HABILITATION THESIS

RESEARCH AND CONTRIBUTION IN ADVANCED DATA ANALYSIS AND KNOWLEDGE DISCOVERY

Summary

Associate Professor PhD. Mirela DANUBIANU

Faculty of Electrical Engineering and Computer Science
Computers Department

2017

Executive Summary

An undeniable reality of our days is that during the last years we have witnessed an explosive growth in the volume of data stored in databases or in data warehouses, growth generated mainly by the rapid progress in data generation, acquisition and storage technologies. This is the fact, which induced idea that these huge amounts of data hide information difficult to be detected at first sight or by using classical query tools. For this reason the interest in finding some methods for automatic or semi-automatic extraction of information from large data sets increased. Thus, the process of knowledge discovery in databases and data mining emerged. With time, data mining and knowledge discovery from data became crucial research areas with important applications in science, engineering, medicine, economics and education.

This habilitation thesis presents the candidate's professional activities and scientific achievements acquired after getting the doctoral degree in computer science, in September 2006, at "Ştefan cel Mare University" from Suceava, until now. Doctoral research aimed to contribute to the development of data mining and knowledge management methods and techniques. For doctoral degree we addressed mainly the issue of association rules discovery in various contexts, and we applied the proposed algorithms to find association rules, and to interpret their meanings for a company providing public utilities. Subsequently, naturally, we continued the research on data mining methods and techniques. This time we paid attention to classification, and more recently, to current trends in Big Data organization and analytics, and to the issues raised by the emerging data science.

The thesis is divided into three parts. Presented results are reproduced from works already published or accepted for publication in the very near future, and addresses both theoretical and applied research in different fields such as tourism, healthcare, education or social issues.

The first part is a short review of author's contributions and of their impact.

Chapter 1 presents the context in which the research was conducted, a summary of the main contributions, a synthesis of our works citations and the associated Hirsh indices, calculated in three databases recognized in the field - ISI Web of Science, Scopus and Google Scholar. They are also mentioned the achieved distinctions, the main grants and projects in which the author acted as project manager, as responsible from a partner or as technical expert, and as member of the implementation team. Finally we mention some activities performed for the benefit of the academic community, such as participation in scientific or organizing committees for various conferences, moderating sections at scientific events, conducting reviews of papers submitted for presentation and publication in various journals or representative of the university in relationship with prestigious companies in the field.

Chapter 2 presents the evolution of modern organization and data processing. First we presented the concept of the data warehouse. Then we described the process of knowledge discovering in data. It has been emphasized that it is a complex process consisting of several stages, described CRISP-DM model, whose central step is data mining. Finally it explores the relationship between data mining and decision making processes.

In the next chapters results of applied research carried out in this period are presented.

Chapter 3 deals with the issues raised by the use of data warehousing and data mining techniques in the tourism area. Starting from traditional information systems used in the field,

it is argued the opportunity to develop a data warehouse for the Romanian tourism industry and a framework for its implementation is presented. Also, as a novelty for 2006-2008 years is a study on the possibility of improving customer relationship management in the hospitality industry by using data mining techniques.

Chapter 4 deals with a topic that we researched for a long time this period, namely, the use of advanced information systems for assisting and improving personalized therapy of speech disorders. This research was financially supported by two projects: "System for personalized therapy of speech disorders TERAPERS", contract 56 CEEX-II-03 / 27.07.2006, CEEX Research Excellence 2006, code MEC 9714 and "Progress and development through postdoctoral research and innovation in engineering and applied sciences - Pride "; contract no. POSDRU / 89 / 1.5 / S / 57 083, subcontract 8205 / 17.06.2010. We started from the presentation of TERAPERS system, which was intended to assist speech therapists to design and implement personalized therapies to treat children suffering from dyslalia. As a result of the successful use of this system by specialists from Regional Center for Speech Therapy of Suceava, the idea to optimize the therapeutically path and to target efforts to correct speech disorders so as to reduce both treatment costs and the time during these problems are corrected or improved emerged. It has been shown that data mining can provide useful information for this purpose. Therefore, we designed Logo DM a dedicated data mining system aiming to allow the analysis of data collected through TERAPERS. A special attention was paid to the preparatory phase of the data sets to be suitable for classification algorithms respectively for association rules discovery. We treated problems of data integration, data transformation and data dimensionality reduction. For the last problem we proposed a combined filter-wrapper architecture. Also related to this issue we have shown that a possibility to improve the therapy performance can be achieved by combining data mining techniques with a knowledge-based system. Last but not least, it should be noted that by the nature of the extracted information, which often can be unexpected and surprising, data mining can be a source of continuous improvement of speech therapists skills and knowledge.

Chapter 5 covers the application of data mining in higher education. Recently, in academia it started the extensive use of learning management systems that allow massive data collection both about students' outcomes, and about their learning behavior. The analysis of these data may allow predicting students' results for a certain course depending on some aspects such as: the results obtained in prerequisites courses or the evaluation type and structure. It can also lead to interesting conclusions on how it should be approached the teaching / learning process, how should be properly structured the assessments, which are the items which must be emphasized and, why not, it can get even a curriculum review.

Chapter 6 addresses a maximum actuality issue - the migration. Concerning this, we conducted a research on how data mining techniques can allow us to predict the number of migrants in a timeframe. We made a case study, in compliance with CRISP-DM model stages, that aims to predict the intention of Roma individuals who migrated to France and settled in Rennes, to return to Romania in a period of five years. The raw data were collected from some questionnaires completed by them during a sociological survey.

Chapter 7 can be thought of as a journey to new frontiers on the data organization, processing and analysis. We addressed issues related to the concept of Big Data and Big Data analysis applied to academic environment and social networks. The issues presented are the beginnings of research on Big Data and NoSQL we intend to achieve in the future.

The second part presents some directions for academic career development. It targets above all to underline the future research directions.

The third part contains the list of works referred in the paper.

Rezumat

O realitate de necontestat a zilelor noastre este faptul că în ultimii ani am asistat la o creştere explozivă a volumului datelor stocate în baze de date sau depozite de date, creştere generată, în principal, de progresul rapid în domeniul producerii şi achiziţiei digitale a datelor, precum şi al tehnologiilor de stocare. Este un fapt care a indus ideea că aceste cantităţi uriaşe de date ascund informaţii dificil de detectat la prima vedere sau prin folosirea instrumentelor clasice de interogare. Ca urmare a crescut interesul pentru găsirea unor metode de extragere automată sau semiautomată a cunoştinţelor din seturi mari de date şi aşa au apărut procesele de descoperire a cunoştinţelor din date şi explorarea datelor.

Cu timpul, explorarea datelor şi descoperirea cunoştinţelor din date au devenit domenii fundamentale de cercetare, cu aplicaţii importante în ştiinţă, inginerie, medicină, economie şi educaţie.

Teza de abilitare prezintă activitatea profesională a candidatului și realizările sale științifice obținute după dobândirea titlului de doctor în știința calculatoarelor, în septembrie 2006 în cadrul Universității "Ștefan cel Mare" din Suceava, până la data depunerii acesteia. Cercetarea doctorală a vizat contribuții la dezvoltarea metodelor și tehnicilor de explorare a datelor și managementului cunoștințelor. In teză am abordat cu precădere problematica descoperirii regulilor de asociere în diferite contexte și am aplicat algoritmii propuși pentru găsirea regulilor de asociere și interpretarea informațiilor obținute pentru o companie ce lucrează în domeniul furnizării de utilități publice. Ulterior, în mod natural, am continuat activitatea de cercetare în direcția metodelor și tehnicilor de explorare a datelor, acordând, de data aceasta, atenție deosebită clasificării, și mai recent, tendintelor actuale de explorare a datelor în organizare de tip Big Data precum și a problemelor ridicate de nou-apăruta știință a datelor.

Teza este structurată pe trei părți. Rezultatele prezentate sunt reproduse din lucrări deja publicate sau acceptate spre publicare în viitorul foarte apropiat și adresează atât probleme teoretice cât şi cercetări aplicative în domenii diferite precum turismul, îngrijirea sănătăţii, educaţia sau probleme de ordin social.

Prima parte este destinată trecerii în revistă a contribuțiilor autorului și a impactului acestora.

Capitolul 1 prezintă contextul în care s-au desfășurat cercetările, un sumar al principalelor contribuții, o sinteză a citărilor și indicii Hirsh asociați, calculaţi în cele trei baze de date recunoscute în domeniu- ISI Web of Science, Scopus și Google Scholar. Sunt amintite distincțiile primite, principalele granturi și proiecte în care autorul a avut calitatea de director de proiect, responsabil din partea unui partener, expert tehnic sau membru în echipele de implementare. In final, sunt menționate câteva activități prestate în folosul comunității academice, precum participări în comitetele științifice sau de organizare a diferitelor conferințe, moderarea de secțiuni în cadrul manifestărilor științifice, efectuarea de recenzii pentru lucrări propuse spre prezentare și publicare în diferite jurnale sau reprezentant al universității în relația cu firme de prestigiu în domeniu.

Capitolul 2 prezintă o evoluție a sistemelor moderne de organizare și prelucrare a datelor. Pentru început este prezentat conceptul de data warehouse. Apoi este descris procesul de descoperire a cunoştinţelor din date. S-a accentuat faptul ca acesta este un proces complex alcătuit din mai multe etape, şi s-a descris modelul CRISP-DM, a cărui etapă centrală este data mining. In final se analizează relaţia dintre data mining şi procesele decizionale.

În următoarele capitole sunt prezentate rezultate ale cercetărilor aplicative desfașurate în perioada menționată.

Capitolul 3 tratează probleme ridicate de introducerea tehnicilor de data warehousing și data mining în domeniul turismului. Plecând de la sistemele informatice clasice utilizate în domeniu, se argumentează oportunitatea dezvoltării unui data warehouse la nivelul sectorului turismului românesc și se prezintă un cadru de lucru pentru realizarea acestuia. De asemenea, ca o noutate pentru perioada anilor 2006-2008 se face un studiu privind posibilitatea îmbunătățirii managementului relațiilor cu clienții în industria hotelieră prin utilizarea tehnicilor de data mining.

Capitolul 4 se ocupă de un subiect asupra căruia am făcut cercetări o lungă perioadă în intervalul vizat de teză, și anume, utilizarea sistemelor informatice avansate pentru terapia personalizată a tulburărilor de vorbire. Aceste cercetări au fost susținute financiar de două proiecte: „Sistem pentru terapia personalizata a tulburarilor de expresie lingvistica TERAPERS", Contract: 56–CEEX–II–03/27.07.2006, Programul CEEX Cercetare de Excelenta 2006, cod MEC 9714 şi "Progres si dezvoltare prin cercetare si inovare post-doctorala in inginerie si stiinte aplicate – PriDE"; Contract nr. POSDRU/89/1.5/S/57083, Subcontract 8205/17.06.2010. Am pornit de la prezentarea sistemului TERAPERS, care a avut ca scop asistarea experților logopezi în proiectarea și implementarea de terapii personalizate pentru tratarea copiilor care suferă de dislalie. Ca urmare a succesului înregistrat în folosirea acestui sistem de către specialiștii Centrului Regional de Terapie Logopedică Suceava, a apărut ideea de a optimiza traseul terapeutic și de a ținti eforturile de corectare a problemelor de vorbire astfel încât să se reducă atât costurile tratamentului cât și timpul în care aceste problem sunt corectate sau ameliorate. S-a demonstrat că explorarea datelor poate furniza informații utile în acest sens. Ca urmare, s-a proiectat un sistem dedicat de data mining – Logo-DM cu scopul de a permite explorare datelor colectate prin sistemul TERAPERS. O atenție deosebită a fost acordată etapei de pregătire a datelor pentru a se constitui seturi potrivite pentru algoritmii de clasificare respectiv de descoperire a regulilor de asociere. Am tratat probleme de integrare, de transformare a datelor și de reducere a dimensionalității acestora. Pentru aceasta ultimă problemă am propus o arhitectură combinată *filter-wrapper*. Tot legat de acest subiect am arătat că o posibilitate de îmbunătățire a performanțelor terapiei o poate constitui combinarea tehnicilor de data mining cu un sistem bazat pe cunoștințe. Nu în ultimul rând, trebuie subliniat faptul că, prin natura informațiilor extrase, care de multe ori pot fi neașteptate și surprinzătoare, data mining poate constitui o sursă de perfecționare continuă a specialiștilor logopezi.

Capitolul 5 vizează utilizarea explorării datelor în domeniul învățământului superior. In ultimul timp, la acest nivel au inceput să se folosească pe scară largă sisteme informatice pentru gestiunea procesului de învățare care au permis colectarea unor masive de date referitoare la rezultatele obţinute, dar şi la comportamentul studenților legat de procesul de învățare. Analiza acestor date poate permite predicția rezultatelor unui student la o disciplină, funcție de rezultatele obținute la disciplinele anterioare, de tipul sau structura evaluării. De asemenea poate conduce la concluzii interesante privind modul în care trebuie abordat procesul de predare/învățare, cum trebuie structurată evaluarea și care sunt elementele asupra cărora trebuie insistat și, de ce nu, se poate ajunge chiar la o revizuire a curriculei.

Capitolul 6 abordează o problemă de maximă actualitate – migrația. Legat de aceasta, am realizat o cercetare asupra modului în care tehnicile de data mining ne pot permite să facem o predicție a numărului de migranți într-un interval de timp. Am făcut un studiu de caz, respectând etapele modelului CRISP-DM, referitor la predicția intenției indivizilor de etnie romă, care au migrat în Franța și s-au stabilit la Rennes, de a reveni în Romania într-un interval de 5 ani. Datele brute au fost colectate de pe chestionarele completate de acestia cu ocazia unei cercetări sociologice.

Capitolul 7 poate fi gândit ca o incursiune către noi frontiere privind organizarea, prelucrarea și analiza datelor. Am abordat aspecte legate de conceptul Big Data și analize Big Data în domeniul academic și al rețelelor sociale. Aspectele prezentate constituie începuturile unor cercetari privind Big Data si NoSQL pe care intenționăm să le realizăm în viitor.

Partea a doua prezintă câteva direcții de dezvoltare a carierei academice, și urmărește cu precădere activitatea și direcțiile de cercetare.

Partea a treia conține lista de lucrări referite pe parcursul lucrării.