

MMIS-07, 08: Mining Multiple Information Sources Workshop Report

Xingquan Zhu
Dept. of Computer Science &
Eng.
Florida Atlantic University
Boca Raton, FL 33431
xqzhu@cse.fau.edu

Ruoming Jin and Yuri
Breitbart
Dept. of Computer Science
Kent State University
Kent, OH 44242
{jin,yuri}@cs.kent.edu

Gagan Agrawal
Dept. of Computer Science &
Eng.
Ohio State University
Columbus, OH 43210
agrawal@cse.ohio-
state.edu

ABSTRACT

In this report, we summarize the research issues, contents, and outcomes of the two recent workshops on Mining Multiple Information Sources (MMIS-07, 08) collocated with the 13th and the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-07 and KDD-08). We first summarize the research issues and topics which in workshop co-chairs' view are the major challenges for mining multiple information sources. Then we briefly introduce the content of the contributed papers in two years' program, along with the introduction to three keynote talks given by the invited speakers.

Keywords

Data Mining, Knowledge Discovery, Multiple Information Sources, Multi-Source Data Mining

1. PREFACE

The first ACM International Workshop on Mining Multiple Information Sources (MMIS) was organized and collocated with the 13th ACM KDD Conference on August 12 2007 in San Jose, CA. The main challenges identified by the workshop co-chairs, then, were threefold: (1) how to efficiently identify quality knowledge from a single data source, where patterns reveal local knowledge for each particular data repository, commonly referred to as *local patterns*; (2) how to integrate and unify multiple information sources into one single view such that previous unseen patterns can be discovered, commonly referred to as *global patterns*; and (3) how to discover the relationships of the patterns hidden across multiple information sources, where the features of the patterns (pattern frequencies and their utilities) across different data repositories define inter-repository relationships, which we refer to as *inter patterns*. The program committee selected seven papers which covered multi-source data mining applications and solutions from traditional machine learning, business intelligence, life science, and software engineering.

After the organization of the MMIS-07, the workshop co-chairs expanded the themes of the multi-source mining to address broader areas of multi-source data mining challenges.

Such as data integration for multiple information sources, clustering multi-source data, supervised learning with multiple data sources, association rule mining and stream data mining under multi-source data environments. Such changes helped the workshop co-chairs successfully organized the second MMIS workshop collocated with the 14th ACM KDD Conference on August 24 2008 in Las Vegas, NV. The committee selected six papers which addressed the multi-source mining from clustering, frequent itemsets, and supervised learning perspectives.

2. INTRODUCTION

As data collection sources and channels continuously evolve, mining and correlating information from multiple information sources has become a crucial step in data mining and knowledge discovery. On one hand, comparing patterns from different databases and understanding their relationships can be extremely beneficial for applications such as Bioinformatics, Sensor Networking, and Business Intelligence. On the other hand, many data mining and data analysis tasks such as classification, regression, and clustering, can significantly improve their performance if information from different sources can be properly leveraged and if the mining process has the power to survey all the data sources involved.

Unleashing the full power of multiple information sources is, however, a very challenging problem, considering that schemas used to represent each data collection might be different (data heterogeneity), data distributions and patterns underlying different data sources may undergo continuous changes (concept evolving), and mining tasks for each data source might also be different (mining diversity). Even though existing researches have demonstrated several approaches to utilize multiple information sources, these methods are still rather ad-hoc and inadequately address some of the fundamental research issues in this field: (1) Harnessing Complex Data Relationship: Multiple information sources represent a collection of highly correlated data, issues such as data integration, data integration, model integration, and model transferring across different domains, play fundamental roles in supporting KDD from multiple information sources; (2) Integrative and Cooperative Mining: For heterogeneous information sources with diverse mining tasks, the mining should be able to unify all data to gener-

ate enhanced global models, as well as help individual data collections to cooperatively achieve their respective mining goals; and (3) Differentiation and Correlation: Differentiate and coordinate the difference between data sources at the knowledge level is one crucial step for users to gain a high-level understanding of their data.

The aim of the MMIS workshop is to bring together data mining experts to revisit the problem of pattern discovery from multiple information sources, and identify and synthesize current needs for such purposes. Representative questions to be addressed include but are not limited to:

- **Harnessing Complex Data Relationship:** (a) Database similarity assessment. (b) Automatic schema mapping and relationship discovery. (c) New mapping framework for multiple information sources. (d) Data source classification and clustering. (e) Data cleansing, data preparation, data/pattern selection, conflict and inconsistency resolution
- **Integrative and Cooperative Mining:** (a) Model integration for heterogeneous information sources. (b) Mode transferring across different data domains. (c) Incremental and scalable data mining algorithms. (d) Multi-tasks multi-source co-learning for multiple information sources
- **Differentiation and Correlation:** (a) Local pattern analysis and fusion. (b) Global pattern synthesizing and assessment. (c) Merging local rules for global pattern discovery. (d) Pattern summarization from multiple datasets. (e) Multi-dimensional pattern search and comparison. (f) Pattern comparison across multiple data sources. (g) Inter pattern discovery from complex data sources
- **Stream data mining algorithms:** (a) Clustering and classification of data of changing distributions. (b) Data stream processing, storage, and retrieval systems. (c) Sensor networking
- **Security and privacy issues in multiple information sources**
- **Interactive data mining systems:** (a) Query languages for mining multiple information sources. (b) Query optimization for distributed data mining. (c) Distributed data mining operators in supporting interactive data mining queries.

3. MMIS-07 PROGRAM

MMIS-07 workshop program consists of two sections (4 regular papers and 3 short papers) and one invited keynote talk. The keynote speaker Dr. Xianghong Jasmine Zhou from the Department of Molecular and Computational Biology, University of Southern California, addressed the problem of functions, networks, and phenotypes by integrative genomics. In her talk, she discussed computational and statistical methods for integrative analysis of diverse genomic sources, including microarray data, proteomics data, sequence data, and the text data. This includes the algorithms for network-based data mining and methods for the integrative analysis of cross-platform microarray data.

The first section of the MMIS-07 program consists of the following three papers:

- “A Combinatorial Fusion Method for Feature Mining”, (*Tian, Weiss, Hsu, and Ma*), Fordham University, USA
- “Mining Vector-Item Patterns for Annotating Protein Domains Paper”, (*Wu and Denton*), North Dakota State University, USA
- “Combining Mining Results from Multiple Sources in Clinical Trials and Microarray Applications”, (*Altıparmak, Ozturk, Erdal, Ferhatosmanoglu, and Trost*), Ohio State University, USA

The authors of the first paper (“A Combinatorial Fusion Method for Feature Mining”) proposed to use information fusion to improve the performance of a classifier by constructing (“fusing”) new features that are combinations of existing numeric features. The quality of the fused features is measured with respect to the performance in classifying minority-class example, which makes the proposed method effective for imbalanced datasets. In the second paper (“Mining Vector-Item Patterns for Annotating Protein Domains Paper”), the authors indicated that the diversity of information collected in Bioinformatics and other application areas makes it increasingly important to develop techniques for associating data of different types. Consequently, they introduced an algorithm to find patterns involving items and vector data, and the proposed method was evaluated on yeast cell cycle gene expression data in combination with domain annotations from the Interpro database and on time series data. In the third paper (“Combining Mining Results from Multiple Sources in Clinical Trials and Microarray Applications”), the authors proposed a mining framework to gather high quality information from heterogeneous and multi-dimensional time series data and applied it to two important classes of biomedical data mining applications: Clinical Trials and Microarray Gene Expressions. Experimental results demonstrated that the quality of the mining results can be continuously improved with the number of data sets and/or metrics used for mining.

The second section of the MMIS-07 program consists of the following four papers.

- “Finding New Customers Using Unstructured and Structured Data”, (*Melville, Liu, Lawrence, Khabibrakhmanov, Pendus, and Bowden*), IBM T.J. Watson Research Center, USA
- “Incorporating Background Knowledge from the World Wide Web for Rule Evaluation using the Minimum Discriminative Information Principle”, (*Fodeh and Tan*), Michigan State University, USA
- “Integrating Projects from Multiple Open Source Code Forges”, (*Conklin*), Elon University, USA
- “An Ensembled based Bayesian Network Learning Algorithm on Limited Data”, (*Liu, Tian, and Zhu*), Beijing University of Posts and Telecommunications, China

The first paper (“Finding New Customers Using Unstructured and Structured Data”) addressed a very interesting

and practical problem on identifying new customers for any sales-oriented businesses. In their paper, the authors demonstrated that the content of company web sites can often be a rich source of information in identifying particular business alignments. The authors proposed solution to employ supervised learning to build effective predictive models on unstructured web content as well as on structured firmographic data. In addition, they also explored methods to leverage the strengths of both sources by combining these data sources. The authors of the second paper (“Incorporating Background Knowledge from the World Wide Web for Rule Evaluation using the Minimum Discriminative Information Principle”) presented a methodology for augmenting background information from an authoritative source on the World Wide Web to the subjective evaluation of association rules. In the paper, the authors focused on the problem of mining association rules in the medical domain, where background information automatically acquired from the MEDLINE database of biomedical citations was used to evaluate the quality of association rules extracted from an electronic medical records (EMR) database. In the third paper (“Integrating Projects from Multiple Open Source Code Forges”), the authors presented a method for integrating data of the open source projects through project (entitie) matching across multiple code forges. The authors of the last paper (“An Ensembled based Bayesian Network Learning Algorithm on Limited Data”) addressed the problem of Bayesian Network learning for limited data. In their paper, they proposed a sampling method, namely Root Nodes based Sampling, based ensemble framework and component integration method to support BN learning (due to the visa problem, the authors of this paper were unable to attend the workshop and present their work).

4. MMIS-08 PROGRAM

Similar to the last year, the program of MMIS-08 workshop also has two sections. In addition, we invited two keynote speakers this year, one from academia and one from industry.

The keynote talk of Dr. Naren Ramakrishnan (from Virginia Tech) addressed the problem of harnessing multiple information sources using compositional data mining. The theme is to build a compositional approach to building complex data mining applications from simple algorithms, which will enable users to capture complex patterns as compositions of simpler patterns. Two basic pattern classes: re-descriptions and biclusters, were addressed in the talk, where re-descriptions identify patterns within a domain and biclusters identify patterns across domains. Given a relational database and its schema, the schema can be automatically compiled into a compositional data mining program, and compositional patterns can be efficiently computed without ‘wasteful’ data mining.

The second keynote speaker Dr. Haixun Wang (from IBM Thomas J. Watson Research Center) addressed the problem of resource allocation in mining multiple data streams, namely load shedding in stream process. Due to the large volume and the high speed of streaming data, mining algorithms must cope with the effects of system overload. How to realize maximum mining benefits under resource constraints becomes a challenging task. In his talk, Dr. Wang proposed

a load shedding scheme for classifying multiple data streams with focus the following two problems: i) how to classify data that are dropped by the load shedding scheme? and ii) how to decide when to drop data from a stream? The proposed method uses a quality of decision (QoD) metric to measure the level of uncertainty in classification when exact feature values of the data are not available because of load shedding. A Markov model is used to predict the distribution of feature values and we make classification decisions using the predicted values and the QoD metric. Consequently, resources are allocated among multiple data streams to maximize the quality of classification decisions.

The first section of the MMIS-08 program consists of the following three papers:

- “Signalling Potential Adverse Drug Reactions from Multiple Administrative Health Databases”, (*Jin, Chen, He, Kelman, McAullay, and OKeefe*), CSIRO, Australia
- “An Exploration of Understanding Heterogeneity through Data Mining”, (*Liu and Dou*), University of Oregon, USA
- “Multiclass Multifeature Split Decision Tree Construction in a Distributed Environment”, (*Ouyang, Patel, and Sethi*), Oakland University, USA

The work reported in the first paper (“Signalling Potential Adverse Drug Reactions from Multiple Administrative Health Databases”) was motivated by the real-world challenge of detecting Adverse Drug Reactions (ADRs), which is a leading cause of hospitalization and death worldwide, from multiple administrative health databases. To signal unexpected and infrequent patterns characteristic of ADRs, the authors proposed to discover Unexpected Temporal Association Rules and further suggested the corresponding interestingness measure, unexlev. In the paper, the authors reported a new algorithm, HUNT, for highlighting infrequent and unexpected patterns by comparing the rankings of the patterns based on the unexlev measure and the traditional leverage measure. Notice that internet and Web have resulted in many distributed information resources which in general are structurally and semantically heterogeneous, whereas the heterogeneity itself has not yet been studied in a formal way, the authors of the second paper (“An Exploration of Understanding Heterogeneity through Data Mining”) provided a brief survey on various ways to categorize heterogeneity in the literature, and then performed a case study on detecting a specific class of heterogeneity in the setting of Semantic Web ontologies. The one that can be discovered by only data-driven approaches. The authors of the third paper (“Multiclass Multifeature Split Decision Tree Construction in a Distributed Environment”) addressed the problem of decision tree induction in distributed settings. The paper provided solutions to construct compact decision trees using multifeature splits for distributed data.

The second section of the MMIS-08 program consists of the following three papers.

- “A Novel Approach for Discovering Chain-Store High Utility Patterns in a Multi-Stores Environment”, (*Lan and Tseng*), National Cheng Kung University, Taiwan

- “Large Scale Security Log Sources Integration: An Ensemble Method”, (*Mao, Wen, Li, Jia, and Wu*), National University of Defense Technology, China
- “Applying MDA to Integrate Mining Techniques into Data Warehouses: A Time Series Case Study”, (*Par-dillo, Mazn, Zubcoff, and Trujillo*), University of Alicante, Spain

In the first paper (“A Novel Approach for Discovering Chain-Store High Utility Patterns in a Multi-Stores Environment”), the authors argued that existing methods on utility mining were mostly designed for centralized databases but not suitable for the environment with multiple data sources like a chain-store enterprise. Consequently, the authors proposed to discover Chain-Store High Utility Pattern that contains not only individual profit and quantity of items but also common selling periods and stores of items in a multi-stores environment. The authors of the second paper (“Large Scale Security Log Sources Integration: An Ensemble Method”) proposed an ensemble framework for large scale security log sources integration. The proposed efforts intended to solve two challenges in integrating heterogeneous data from different monitoring sensors, this includes (1) how to unify the heterogeneous log data at the schema-level; and (2) how to understand the large scale instances with different encoding format? (Due to the visa problem, the authors of this paper were unable to attend the workshop. Their paper was presented by one of the workshop co-chairs instead). In the last paper (“Applying MDA to Integrate Mining Techniques into Data Warehouses: A Time Series Case Study”), the authors complained that existing data mining lacks a modeling architecture that allows analysts to consider it as a truly software-engineering process. Consequently, the authors proposed a model driven approach based on (i) a conceptual modeling framework for data mining, and (ii) a set of model transformations to automatically generate analysis models for data mining. As a result, the analysts can concentrate on understanding the analysis problem via conceptual data-mining models instead of wasting efforts on low-level programming tasks related to the underlying-platform technical details.

5. CONCLUSION

All together, the papers selected in the MMIS-07 and MMIS-08 proceedings represent a subset of existing research on knowledge discovery from multiple data sources. The examples and applications reported in the papers spanned numerous domains from life science, business intelligence, network security, software engineering, and traditional artificial intelligence and machine learning. The algorithms proposed in the paper covered numerous data mining areas such as supervised learning, clustering, association rule mining, and data streams. This asserts that as data mining is rapidly becoming a useful tool for different applications, the needs of collaborating multiple data sources for effective data mining are indeed a critical issue facing the data mining community.

We wish to provide the MMIS workshop series as a platform to encourage a wide community of researchers to communicate and exchange ideas in areas of mining multiple information sources. We expect to continue this workshop in the following years and hope our efforts can eventually lead to

a break-through to bring multi-source data mining to the next level.

6. ACKNOWLEDGEMENT

We are grateful to the members of the MMIS workshop program committee for their constructive comments in organizing the workshops and finishing all the reviews in a very short amount of time. We also thank all keynote speakers for their support and excellent presentations. Finally, we would also like to thank all the authors who submitted their papers to the workshops. It would be impossible to make success workshops without their contributions.

7. MMIS-07, 08 WORKSHOP CO-CHAIRS

Xingquan Zhu (07, 08) - Florida Atlantic University, USA
 Ruoming Jin (07, 08) - Kent State University, USA
 Gagan Agrawal (07) - Ohio State University, USA
 Yuri Breitbart (08) - Kent State University, USA

8. MMIS-07, 08 WORKSHOP PC MEMBERS

Aalid G. Aref (08) - Purdue University, USA
 Henrique Andrade (07) - IBM T.J. Watson, USA
 Philip Chan (07, 08) - Florida Institute of Technology, USA
 Ian Davidson (07) - SUNY at Albany, USA
 Dejing Dou (08) - University of Oregon, USA
 Christopher Jermaine (08) - University of Florida, USA
 Taghi M. Khoshgoftaar (07, 08) - Florida Atlantic University, USA
 Tao Li (07, 08) - Florida International University, USA
 Huan Liu (07, 08) - Arizona State University, USA
 Gary M. Weiss (07) - Fordham University, USA
 Prem Melville (08) - IBM T.J. Watson, USA
 Xintao Wu (07, 08) - UNC Charlotte, USA
 Jieping Ye (08) - Arizona State University, USA
 Shichao Zhang (07, 08) - University of Technology, Sydney
 Aoying Zhou (08) - Fudan University, China
 Zhi-hua Zhou (07, 08) - Nanjing University, China

About the Authors

Xingquan Zhu is an Assistant Professor in the Department of Computer Science and Engineering at Florida Atlantic University, Boca Raton, FL. He received his Ph.D degree in Computer Science from Fudan University, Shanghai, China, in 2001. From February 2001 to October 2002, he was a Postdoctoral Associate in the Department of Computer Science, Purdue University, West Lafayette, IN. From October 2002 to July 2006, He was a Research Assistant Professor in the Department of Computer Science, University of Vermont, Burlington, VT. His research interests include data mining, machine learning, multimedia systems, and information retrieval. Since 2000, he has published over 70 technical papers in referred journals and conference proceedings.

Ruoming Jin is currently an assistant professor in the Computer Science Department at Kent State University. He received a BE and a ME degree in computer engineering from Beihang University (BUAA), China in 1996 and 1999, respectively. He earned his MS degree in computer science from University of Delaware in 2001, and his Ph.D. degree

in computer science from the Ohio State University in 2005. His research interests include data mining, databases, processing of streaming data, bioinformatics, and high performance computing. He has published more than 50 papers in these areas. He is a member of ACM and SIGKDD.

Gagan Agrawal is a Professor of Computer Science and Engineering at the Ohio State University. He received his B.Tech degree from Indian Institute of Technology, Kanpur, in 1991, and M.S. and Ph.D degrees from University of Maryland, College Park, in 1994 and 1996, respectively. His research interests include parallel and distributed computing, compilers, data mining, grid computing, and data integration. He has published more than 140 refereed papers in these areas. He is a member of ACM and IEEE Computer Society. He received a National Science Foundation CAREER award in 1998.

Yuri Breitbart is Ohio Board of Regents Distinguished Professor of Computer Science in the Department of Computer Science at Kent State University. Prior to that he was a Member (and later a Distinguished Member) of Technical Staff at the Bell Laboratories at Murray Hill, New Jersey. From 1986 to 1996 he was a Professor in the Department of Computer Science at University of Kentucky and was the Department Chair there for the first seven years. Prior to 1986, he was leading the Database Research group first at ITT Research Center and then at Amoco Production Company Research Center. He has been consulting for numerous companies and among them IBM, Boeing, Amoco, HP, and Bell Labs. Dr. Breitbart's research is in the area of distributed information system, network management systems and data mining. Dr. Breitbart has received DSc degree in Computer Science from the Department of Computer Science at Israel Institute of Technology (Technion). He is a Fellow of the ACM and member of IEEE Computer Society and SIGMOD. He has served on numerous program committees and NSF panels.