Studies in Big Data

Volume 1

Series Editor
Janusz Kacprzyk, Warsaw, Poland

For further volumes:
http://www.springer.com/series/11970
Data Mining and Knowledge Discovery for Big Data

Methodologies, Challenge and Opportunities
The field of data mining has made significant and far-reaching advances over the past three decades. Because of its potential power for solving complex problems, data mining has been successfully applied to diverse areas such as business, engineering, social media, and biological science. Many of these applications search for patterns in complex structural information. This trans-disciplinary aspect of data mining addresses the rapidly expanding areas of science and engineering which demand new methods for connecting results across fields. In biomedicine for example, modeling complex biological systems requires linking knowledge across many levels of science, from genes to disease. Further, the data characteristics of the problems have also grown from static to dynamic and spatiotemporal, complete to incomplete, and centralized to distributed, and grow in their scope and size (this is known as big data). The effective integration of big data for decision-making also requires privacy preservation. Because of the board-based applications and often inter-disciplinary, their published research results are scattered among journals and conference proceedings in different fields and not limited to such journals and conferences in knowledge discovery and data mining (KDD). It is therefore difficult for researchers to locate results that are outside of their own field. This motivated us to invite experts to contribute papers that summarize the advances of data mining in their respective fields. Therefore, to a large degree, the following chapters describe problem solving for specific applications and developing innovative mining tools for knowledge discovery.

This volume consists of nine chapters that address subjects ranging from mining data from opinion, spatiotemporal databases, discriminative subgraph patterns, path knowledge discovery, social media, and privacy issues to the subject of computation reduction via binary matrix factorization. The following provides a brief description of these chapters.

Aspect extraction and entity extraction are two core tasks of aspect-based opinion mining. In Chapter 1, Zhang and Liu present their studies on people’s opinions, appraisals, attitudes, and emotions toward such things as entities, products, services, and events.
Chapters 2 and 3 deal with spatiotemporal data mining (STDM) which covers many important topics such as moving objects and climate data. To understand the activities of moving objects, and to predict future movements and detect anomalies in trajectories, in Chapter 2, Li and Han propose Periodica, a new mining technique, which uses reference spots to observe movement and detect periodicity from the in-and-out binary sequence. They also discuss the issue of working with sparse and incomplete observation in spatiotemporal data. Further, experimental results are provided on real movement data to verify the effectiveness of their techniques.

Climate data brings unique challenges that are different from those experienced by traditional data mining. In Chapter 3, Faghmous and Kumar refer to spatiotemporal data mining as a collection of methods that mine the data’s spatiotemporal context to increase an algorithm’s accuracy, scalability, or interpretability. They highlight some of the singular characteristics and challenges that STDM faces with climate data and their applications, and offer an overview of the advances in STDM and other related climate applications. Their case studies provide examples of challenges faced when mining climate data and show how effectively analyzing the spatiotemporal data context may improve the accuracy, interpretability, and scalability of existing methods.

Many scientific applications search for patterns in complex structural information. When this structural information is represented as a graph, discriminative subgraph mining can be used to discover the desired pattern. For example, the structures of chemical compounds can be stored as graphs, and with the help of discriminative subgraphs, chemists can predict which compounds are potentially toxic. In Chapter 4, Jin and Wang present their research on mining discriminative subgraph patterns from structural data. Many research studies have been devoted to developing efficient discriminative subgraph pattern-mining algorithms. Higher efficiency allows users to process larger graph datasets, and higher effectiveness enables users to achieve better results in applications. In this chapter, several existing discriminative subgraph pattern-mining algorithms are introduced, as well as an evaluation of the algorithms using real protein and chemical structure data.

The development of path knowledge discovery was motivated by problems in neuropsychiatry, where researchers needed to discover interrelationships extending across brain biology that link genotype (such as dopamine gene mutations) to phenotype (observable characteristics of organisms such as cognitive performance measures). Liu, Chu, Sabb, Parker, and Bilder present path knowledge discovery in Chapter 5. Path knowledge discovery consists of two integral tasks: 1) association path mining among concepts in multipart phenotypes that cross disciplines, and 2) fine-granularity knowledge-based content retrieval along the path(s) to permit deeper analysis. The methodology is validated using a published heritability study from cognition research and obtaining comparable results. The authors show how pheno-mining tools can greatly reduce a domain expert’s time by several orders of magnitude.
when searching and gathering knowledge from published literature, and can facilitate derivation of interpretable results.

Chapters 6, 7 and 8 present data mining in social media. In Chapter 6, Bhattacharyya and Wu, present “InfoSearch: A Social Search Engine” which was developed using the Facebook platform. InfoSearch leverages the data found in Facebook, where users share valuable information with friends. The user-to–content link structure in the social network provides a wealth of data in which to search for relevant information. Ranking factors are used to encourage users to search queries through InfoSearch.

As social media became more integrated into the daily lives of people, users began turning to it in times of distress. People use Twitter, Facebook, YouTube, and other social media platforms to broadcast their needs, propagate rumors and news, and stay abreast of evolving crisis situations. In Chapter 7, Landwehr and Carley discuss social media mining and its novel application to humanitarian assistance and disaster relief. An increasing number of organizations can now take advantage of the dynamic and rich information conveyed in social media for humanitarian assistance and disaster relief.

Social network analysis is very useful for discovering the embedded knowledge in social network structures. This is applicable to many practical domains such as homeland security, epidemiology, public health, electronic commerce, marketing, and social science. However, privacy issues prevent different users from effectively sharing information of common interest. In Chapter 8, Yang and Thuraisingham propose to construct a generalized social network in which only insensitive and generalized information is shared. Further, their proposed privacy-preserving method can satisfy a prescribed level of privacy leakage tolerance that is measured independent of the privacy-preserving techniques.

Binary matrix factorization (BMF) is an important tool in dimension reduction for high-dimensional data sets with binary attributes, and it has been successfully employed in numerous applications. In Chapter 9, Jiang, Peng, Heath and Yang propose a clustering approach to updating procedures for constrained BMF where the matrix product is required to be binary. Numerical experiments show that the proposed algorithm yields better results than that of other algorithms reported in research literature.

Finally, we want to thank our authors for contributing their work to this volume, and also our reviewers for commenting on the readability and accuracy of the work. We hope that the new data mining methodologies and challenges will stimulate further research and gain new opportunities for knowledge discovery.

Los Angeles, California

June 2013

Wesley W. Chu
Contents

Aspect and Entity Extraction for Opinion Mining .........  1
Lei Zhang, Bing Liu

Mining Periodicity from Dynamic and Incomplete
Spatiotemporal Data ................................................  41
Zhenhui Li, Jiawei Han

Spatio-temporal Data Mining for Climate Data: Advances,
Challenges, and Opportunities .............................  83
James H. Faghmous, Vipin Kumar

Mining Discriminative Subgraph Patterns from Structural
Data..........................................................................  117
Ning Jin, Wei Wang

Path Knowledge Discovery: Multilevel Text Mining
as a Methodology for Phenomics ..........................  153
Chen Liu, Wesley W. Chu, Fred Sabb, D. Stott Parker,
Robert Bilder

InfoSearch: A Social Search Engine .........................  193
Prantik Bhattacharyya, Shyhtsun Felix Wu

Social Media in Disaster Relief: Usage Patterns, Data
Mining Tools, and Current Research Directions...........  225
Peter M. Landwehr, Kathleen M. Carley

A Generalized Approach for Social Network Integration
and Analysis with Privacy Preservation .....................  259
Chris Yang, Bhavani Thuraisingham
A Clustering Approach to Constrained Binary Matrix Factorization ................................................ 281
Peng Jiang, Jiming Peng, Michael Heath, Rui Yang

Author Index ................................................ 305