Mining Social-Driven Data

Mikołaj Morzy, PhD

Institute of Computing Science
Faculty of Computing Science and Management
Poznan University of Technology

habilitation thesis submitted to
Poznan University of Technology
in partial fulfillment of the requirements for the
habilitation doctor degree in computer science

Poznan, 2009

Contents

Abstract

The Web 2.0 revolution spreading over the Internet has dramatically changed the way data is gathered and processed by web applications. The static, authoritarian model of the Web has been abandoned in favor of dynamic, community-driven model of user-generated content. Social networks appear abundantly in all domains of human activity, presenting users with limitless volumes of data, information, and knowledge. Unfortunately, unearthing the knowledge hidden in vast repositories of social applications, such as wikis, Internet forums, or the blogosphere, is a difficult and challenging task. Structural complexity, huge volume of data to be processed, stochastic nature of social processes underlying the data, all contribute to the hardness of this task. Data mining methods developed for relational data repositories cannot be simply adapted to social-driven data. New models and algorithms are required for discovering knowledge in social-driven data.

This dissertation introduces a trust-based approach for mining social-driven data. The author examines different types of social-driven data, including blogs, Internet forums, and online auctions, and utilizes common underlying notions of trust and credibility to develop algorithms for mining social-driven data. Using the notions of trust and credibility allows to discover various important patterns in social-driven data. In the blogosphere, trust manifests itself as the ranking of blogs based on their relative influence on the blogosphere. In the domain of Internet forums, mining the social network of participants unveils true social roles attributed to particular participants. Finally, trust and credibility form the foundation of reputation models for the participants of online auctions.

The dissertation presents new models and algorithms for mining social-driven data. All algorithms have been implemented and their effectiveness has been verified by thorough experiments. The results of the experimental evaluation of models and algorithms allowed to confirm the main thesis of the dissertation, namely, that trust and credibility were the most important and crucial notions used to create relationships in social-driven data. In addition, it has been proven that trust and credibility might be discovered automatically by using data mining methods on the underlying social networks.

Foreword

The presented habilitation thesis is the result of author's work conducted on risk, credibility, and reputation in social networks. The research has been carried out in the Institute of Computing Science at Poznan University of Technology throughout the last four years. Concurrently with the project, four BSc thesis and five MSc thesis have been completed, as well as one PhD thesis has been started. All essential results presented in the dissertation are also available in the form of conference and journal publications, freely downloadable from the author's webpage [152].

The general themes of the dissertation are trust and credibility in social-driven data. The dissertation is not merely a simple compilation of conference and journal papers, because these manuscripts focus on narrowly defined subjects and lack a general unifying framework. Therefore, the author has decided to use the material from these publications selectively, enriching it with detailed descriptions and commentaries, adding to each chapter relevant information on the related work, and providing the reader with a broad introduction to the contemporary world of new media, new forums and new markets. The reader may thus consider this dissertation as an example of a cumulative habilitation thesis, with subsequent chapters being much edited previous author's works on the subject. The original papers, on which this habilitation thesis is based, are all published in international conference proceedings and journals, and thus are written in English. For the reasons of linguistic coherence and to make the thesis accessible to a much broader audience, English has been selected as the preferred language for the entire dissertation.

Preface

Since the completion of my PhD thesis back in June, 2004, I have been gradually turning my attention from core data mining algorithms [88, 89, 93, 95, 127] towards investigating new data mining frontiers, most notably, in the field of social networks. Astonished with an unprecedented viral spread of social applications, combined with enormous advancements in both hardware and software underlying the Web 2.0 revolution, I have decided to focus on these new developments and seek opportunities to employ data mining techniques to discover interesting knowledge in vast volumes of data constituting the world of Web 2.0. As a person originating from a database community, my perspective is strongly biased towards structure, constraints and operations of a model. Most researchers active in the social network domain come from distributed processing, statistics, or graph theory. I hope that my database theory background provides an interesting insight into the discussed subject.

This dissertation summarizes the research conducted over the period of four years in the domains so diverse as blogs, Internet forums, online auctions, and moving object databases. As a consequence, the variety of subjects presented in the dissertation may induce incoherence and disjointedness between chapters. I have linked the chapters by the common theme of trust and confidence, the crucial notions from which other important social measures are derived. Also, in order to improve the quality of the presentation and the readability of the dissertation, I have used a strict template for all the chapters. Each chapter starts with a brief introduction and continues with a section on related work. Then, the main body of the chapter is presented, followed by experiments. At the end of each chapter the reader finds a section with conclusions and the future work agenda.

The dissertation is divided into three main parts. Part I, "The World of the New", consists of chapters presenting the background for the research. Chapter 1 presents an introduction to the Web 2.0 revolution and describes new forms of social activities (blogs, Internet forums, and online auctions). Based on this introduction, the motivation for the conducted research is presented and the main thesis of the dissertation is formulated. In Chapter 2 the characteristics of the data produced by Web 2.0 applications are described. These characteristics are crucial for understanding the research

directions presented in the forthcoming chapters. After describing the data used throughout the research, basic definitions are formulated and basic notions are formalized. In Part II, "Mining of the New", the main research contribution of the dissertation is presented. Chapter 3 contains the results of the research conducted on blogs and the blogosphere. In Chapter 4 Internet forums are presented and the results of the research investigating this social structure are communicated. Finally, Chapter 5 reports on the results of the research on credibility and reputation in online auctions. Part III consists of two chapters. Chapter 6 presents findings in the domain of mining moving object databases. Negative patterns are described in Chapter 7. The dissertation concludes in Chapter 8 with closing remarks. Presentations of prototype implementations are given in appendixes. Appendix A presents Trendoo, a prototype blog mining engine. Foruminer, a system for Internet forum mining, analysis, and social role discovery, is presented in Appendix B. New algorithms for mining reputation and credibility of online auction participants are implemented in Presto, an experimental online auction platform and simulator, described in Appendix C. Finally, the research on mining moving object databases resulted in the creation of Moppy, a moving object position prediction and visualization tool. Moppy is described in detail in Appendix D. Finally, in the Afterword, the reader may find a detailed list of people who participated in this research and the acknowledgment of their contribution.

The dissertation covers a broad spectrum of subjects relating to social networks and Web 2.0. Parts of the dissertation provide a necessary introduction to the discussed subjects. Obviously, these introductory parts cannot be considered an original contribution of the dissertation. To help the reader better distinguish between the general purpose description and the description of ideas, concepts, algorithms, and experiments constituting the original contribution of the dissertation, below the list of original contribution is explicitly provided:

- the concept of using semantic links between blogs to discover influence relations between blogs,
- the concept of using blog network to find entities advertised by early adopting and influential blogs,
- the concept of a dual feature vector representation of blog items,
- the experimental evaluation of the proposal consisting in crawling the blogosphere to discover influential blogs,
- the framework for mining Internet forums,
- the definitions of indexes and social roles in Internet forum communities,

- the experimental evaluation of the proposal consisting in crawling selected Internet forums, computing relevant indexes and discovering social roles among Internet forum contributors,
- the concept of the credibility of participants of online auctions and the algorithm for mining the credibility of participants,
- the concept of the density reputation measure for sellers in online auctions,
- the concept of the S-graph for mining reputation and automatic recommendation for participants of online auctions,
- the concept of the negative reputation of participants of online auctions,
- the concept of the implicit feedback in feedback-based reputation systems,
- the algorithm for mining positive and negative reputation of participants of online auctions,
- the algorithm to discover the implicit feedback in online auction data,
- the experimental evaluation of the proposed algorithms for mining reputation of participants of online auctions,
- the concept of movement rules and the model for position prediction based on movement rules,
- the AprioriTraj and Traj-PrefixSpan algorithms to discover movement rules from a large body of moving objects data,
- the experimental evaluation of the proposed prediction model,
- the concept of dissociation rules and the design of the DI-Apriori family of algorithms to discover dissociation rules from transactional data,
- the experimental evaluation of the dissociation rules model.

The last four years of my life have been marked with ever increasing satisfaction about my work, invaluable inspiration from my colleagues and students, and remarkable developments in my personal life. This dissertation serves as the completion of this phase and, hopefully, as the commencement of new daring projects, new research ideas, and new acquaintances. The finalization of the dissertation would not be possible without help and assistance from many people around me. As for contributing students, I have added the Afterword with acknowledgments and detailed description of their

contribution at the end of the dissertation. I am particularly indebted to all the staff in the Division of Computing Systems at the Institute of Computing Science, Poznan University of Technology, for many hours of fruitful discussions and critiques. The support I received from my family allowed me to concentrate on my work, without having to worry about other things. I am taking this opportunity to express my gratitude. Dagmara was always tender and supportive, gently prodding her husband. The parents were patiently waiting for the dissertation to finish and they provided great mental support, for which I am very much obliged. The grandparents showed unconditional confidence and trust in their grandson, presenting me with heart and positive attitude. Last but not least, Agatka would constantly remind me that there are things incomparably more important than research and dissertation, such as playing hide and seek, making lengthy strolls in our neighborhood, or answering a hundred "why?" questions in a row. To them I dedicate this work.