Universitatea „Lucian Blaga” din Sibiu
Facultatea de inginerie “Hermann Oberth”
Catedra de calculatoare şi automatizări

Extragerea unor trăsături relevante din masive de date ne-structurate semantic

Referat de doctorat nr. 3

Titlul tezei: „Contribuții la extragerea automată de cunoştințe din masive de date”

Autor:
asist. drd. ing. Daniel MORARIU

Coordonator ştiințific:
prof. univ. dr. ing. Lucian N. VINTAN

SIBIU, 2006

1 Rezumat

In ultimii ani, datorită impactului calculatoarelor în toate domeniile, cantitatea de informaţii în format text devine tot mai mare. Recunoaşterea şi clasificarea automată a acestora devine tot mai necesară. Acestea sunt considerate ca fiind date semi-structurate deoarece ele sunt mai degrabă nestructurate decât complet structurate. Ele conţin o mică parte de organizare a informaţiilor, care în majoritatea cazurilor nu este completată de către autorul documentului.

Tehnicile tradiţionale pentru recunoaşterea informaţiilor devin inadecvate pentru aplicaţiile de căutare în depozite de date. De obicei, doar o mică parte din toate documentele disponibile sunt relevante pentru utilizator la un moment dat. Fără a şti ce conţine documentul, este dificil să formulăm interogări pentru analiza şi extragerea informaţiilor necesare. Pentru a putea extrage doar acele documente relevante, utilizatorii au nevoie de componente care să îi ajute să poată compara documentele, cum ar fi eficienţa şi relevanţa lor în raport cu o anumită interogare, sau pentru a putea găsi şabloane pentru a le putea indexa şi regăsi mai uşor.

Un număr impresionant de documente se găsesc pe WEB, iar utilizatorul are nevoie de componente de clasificare automată a acestora pentru a le putea gestiona. Gestiunea lor a devenit o problemă importantă în ultima perioada. Devine esenţială existenţa unor programe de organizare automată a documentelor în clase pentru a facilita analiza şi recunoaşterea acestora. O posibilă procedură generală pentru rezolvarea acestei probleme este alegerea unei mulţimi de documente pre-clasificate (mult mai mică în comparaţie cu documentele existente) şi considerarea acestora ca fiind documente de antrenament. Mulţimea de antrenament este apoi analizată pentru a putea extrage o schemă de clasificare. Această schemă este finisată utilizând o mulţime de documente de test. După aceasta, schema astfel obţinută poate fi utilizată pentru clasificarea celorlalte documente existente. Analiza clasificării decide care mulţime de perechi de atribut-valoare are o putere de discriminare mai mare în determinarea claselor. O metodă eficientă pentru clasificarea documentelor este de a exploata clasificările bazate pe asocieri, unde documentele sunt clasificate pe baza unei mulţimi de asocieri şi a frecvenţei de şabloane. Metodele de clasificare pe baza asocierilor respectă următoarele etape: (1) cuvintele cheie şi termenii pot fi extraşi prin tehnici de recunoaştere a informaţilor şi tehnici de analiză a asocierilor simple; (2) ierarhiile de concepte de cuvinte cheie sau termeni pot fi obţinute utilizând termenii claselor disponibile, încrederea în cunoştinţele expertului sau unele sisteme de clasificare pe bază de cuvinte cheie. Documentele din mulţimea de antrenament pot fi de asemenea clasificate în ierarhi de clase. Metodele de minerit pe baza asocierii termenilor pot fi apoi aplicate pentru a descoperi mulţimi de termenii de asociere care pot fi utilizaţi pentru a maximaliza diferenţele dintre două clase de documente. Acestea produc o mulţime de reguli de asociere pentru fiecare clasă de documente. Astfel de reguli de asociere pot fi ordonate pe baza frecvenţei lor de apariţie şi a puterii de discriminare, utilizate apoi pentru clasificarea de noi documente.

Clasificarea documentelor text este un proces foarte general incluzând o mulţime de cerinţe care trebuie realizate pentru rezolvarea problemei. Fiecare dintre aceste cerinţe are o influenţă foarte mare asupra rezultatului final al clasificării. Acest proces poate fi privit ca un flux de date în mai multe etape. În care, în fiecare etapă, văzută ca o cutie neagră, se primesc informaţii, se procesează şi se transferă mai departe. Fiecare etapă din acest flux de date poate avea unul sau mai mulţi algoritmi ataşaţi. La un moment dat, pentru fiecare etapă putem alege unul dintre algoritmii ataşaţi

cu diferiți parametrii de intrare. În primul meu referat am prezentat câteva părți ale fluxului; în acest referat am îmbunătățit anumite părți deja existente şi am finalizat fluxul de date.

Astfel, în primul referat am prezentat unele tehnici de pre-procesare a documentelor text. În special am prezentat pre-procesarea bazei de date Reuters-2000. Am continuat cu o scurtă introducere a procesării paginilor web, axându-mă pe diferențele dintre acestea şi documentele text. În această etapă datele de intrare sunt reprezentate prin documente text (fişiere text sau pagini web) şi datele de ieşire sunt reprezentate de vectori de trăsături care caracterizează fiecare document din mulțimea de documente. Trăsăturile din aceşti vectori sunt de fapt frecvențele de apariție ale cuvintelor în documentul respectiv. Această etapă conține un modul de eliminare a cuvintelor irelevante din punct de vedere semantic (stop-words), un modul de extragere a rădăcinii cuvintelor pentru limba engleză şi un modul de numărare a cuvintelor. Datorită dimensiunii impresionante a vectorilor rezultați această etapă continuă cu o etapă de selectare a trăsăturilor relevante din aceşti vectori. Astfel, în cel de-al doilea referat am prezentat trei metode de selecție a trăsăturilor relevante: selecția aleatoare, selecția bazată pe câştigul informațional şi selecția bazată pe vectori suport. O nouă metodă de selecție a trăsăturilor caracteristice relevante bazată pe algoritmi genetici este prezentată şi implementată în acest referat.

În cel de-al doilea referat am prezentat în detaliu algoritmul utilizat pentru clasificarea documentelor, bazat pe vectorii suport (SVM) şi învățarea utilizând nuclee. Acolo m-am axat în general pe procesul de clasificare a documentelor. De asemenea am prezentat o adaptare a acestui algoritm pentru a putea fi folosit şi în contextul învățării nesupervizate. Apoi am prezentat o metodă nouă de corelare a parametrilor nucleelor şi îmbunătățirile aduse aplicației noastre în comparație cu o implementare a acestei tehnici des folosită în literatură, LIBSVM. În acest referat vom prezenta rezultatele care justifică corelarea parametrilor din nuclee precum şi metodologia de alegere a acestor corelații.

În ultima etapă din acest flux prezentăm o tehnică de dezvoltare a unui meta-clasificator pentru a îmbunătății acuratețea clasificării. Clasificatorii de bază din acest meta-clasificator utilizează tehnica vectorilor suport. Această metodă este prezentată în acest referat şi încearcă să obțină rezultate mai bune decât oricare dintre clasificatorii de bază componenți.

Acest referat conține unele contribuții care îmbunătățesc fluxul clasificării făcându-l mult mai fiabil pentru lucrul cu mulțimi mari de date. Una dintre acestea este abilitatea aplicației noastre de a lucra cu mult mai multe documente. Rezultatele prezentate până acum sunt obținute pe o dimensiune mică a setului de date în comparație cu toate datele existente în baza de date. Utilizarea tuturor datelor existente ar face procesul de învățare ineficient din punct de vedere al timpului. În acest referat am prezentat o metodologie care face ca aplicația noastră să fie capabilă să lucreze cu o dimensiune impresionantă a datelor de intrare şi cu pierderi cât mai mici din punct de vedere al acurateței clasificării. Această metodologie are două obiective majore antagonice: o dimensiune mare a datelor de intrare şi un timp de învățare minim pentru o clasificare optimă.

Pentru a avea date cât mai realiste în cercetarea noastră, ar trebui să prezentăm media rezultatelor obținute pe mai multe perechi de mulțimi de date de antrenare-testare. Pentru fiecare pereche de mulțimi ar trebui calculată separat acuratețea clasificării. În toate rezultatele am prezentat acuratețea obținută pe o singură pereche de mulțimi de antrenare-testare. În ultimul capitol al acestui referat am prezentat o parte din rezultatele obținute şi pe o altă pereche de mulțimi. Am făcut aceste experimente pentru a vedea dacă rezultatele noastre sunt sau nu apropiate de rezultatele ce s-ar obține în urma calculului mediei. Aceste rezultate ne vor furniza nivelul de încredere al aplicației noastre din punct de vedere al acurateței prin calcularea mediei acurateților obținute pe mai multe perechi de mulțimi de antrenare–testare.

Experimentele prezentate în acest referat sunt efectuate folosind colecţia de documente Reuters 2000, care are o dimensiune de 948 Mb de ştiri în format comprimat. Colecţia include un număr de 806791 ştiri publicate de agenţia de presă Reuters în perioada 20.07.1996 - 19.07.1997. Statistic vorbind articolele au 9822391 paragrafe, 11522874 propoziţii şi au 310033 rădăcini distincte de cuvinte. Documentele sunt clasificate de Reuters după trei categorii distincte. După *Regiunea* la care se referă articolul, existând 366 regiuni distincte. După un *Cod Industrial* în funcţie de ramura industrială la care se referă articolul, existând un număr de 870 coduri industriale. Şi după *Categorie* care este propusă de către Reuters, existând 126 categorii distincte. Douăzeci şi trei dintre aceste categorii nu au nici un document ataşat. Datorită acestei dimensiuni imense a datelor de intrare am prezentat rezultatele doar pe o submulţime a acestora. Astfel din toate aceste documente am selectat doar acelea pentru care codul industrial este „System software”. Am obţinut un număr de 7083 documente care pot fi reprezentate folosind 19038 trăsături distincte şi 68 categorii. Am reprezentat documentele ca şi vectori de frecvenţe de termeni aplicând o filtrare a cuvintelor de legătură prin folosirea unei liste standard de 510 cuvinte (pentru limba engleză) şi am extras rădăcina cuvintelor rămase. Din aceste 68 de categorii rezultate am eliminat acele categorii care sunt slab sau excesiv reprezentate. Astfel am eliminat acele categorii care conţin mai puţin de 1% din numărul total de documente (7083). De asemenea am renunţat la acele categorii care conţin mai mult de 99% din documente. Eliminarea a fost necesară deoarece păstrând aceste categorii există riscul ca algoritmul nostru de învăţare să înveţe doar o singură categorie şi să clasifice toate documentele în acea categorie, ignorând celelalte categorii existente. După acest pas am obţinut 24 de categorii diferite şi 7053 documente care au fost împărţite aleator în setul de antrenament (4702) şi cel de test (2531). În partea de extragere a trăsăturilor am luat în considerare atât întreg conţinutul articolului cât şi titlul articolului.

Capitolul doi conţine experimentele care au condus la alegerea corelaţiilor dintre parametrii nucleelor. Capitolul trei conţine o nouă metodă de selecţie a trăsăturilor caracteristice bazată pe algoritmi genetici utilizând o funcţie de calcul a acurateţei bazată pe tehnica vectorilor suport. Următorul capitol încheie procesul de clasificare automată a documentelor prin prezentarea câtorva metode pentru implementarea unui meta selector care să conducă la îmbunătăţirea acurateţei finale a clasificării. Capitolul cinci prezintă influenţa unei mari cantităţi de date de intrare asupra algoritmului nostru. Tot aici prezentăm o strategie care permite algoritmului nostru să lucreze cu date de intrare mari într-un timp mic şi cu pierderi minore. În capitolul 6 prezentăm câteva rezultate obţinute pe o altă împărţire a setului de date în date de test şi date de antrenament. Ultimul capitol prezintă şi dezbate cele mai importante rezultate obţinute şi propune câteva tendinţe de viitor.

Mulțumiri

În încheierea acestei introduceri aş vrea să exprim sincerele mele mulțumiri către conducătorul meu de doctorat prof. dr. ing. Lucian VINȚAN pentru coordonarea ştiințifică pe perioada de pregătire a acestui doctorat, pentru discuțiile extrem de stimulative pe care le-am avut şi pentru întreg sprijinul acordat. De asemenea, rămân îndatorat pentru sprijinul profesional competent acordat mie cu generozitate, încă de la începuturile perioadei de pregătire prin doctorat, de către profesorii: mat. Ioana MOISIL, ing. Boldur BĂRBAT, ing. Daniel VOLOVICI, ing. Dorin SIMA şi ing. Macarie BREAZU.

În acelaşi timp aş dori să mulțumesc firmei SIEMENS AG, CT IC MUNCHEN, Germany, în special vicepreşedintelui Dr. h. c. mat. Hartmut RAFFLER, pentru sugestiile profesionale foarte folositoare şi pentru suportul financiar acordat pe parcursul acestui program de doctorat. Aş dori să mulțumesc tutorelui meu de la firma SIEMENS, Dr. Volker TRESP, Senior Principal Research Scientist în Neural Computation, pentru sprijinul ştiințific acordat şi pentru îndrumarea în acest imens şi dificil domeniu al cercetării. De asemenea aş dori să mulțumesc d-lui Dr. Kai Yu de la aceeaşi companie pentru informațiile furnizate în dezvoltarea ideilor mele. La sfârşit aş dori să mulțumesc tuturor acelora care m-au ajutat în pregătirea acestui referat.