Extragerea şi procesarea cunoştinţelor din date structurate şi nestructurate

Teză de abilitare

Mihaela Dînşoreanu

Rezumat

Această teză de abilitare prezintă activitatea profesională a candidatei precum şi activitatea ştiinţifică care a fost desfăşurată după susţinerea tezei de doctorat la Universitatea Tehnică din Cluj-Napoca în 9.07.2004, urmată de confirmarea titlului de doctor prin Ordinul Ministerului Educaţiei şi Cercetării nr. 445/2.08.2004. Activitatea de cercetare din domeniul tezei a avut ca şi domeniu proiectarea sistemelor distribuite cu agenţi cooperanţi. Modelele şi metodele propuse au fost aplicate pe un studiu de caz în cadrul unui sistem de învăţare la distanţă. După finalizarea tezei de doctorat candidata a rămas interesată de domeniul inteligenţei artificiale, participând ca şi membru în câteva proiecte naţionale în domeniu. În cadrul acestora, s-au analizat şi propus soluţii de data mining (analiza datelor) aplicate în domeniul medical, soluţii de extragere şi reprezentare a cunoştinţelor aplicate în contextul documentelor arhivistice redactate în limbaj natural. Mai recent, preocupările candidatei s-au orientat înspre aspecte ale reprezentării şi integrării semantice a datelor, extragerea cunoştinţelor prin aplicarea diferitelor tehnici de învăţare automată şi procesarea cunoştinţelor. Domeniile adresate, realizările teoretice şi practice sunt rezumate grafic în cadrul tezei.

a. *Reprezentarea şi integrarea semantică a datelor*.

În acest context, candidata a analizat şi propus soluţii pentru integrarea semantică a datelor importate din diferite surse de date structurate (de ex. baze de date relaţionale, fişiere XML, fişiere CSV), proiectarea automată a structurilor de date unificate precum şi proiectarea automată a proceselor ETL corespunzătoare. Aceste soluţii se bazează pe meta-datele extrase din sursele de date, relaţii semantice (sinonimie) existente în WordNet precum şi algoritmi de similaritate lexicală (de ex. Jaro-Winkler) pentru a construi o resursă adnotată semantic utilizată în cadrul algoritmului de fuziune semantică. Obiectivul urmărit este de a produce informaţie cât mai exactă pentru integrarea datelor eterogene şi generarea dinamică a proceselor ETL. S-au realizat experimente pe diferite dimensiuni de seturi de date pentru a evalua scalabilitatea şi performanţa soluţiei.

b. *Extragerea cunoştinţelor*.

Cele mai recente abordări ale candidatei în extragerea cunoştinţelor s-au concentrat pe surse de date nestructurate, şi anume text scris în limbaj natural. Unul din obiectivele urmărite a fost identificarea topicii în vederea definirii contextului pentru a furniza recomandări senzitive la context. Un alt obiectiv urmărit a fost identificarea opiniilor în text. Pentru aceste obiective s-au explorat mai multe dimensiuni ale soluţiilor: supervizate vs. nesupervizate, dependente de domeniu vs. independente de domeniu, soluţii specifice unei limbi (engleză, română) şi posibilităţi de generalizare. Am propus un model unificat de topici bazat pe Latent Dirichlet Allocation (LDA) pentru a defini contextul precum şi un set de metrici de distanţă contextuală şi comportamentală pentru a măsura distanţa între conţinut (recomandări) şi context. Pentru problema identificării opiniei am propus un flux de procesare precum şi o abordare de învăţare bazată pe un set original de meta-trăsături ale elementelor de text (combinaţii de părţi de vorbire ca şi bigrame). Soluţia supervizată propusă a produs rezultate bune pentru documente în limba engleză şi am propus o adaptare pentru limba română bazată

pe un set de reguli ale gramaticii limbii române. Pentru aceeași problemă a extragerii de opinii am propus și o soluție nesupervizată bazată pe un algoritm de propagare dublă. Soluția a inclus un set extins de reguli pentru dubla propagare precum și un set minimal de cuvinte ”sământă” foarte generale (2 cuvinte). Soluția s-a dovedit a performa comparabil cu alte soluții documentate în literatură.
Un alt aspect pe care l-am considerat a fost extragerea de opinii din conținut publicat în rețele sociale cum ar fi Twitter. Acest tip de conținut pune probleme specifice: în multe cazuri nu este scris într-un limbaj corect, lungimea este limitată la 140 caractere și conține de regulă limbaj argotic, abrevieri, simboluri speciale, emoticoane etc. În mod normal o opinie include elemente ca și subiectul opiniei (entitatea despre care se emite opinia), deținătorul opiniei și opinia însăși. În cazul conținutului din rețele sociale, deținătorul opiniei este implicit utilizatorul curent dar, deși de cele mai multe ori mesajul include o polaritate a sentimentului, nu este întotdeauna implicat un subiect. Datorită conținutului extrem de divers existent, soluțiile supervizate nu performează acceptabil în clasificarea lor de cele mai multe ori. De aceea am investigat soluții nesupervizate, independente de un set de antrenare, pentru clasificarea polarității conținutului din rețeaua socială Twitter. Aceste soluții au utilizat resurse lexicale și semantice existente și au fost testate pe patru seturi de date de evaluare utilizate în competiția SemEval din 2013 și 2014. Fluxul de procesare propus, precum și regulile propuse pentru clasificare au performat comparabil cu alte abordări prezentate în competiție.
c. *Procesarea cunoștințelor*
Am investigat metode de a deriva informație suplimentară din cunoștințele extrase. Astfel, am propus metode pentru detectarea contradicțiilor în opinii, pentru detectarea comunităților bazate pe opinii și pentru analiza fluxurilor de date în domeniul financiar. Primele două metode se bazează pe rezultatele soluțiilor de identificare a opiniilor. Metoda de detectare a contradicțiilor identifică opiniile exprimate asupra aceluiași subiect evaluând distanța între polaritățile opiniilor exprimate. Metoda identifică de asemenea forme de bază ale negației pentru detectarea contradicțiilor. Soluția de detectare a comunităților are ca obiectiv identificarea deținătorilor de aceleași opinii. Această soluție integrează și informații din rețele sociale și utilizează două metode de integrare a datelor multidimensionale: Network Integration și Partition Integration. Am propus și analizat un set de 9 funcții de similaritate pentru obiective multiple care au condus la următoarele concluzii: funcțiile liniare performează slab pentru seturile de date cu obiective multiple și funcțiile care calculează similarități medii au o mai mare rezistență la zgomot.
Toate soluțiile propuse și rezultatele obținute sunt detaliate în capitolul de Realizări științifice.

Direcțiile de cercetare viitoare ale candidatei se orientează spre metodele specifice seturilor mari de date (Big Data) aplicate în contextul soluțiilor de rețele de lucruri (IoT). Problemele pe care ne propunem să le adresăm sunt:

- Alinierea semantică a meta-datelor care descriu datele emise de sursele de date eterogene, dinamice cum sunt echipamente, senzori, aparate electrocasnice.
- Dezvoltarea de modele eficiente de reprezentare și metode de procesare a datelor cu accent pe fluxurile de date.
- Integrarea tehnicilor de învățare automată și a resurselor semantice în metodele de analiză a datelor/fluxurilor de date pentru a obține extragere de cunoștințe senzitivă la context.

Detaliile despre direcțiile de cercetare viitoare se regăsesc în teză în capitoul Planuri de dezvoltare științifică.