir de cuvinte cheie, până la regăsirea unor cunoştinţe care să corespundă nevoilor de informare ale utilizatorilor. O direcţie de dezvoltare importantă a motoarelor de căutare este aceea de a oferi suport şi pentru alte limbi decât cele de circulaţie internaţională. Google depune eforturi intense în acest sens, oferind utilizatorilor atât posibilitatea restricţionării căutărilor la o singură limbă (din peste 40 disponibile) dar şi traducerea automată a anumitor pagini în limba selecţionată, inclusiv română.

Motoarele curente de căutare pot analiza doar o mică parte a webului reprezentată de partea „vizibilă” a acestuia. O cantitate uriaşă de valoroase

informaţii ştiinţifice sau de altă natură se regăseşte în partea „invizibilă" a webului, reprezentată de paginile care nu au putut fi indexate de motoarele de căutare. Această parte a webului cuprinde documentele al căror format nu este recunoscut şi nu poate fi procesat de către motoarele de căutare, dar şi baze de date existente pe internet care nu pot fi indexate de acestea din motive de securitate sau de reprezentare a datelor. Mărimea webului „invizibil" nu poate fi precis estimată: Marcus Zillman estima la sfârşitul anului 2007 mărimea acestuia ca fiind de aproximativ 900 miliarde pagini, în comparaţie cu 20 de miliarde de pagini care erau indexate de motoarele de căutare la momentul respectiv (Zillman, 2007). În acest context, dezvoltatorii motoarelor de căutare trebuie să găsească soluţii şi metode adecvate pentru străpungerea barierelor ce separă webul „invizibil" de cel „vizibil".

Pe 28 iulie 2008 a fost lansat un nou motor de căutare denumit „Cuil" care promite o abordare diferită a modului de realizarea a căutărilor pe internet. Dacă motoarele de căutare clasice se bazează pe tehnici de analiză a legăturilor şi pe statistici de trafic, Cuil se orientează spre analiza contextului fiecărei pagini şi a conceptelor existente în fiecare interogare, după care organizează căutările similare în grupuri, sortându-le pe categorii. Cuil prezintă utilizatorilor rezultatele pe baza conţinutului documentelor şi nu a popularităţii acestora.

80% din informațiile companiilor sunt stocate în documente text(Tan, Text Mining: The state of art and the challenges, 1999). Aceste informații pot fi structurate, reprezentate prin tabele, atribute, numere generate de anumite aplicații, sau nestructurate, așa cum se regăsesc în e-mailuri, rapoarte sau diverse documente. Informațiile structurate au un caracter repetitiv fiind, în esență, numerice și, deci, mai ușor de identificat și prelucrat. Ele sunt rezultatul unor tranzacții cu caracter repetitiv (completarea unui ciclu de producție, vânzarea unui produs, încasarea unui ordin de plată) care generează date structurate despre activitățile la care se referă. Singurele elemente care diferă de la o tranzacție la alta sunt seturile de valori luate de tipurile de date.

În contrast, informațiile nestructurate nu au format, structură sau repetabilitate strict definite fiind, în general, scrise în limbaj natural liber. Scrierea unui e-mail nu este condiționată din punctul de vedere al formatului sau conținutului. Un e-mail poate fi scris de oricine, în orice format, utilizând orice limbă şi poate conţine orice subiect. La fel, există şi alte tipuri de documente text care conțin foarte multe informații şi cunoştinţe dar nu sunt condiționate de nici un fel de formatare: articole, cărţi, pagini web, rapoarte.

1.1. CARACTERISTICILE DOCUMENTELOR UNEI ORGANIZAŢII

Într-o organizaţie complexă cum ar fi, de exemplu, o universitate, se regăseşte o mare varietate de documente nestructurate ale căror caracteristici sunt neuniforme: cursuri, e-mailuri, contracte, rapoarte financiare.

Cursurile sunt documente de lungime variabilă ce pot conţine terminologie specifică domeniului în care se încadrează. Sunt scrise într-o singură limbă şi sunt identificate prin titularul de curs, specializarea şi anul la care acestea sunt predate.

E-mailurile sunt relativ scurte în comparaţie cu alte documente și pot conține atât informaţii personale cât şi informaţii legate de mediul de afaceri. De obicei sunt scrise într-o singură limbă, dar pot conţine şi texte scrise în diferite limbi. Sunt identificate prin adresa de e-mail a destinatarului, adresa de e-mail a expeditorului, data şi ora trimiterii.

Contractele abundă în terminologie juridică şi pot varia foarte mult în mărime. Sunt scrise într-un stil formal, putând fi traduse în mai multe limbi şi se identifică prin titulari de contract, obiectul contractului, data semnării, perioadă de valabilitate şi instituţia care îl legiferează.

Rapoartele financiare conţin informaţii legate de mediul de afaceri care includ datele financiare ale organizaţiei pe o anumită perioadă de timp, date referitoare la manageri şi acţionari precum şi la relaţionările organizaţiei cu alte organizaţii. Aceste rapoarte pot include şi grafice şi indicatori financiari dar şi previziuni referitoare la evoluţia financiară a organizaţiei. Sunt identificate prin organizaţia la care fac referire, perioada pentru care au fost făcute, organizaţia care le-a realizat şi data la care au fost realizate.

Acestea reprezintă doar câteva tipuri de documente text nestructurate existente într-o organizaţie educaţională, fiecare având propriile caracteristici.

Inmon şi Nesavich au propus o serie de caracteristici care evidenţiază lipsa de uniformitate a documentelor text nestructurate din interiorul unei organizaţii (Inmon & Nesavich, 2007): corelarea cu profilul de activitate al organizaţiei, modalitatea de redactare, terminologia utilizată, gradul de repetitivitate şi volumul datelor conţinute, numărul documentelor, perioada de păstrare, mediul de stocare, posibilitatea de actualizare.

Documentele pot fi direct sau indirect corelate cu profilul de activitate al organizaţiei. În cazul unei organizaţii educaţionale, cursurile, planurile de învăţământ, contractele şi rapoartele financiare pot fi considerate documente direct corelate cu profilul de activitate al acesteia. Rapoartele de evaluare a

resurselor umane, fişele angajaţilor sau e-mailurile sunt documente indirect corelate cu profilul de activitate al organizaţiei.

Modalitatea de redactare a documentelor poate fi formală sau informală. Exemple de documente informale sunt scrisorile, e-mailurile sau paginile web. Exemple de documente formale sunt contractele sau rapoartele financiare.

Terminologia utilizată în cadrul unui document se referă la tipul de limbaj utilizat. Unele documente tind să abunde în terminologie specifică unui anumit domeniu putând fi corect înţelese şi interpretate doar de experţi care posedă cunoştinţe anterior dobândite în domeniul respectiv. Rapoartele de cercetare tind să conţină foarte mulţi termeni ştiinţifici iar contractele conţin, de regulă, foarte mulţi termeni juridici. E-mailurile sau paginile web sunt, însă, mult mai generale.

Gradul de repetitivitate a datelor se referă la măsura în care se repetă atributele care reprezintă datele din cadrul documentelor. Paginile web ale site-urilor de comerţ electronic au, de regulă, un grad de repetitivitate ridicat, în timp ce e-mailurile au un grad scăzut de repetitivitate.

Mediul de stocare reprezintă suportul de prezentare a documentelor. Paginile web şi e-mailurile sunt, de regulă, stocate în format electronic în timp ce contractele sunt tipărite pe hârtie. În anumite circumstanţe, documentele pot fi tipărite pe hârtie din formatul electronic sau pot fi transpuse de pe hârtie în format electronic.

Volumul de date asociat documentelor poate varia de la câţiva kilobaiţi la mărimi de ordinul zecilor de megabaiţi în cazul e-mailurilor. În general, documentele tipărite pe hârtie au asociat un volum mic de date, în timp ce documentele în format electronic pot avea asociate volume mari de date.

Numărul de e-mailuri este, în general, mult mai mare decât numărul de contracte.

Perioada de păstrare se referă la cât timp vor fi păstrate documentele. Paginile web nu sunt păstrate, de obicei, pentru o perioadă foarte mare, în timp ce rapoartele anuale sunt păstrate pentru o perioadă foarte lungă de timp.

Posibilitatea de actualizare se referă la posibilitatea de schimbare a conţinutului unui document original, odată ce acesta a fost creat. E-mailurile nu sunt niciodată modificate, în unele ţări acest lucru fiind ilegal. Paginile web însă, sunt actualizate sau modificate cu regularitate.

Tipul de document	Corelare cu profilul de activitate	Modalitate de redactare	Terminologie	Grad de repetitivitate a datelor	Mediul de stocare	Volum de date asociat	Număr documente	Perioada de păstrare	Posibilitatea de actualizare
Pagini web	Directă	Informală	Variază	Variază	Electronic	Mediu	Mare	Variază	Da, Frecventă
E-mail	Indirectă	Informală	Generală	Mic	Electronic	Poate fi foarte mare	Foarte mare	Variază	Nu
Rapoarte financiare	Directă	Formală	Specifică	Ridicat	Hârtie, Electronic	Mic	Mic	Mare	Da, Uneori
Cursuri	Directă	Informală	Variază	Mic	Hârtie, Electronic	Mic	Mare	Mare	Da
Rapoarte de cercetare	Directă	Formală	Specifică	Mic	Hârtie, Electronic	Mic	Mare	Mare	Da
Contracte	Directă	Formală	Specifică	Foarte mic	Hârtie	Mic	Moderat	Foarte mare	Da, Ocazional
Fişe angajaţi	Indirectă	Formală	Generală	Ridicat	Hârtie, Electronic	Mic	Mare	Mare	Da

Tabelul 1. Caracteristicile documentelor text existente într-o organizaţie educaţională