Challenging Research Issues in Data Mining, Databases and Information Retrieval

Aparna S. Varde
Department of Computer Science
Montclair State University
Montclair, NJ, USA
vardea@montclair.edu

ABSTRACT

Data mining research along with related fields such as databases and information retrieval poses challenging problems, especially for doctoral students. The research spreads over a variety of topics such as text mining, semantic web, multilingual information analysis, heterogeneous data management, database learning, digital libraries and more. Much of this research cuts across multiple fields and presents interesting issues for discussion at conferences with a confluence of several tracks. The ACM Conference on Information and Knowledge Management provides an excellent environment for presenting such research problems spanning the three tracks of database systems, information retrieval and knowledge management. This article provides an overview of the dissertation problems presented at a Ph.D. workshop in the ACM Conference on Information and Knowledge Management. The goal of such workshops is to allow students to showcase their creative ideas at an early stage. This enables experts to critique their work and also gives the students an opportunity to exchange their thoughts with each other, besides providing excellent networking opportunities with industry and academia. This article provides an overview of the papers presented at the Ph.D workshop. It serves as a motivation for researchers to delve deeper into the innovative dissertation problems summarized here and the related work in these areas.

Keywords

Dissertation problems, Ph.D. forum, interdisciplinary work, knowledge discovery, database management, web technology

1. INTRODUCTION

This article serves as a short review paper describing the doctoral research issues that formed the content of PIKM 2008, the 2nd Ph.D. Workshop on Information and Knowledge Management. This was in held conjunction with ACM CIKM 2008, the Conference on Information and Knowledge Management, at Napa Valley, CA, USA in October 2008. The PIKM 2008 workshop served as a good forum for doctoral candidates in database systems, information retrieval and knowledge management to present their early work and get feedback from experienced researchers. This workshop was built on the success of the PIKM 2007 workshop, the first in the series, held along with ACM CIKM 2007 (at Lisbon, Portugal in November 2007).

The PIKM 2008 call for papers attracted 29 submissions from 13 countries across Asia, Europe, and North America. The program committee accepted 17 high quality submissions of which 10 were full papers and another 7 were poster papers, covering a vast range of topics spanning database systems, text mining, semantic web, social networks, and human computer interaction (HCI).

The program committee comprised 22 experts from across 7 countries, maintaining a good balance of academia, research labs and the corporate world. The program committee members along with the external reviewers put in tremendous efforts and it was gratifying to see such high quality reviews, considering the relatively short reviewing time frame of approximately 2 weeks.

The workshop did indeed serve its purpose of providing the participants with worthwhile feedback, and enabled the sharing of ideas with fellow students and researchers from around the world. It is hoped that this workshop will encourage many more of its kind, and that it will further enhance research across multiple tracks.

This article outlines the technical contributions of the Ph.D. workshop in the form of the research problems presented along with their proposed solutions and pilot evaluations. Therefore, it encompasses challenging doctoral research issues in the areas of data mining, databases and information retrieval today that would be of interest to a fairly broad audience in industry and academia.

2. RESEARCH PROBLEMS

The dissertation research problems presented at the workshop are described in the following three sections on Data Mining, Databases and Information Retrieval respectively. Although there are overlapping research issues in many of these papers, they are divided into these categories based on their primary contributions to the respective areas.

2.1 Data Mining

A text mining issue pertaining to topic models for text related applications was discussed by Ha Thuc et al. [HS-08]. They covered text models such as aspect model or LDA. They discovered limitations of scalability issues in running the models in mining large corpora and the inability to model the important

concept of relevance which prevents the models from being directly applied for text classification. To overcome these limitations, they introduced a one-scan topic model requiring only a single pass over a corpus for inference and also proposed relevance-based topic models that provided better results than state-of-the-art models.

The problem of concept search in discovering knowledge from non-English and non-European sources was discussed by Riaz [R-08]. Urdu was chosen as an example language because of its unique nature, morphology and a large number of speakers. Named-entity identification was considered to be useful in determining the knowledge being sought by the user. A TREC like evaluation criteria was presented with relevance judgments, test collection and appropriate queries for knowledge discovery.

Wu [W-08] reported the results of database learning in human computer interaction (HCI) errors. Earlier studies on HCI errors focused on improving the system performance. This work proposed an approach to identify the nature of HCI errors in interactions from the user perspective. It analyzed error episodes, recovery trials, and recovery actions, suggesting systematic methods for error recovery by users in different cases. Pilot studies corroborated the usefulness of the proposed approach.

In the work of [AI-08] the areas of clustering was addressed along with the concept of term frequency - inverse document frequency, i.e., tf-idf. In this paper, a technique was proposed based on integrating models to enhance performance in querying and mining data from spontaneous speech. The proposed technique clustered the training topics using tf-idf properties, selecting the best models for each cluster. For test data, the topic cluster was found and a combination of models for this cluster was used. It was found to improve performance as compared to earlier methods in the area.

Song et al. [WSWA-08] dealt with an NP-complete problem in the area of social network mining, proposing a 2-step method for community detection. This involved first analyzing vertex similarity of the network (a microscopic view) and putting a pair of vertices into the same community if they were similar; followed by incrementing modularity of the similarity-based communities. If the number of edges between 2 communities was greater than an expected number based on random choice, the communities were merged. They tested their method on over 20 data sets and did better than existing algorithms.

Knowledge management with respect to navigating humanities resources was the focus of [Sh-08]. They emphasized that in the humanities, events and narratives though important, are not identified and disambiguated by most knowledge organization systems; and that the digitization of artifact collections and development of digital metadata make it imperative to address this issue. They described their research on gazetteers to depict such events; along with the relationships and best practices to use the gazetteers for improving digital resources and services.

A decision support problem pertaining to online communities in social network mining was discussed in the work of Smith [Sm-08]. Online communities are connecting hordes of individuals that generate rich social network data. The social capital residing in these networks is by far unknown and needs to be discovered. This work proposed to create a mathematical model of social capital to incorporate mobilization of social resources. It involved evaluating nodes based on their relationships and attributes, and also on their social resources. The result was a quantitative model for characterizing and providing decision support on how to maximize participation within social networks

2.2 Databases

A SQL database system to solve constraints was discussed in [SW-08]. It involved integrating constraint satisfaction problems (CSP) with databases using the structured query language SQL. A case study was presented on developing such a system. Their SQL-based constraint data engine (SCDE) supported key concepts of "consql", with additional syntax to better align CSP admin with regular SQL. Testing on ACC Basketball Scheduling showed the feasibility of SCDE. Ongoing work included building an environment for end-users to progam and debug CSPs; providing a more interactive database engine for objective functions; and improving the performance of hard constraint problems.

In [JX-08], the problem addressed was the privacy-preserving integration of distributed heterogeneous data. Many applications have huge volumes of data in distributed databases collected over time or produced by large scale scientific experiments. Such data sharing is subject to confidentiality of individuals and institutions. Accordingly, this work proposed a distributed anonymization protocol for independent data providers to develop a virtual anonymized database given horizontally partitioned databases, and a secure query protocol for clients to query virtual databases. It also proposed distributed data integration architecture for querying heterogeneous and private databases.

Haapasalo et al. [HJSS-08] presented the database management problem of concurrency control and recovery given multiversion database structures They proposed the design and implementation of many multiversion index structures with complete concurrency-control along with ARIES-based recovery algorithms. Their experiments involved a multiversion B^+tree as a historical storage, to which committed transaction updates were moved one by one from a separate B^+tree. They also considered using an optimized R-tree to store the multiversion data as 2-D line segments. They implemented a standard B^+tree based solution to store different versions of a data item consecutively; along with a solution based on the existing time-split B^+tree. They concluded that the solution with a multiversion B^+tree was the most efficient. This work won the best paper award in the Ph.D. workshop.

An extended cooperative transaction paradigm for the XML data model was proposed in [GS-08]. In certain areas, e.g., design or media production, authors cooperate on projects and a common data format for communication is XML. This paper addressed the special needs of working together on shared XML graph structures considering early visibility of updates, multi-directional information flow, and parallel tasks. Since state-of-the-art transaction models were not found suitable, they presented a new model incorporating multi-level transactions and dynamic actions to meets the special needs. Their model was found advantageous in terms of providing appropriate concepts for transaction synchronization and resolution of conflicts.

The issue of advanced properties in ontology mapping for web databases was addressed by Stoutenberg in [St-08]. Ontologies support many applications today, such as enhanced search, rapid enterprise integration and cross-domain data sharing. Most state-

of-the-art approaches map the ontologies using similarity and equivalence and very few apply knowledge in upper ontologies. This work developed algorithms to acquire relationships between ontological components beyond similarity and equivalence which encompass hyponymy. It also built algorithms to map ontologies based on relationships specified within ontologies. Their initial test results looked promising with potential for further work on ontology mapping in the web database area.

2.3 Information Retrieval

An important information retrieval issue related to folksnomy systems formed the focus of Abel's work [A-08]. With the advent of Web 2.0, various folksonomy systems such as Flickr and del.icio.us have become popular, allowing users to annotate resources, e.g., images and websites with freely chosen keywords, i.e., tags. The evolving set of such tag assignments, generally user-tag-resource bindings, are called folksonomies. This work analyzed the benefit of additional semantics in folksonomy systems. The GroupMe! folksonomy system was presented which brought additional semantics to tagging systems enabling the grouping of resources. The work introduced group-sensitive ranking algorithms that were better than existing ones. GroupMe! also contributed towards closing the gap between the social and the semantic web. GroupMe! data was stored as RDF and provided based on principles of Linked Data. An architecture was developed to bridge between folksonomies and ontologies using the MOAT (Meaning Of A Tag) framework.

Ingawale [I-08] addressed a Wikipedia related issue. They noted that explanations for 'altruistic' contributor tendencies considering the positivistic paradigm, with roots in organizational psychology, though heavily researched, are not transferable to quantitative models of predictive value, with reference to Wikipedia metrics. They also pointed out that the models using aggregated top-level relationships between Wikipedia entities face the following problem: they assume relationships between entities as inputs to the process, not as emergent phenomena that evolving with the output. They argued for an agent based model of Wikipedia. The main contribution of their work was a diagnostic and/or prescriptive tool for decision makers in organizations using or planning to use Knowledge Management Systems.

Brauer et al. [BLD-08] addressed the problem of mapping enterprise entities to text segments. In recent years, important business information is stored in unstructured formats such as documents and emails. Consider the fact that documents shared among business partners store information on transactions (purchases, invoices etc.). It is challenging to appropriately identify and associate real-world entities in unstructured data with those in structured data e.g., enterprise databases. To solve this problem, Brauer et al. proposed a robust process methodology with 3 phases: extracting entities from documents, generating entity mapping of with structured data, and disambiguating mappings to discover relationships from the enterprise data and the documents' structure.

In [SU-08], the subject of digital libraries was approached. The aim of this research was to find the networks of "academic community" in digital libraries by using different tools and metrics from social network analysis. This study was expected to detect the figureheads responsible for shaping the multi-disciplinary community of digital libraries. It also exposed the factors pertaining to the networking of academic communities. The authors indicated that the application of social network analysis techniques to trace linkages among community members was innovative and hoped that it would be useful in providing insight into the shaping of a community.

A web information retrieval problem related to imprecise regions formed the work of Pasley [P-08]. An overview of their research was presented in terms of defining imprecise regions such as "The Midlands" of Great Britain. They proposed using unstructured data sources found on the web to address this issue. A literature review was conducted on such regions and the experimental plan was developed accordingly. Their study showed the results of 2 fundamental experiments. One of them was on geo-tagging, while the other was on geographic coverage on the web. The results looked promising and encouraged further work on this important information retrieval issue.

3. CONCLUSIONS

This article summarizes the major contributions of PIKM 2008, the 2nd Ph.D. workshop in CIKM. The Ph.D. workshop was a grand success. It served as a good forum for exchange of thoughts and ideas among doctoral students at an early stage of their dissertation. This workshop emerged even more successful than its predecessor PIKM 2007, in terms of the quality and quantity of submissions as judged from the reviews. PIKM 2008 also included a poster session in addition to oral presentations in order to give more students an opportunity to present their work in a rather short span of time.

The best paper award was presented to Tuukka K. Haapasalo, Ibrahim Jaluta, Seppo Sippu, and ljas Soisalon-Soininen for their paper on "Concurrency control and recovery for multiversion database structures". The presentation of the best paper award served as a motivation for enhancing their research as well as encouraging other students to be more competitive in executing their research and presenting it at such events. It indicated that the workshop meets the quality standards of leading conferences where such accolades are given.

The cutting edge research ideas presented at this Ph.D workshop provided the scope for further research in the respective areas, especially in working towards interdisciplinary problems across the various tracks. It is sincerely hoped that we will have more such workshops and doctoral consortia for presentation of dissertation research at international conferences with a good blend of industry and academia.

4. ACKNOWLEDGMENTS

The contributions of Dr. Prasan Roy for serving as the PIKM 2008 workshop co-chair, session chair and reviewer and Dr. Anisoara Nica for serving as a session chair and program committee member of the workshop are gratefully acknowledged. They provided valuable inputs to the students during the Ph.D. workshop in addition to helping with various other issues related to its organization.

All the other program committee members are also thanked for their reviewing efforts. These are: Dr. Indrajit Bhattacharya, Dr. Francisco Couto, Dr. Sreenivas Gollapudi, Dr. Katherine Herbert, Dr. Vagelis Hristidis, Dr. Giti Javidi, Dr. Mouna Kacimi, Dr. Daniel Keim, Dr. Andreas Koeller, Dr. Pawan Lingras, Dr. Bin Liu, Dr. Murali Mani, Dr. Florent Masseglia, Dr. Thomas Neumann, Dr. Jian Pei, Dr. Carolina Ruiz, Dr. Pierre Senellart, Dr. Ranga Raju Vatsavai, Dr. Li Xiong, Dr. Mohammed Zaki amd Dr. Zhongfei Zhang.

Sincere gratitude is also conveyed to the ACM CIKM 2008 organizing committee, especially the General Chair, Dr. James Shanahan and the Workshops Chair, Dr. Gregory Grefenstette, for their support.

5. REFERENCES

[A-08] Fabian Abel: The benefit of additional semantics in folksonomy systems, PIKM 2008, pp. 49-56.

[AI-08] Muath Alzghool, Diana Zaiu Inkpen: Clustering the topics using TF-IDF for model fusion, PIKM 2008, pp. 97-100.

[BLD-08] Falk Brauer, Alexander Löser, Hong-Hai Do: Mapping enterprise entities to text segments, PIKM 2008, pp. 85-88.

[GS-08] Francis Gropengießer, Kai-Uwe Sattler: An extended cooperative transaction model for xml, PIKM 2008, pp. 41-48.

[HJSS-08] Tuukka K. Haapasalo, Ibrahim Jaluta, Seppo Sippu, Eljas Soisalon-Soininen: Concurrency control and recovery for multiversion database structures, PIKM 2008, pp. 73-80.

[HS-08] Viet Ha-Thuc, Padmini Srinivasan: Topic models and a revisit of text-related application, PIKM 2008. pp. 25-32.

[I-08] Myshkin Ingawale: Understanding the wikipedia phenomenon: a case for agent based modeling, PIKM 2008, pp. 81-84.

[JX-08] Pawel Jurczyk, Li Xiong: Towards privacy-preserving integration of distributed heterogeneous data, PIKM 2008, pp. 65-72.

[P-08] Robert Pasley: Defining imprecise regions using the web, PIKM 2008, pp. 105-108.

[R-08] Kashif Riaz: Concept search in Urdu, PIMM 2008, pp. 33-40. [P-08] Robert Pasley: Defining imprecise regions using the web, PIKM 2008, pp. 105-108.

[RV-08] Prasan Roy, Aparna S. Varde (Eds.): Proceedings of the Second Ph.D. Workshop in CIKM, PIKM 2008, Napa Valley, California, USA, October 30, 2008. ACM 2008, ISBN 978-1-60558-257-3.

[SU-08] Monica Sharma, Shalini R. Urs: Network dynamics of scholarship: a social network analysis of digital library community, PIKM 2008, pp. 101-104.

[Sh-08] Ryan Shaw: Event gazetteers for navigating humanities resources, PIKM 2008, pp. 89-92.

[SW-08] Sebastien Siva, Lesi Wang: A SQL database system for solving constraints, PIKM 2008, pp. 1-8.

[Sm-08] Matthew S. Smith: Social capital in online communities, PIKM 2008,. 17-24.

[St-08] Suzette Kruger Stoutenburg: Acquiring advanced properties in ontology mapping, PIKM 2008, pp. 9-16.

[WSWA-08] Yang Wang, Huaiming Song, Weiping Wang, Mingyuan An: A microscopic view on community detection in complex networks, PIKM 2008, pp. 57-64.

[W-08] Lei Wu: Error recovery in human-computer interaction: a preliminary study in a database learning environment. 93-96.

About the author:

Dr. Aparna Varde is a Tenure Track Assistant Professor in the Department of Computer Science at Montclair State University, NJ, USA. She obtained her Ph.D. and M.S. in Computer Science, both from Worcester Polytechnic Institute in Massachusetts, USA and her B.E. in Computer Engineering from the University of Bombay, India. Dr. Varde has been a Visiting Senior Researcher at the Max Planck Institute for Informatics Germany; and a Tenure Track Assistant Professor in the Department of Math and Computer Science at Virginia State University, USA. Her research interests span Data Mining, Artificial Intelligence and Database Management with particular emphasis on multi-disciplinary work. Dr. Varde's research has led to several publications in reputed international journals and conferences as well as trademarked software tools. Her professional activities include serving as a Panelist for NSF; a Reviewer for journals such as the VLDB journal, IEEE's TKDE, Elsevier's DKE and IEEE's Intelligent Systems; Co-chair of ACM CIKM's Ph.D. workshops; Program Committee Member in conferences such as IEEE's ICDM 2008, Springer's DEXA 2008, EDBT 2009 and ER 2009, and SIAM's SDM 2008. Dr. Varde has worked in the corporate world as a Computer Engineer in multi-national companies such as Lucent Technologies and Citicorp. She is currently working on research projects in Text Mining, Scientific Data Mining, Web Databases and Machine Learning.