Advances in Information and Knowledge Management

Aparna Varde
Department of Math and Computer Science
Virginia State University
Petersburg, VA, USA
avarde@vsu.edu
http://dolphin.vsu.edu/~aparna/

Jian Pei
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
jpei@cs.sfu.ca
http://www.cs.sfu.ca/~jpei/

Abstract

Several research areas today overlap between the tracks of databases, information retrieval and knowledge management, such as natural language processing, semantic web, digital libraries, visualization, information quality and data mining. Inter-disciplinary research across these tracks encourages advances in the development of databases, the extraction of information and the discovery of knowledge. This is precisely the focus of our article. We explain the research issues addressed in a Ph.D. workshop recently held at the ACM Conference on Information and Knowledge Management. This workshop had presentations on novel ideas addressing challenges in information and knowledge management. It covered a broad range of topics such as XML architectures, sensor data streams, personal information managers and text pre-processing. In this article, we provide an overview of the research problems and solutions discussed in the Ph.D. workshop. Our article thus describes the latest technological developments in information and knowledge management as seen by academia. This cutting edge technology also finds practical applications in the corporate world.

1 Introduction

PIKM 2007 was the first Ph.D. Workshop in CIKM, the ACM Conference on Information and Knowledge Management [1]. The goal of this workshop was to encourage Ph.D. students to present their dissertation research at a relatively early stage. The targeted students were those with a focus in any of the CIKM research tracks, i.e., databases, information retrieval and knowledge management. Inter-disciplinary work across the three tracks was particularly encouraged.

We noticed that a broad range of topics are addressed in CIKM related to various issues in databases and information systems, information retrieval and semantic web, knowledge management and data mining. It was

thus realized that the CIKM conference with its confluence of several research tracks provides an excellent opportunity for learning about the latest developments in the respective fields. It offers a good environment for a Ph.D. workshop where graduate students can expose their research early in their academic career. Particularly, we thought that such a workshop would encourage further studies across the three major fields addressed by CIKM: databases, information retrieval and knowledge management. This motivated the need for PIKM.

It was observed that PIKM workshop indeed encouraged graduate students to present their dissertation proposals and ongoing research on their dissertation sub-problems. It gave them an opportunity to get feedback on their work from researchers worldwide. The reviewers' comments helped them assess and revise their work as needed. Selected papers were published in the workshop proceedings. This gave considerable recognition to the work of the students allowing them to showcase their research. Moreover, students attending the workshop got an opportunity to critique the work of their peers. This promoted an open discussion similar to that in a doctoral consortium.

A vast array of topics on databases, information retrieval and knowledge management were presented at the PIKM workshop. The areas of interest were similar to those at the CIKM main conference. Although this workshop was the first of its kind, we received over 40 submissions from all over the world which exceeded our expectations. After a very tough review process by a committee comprising international experts, approximately 20 papers were selected for publication in PIKM. These are listed as follows [2].

- *Natural Language Processing for Information Retrieval: the time is ripe (again)*; Matthew Lease.
- *Reusing Relational Sources for Semantic Information Access*; Lina Lubyte.
- *An Architecture for XML Information Retrieval in a Peer-to-Peer Environment*; Judith Winter and Oswald Drobnik.
- *Leveraging Semantic Technologies for Enterprise Search*; Gianluca Demartini.
- *Two-Level Classifier Combination Techniques: A New Paradigm for Network Intrusion Detection*; Morteza Analoui, Behrouz Minaei-Bidgoli and Mohammed Hossein Rezvani.
- *Visualization and Analysis of Large Graphs*; Eloïse Loubier, Wahiba Bahsoun and Bernard Dousset.
- *Mining Semantic Distance Between Corpus Terms*; Ahmad Sayed, Hakim Hacid and Djamel A. Zighed.
- *CUFRES: Clustering Using Fuzzy Representative Events Selection for the Fault Recognition Problem in Telecommunication Networks*; Jacques Bellec and M. Tahar Kechadi.
- *MaxProb and categorization of queries based on linguistic features*; Desire Kompaore and Josiane Mothe.
- *Exploiting Web 2.0 for Knowledge-Based Information Retrieval*; David Meline.
- *Incorporating Quality Aspects in Sensor Data Streams*; Anja Klein
- *Mutually Beneficial Learning with Application to Online News Classification*; Lei Wu, Zhiwei Li, Mingjing Li, Wei-Ying Ma and Nenghai Yu.
- *Temporal Constraints for Rule-based Event Processing*; Karen Walzer, Alexander Schill and Alexander Löser.
- *Managing Highly Correlated Semi-structured Data: Architectural Aspects of a Digital Archive*; Alf-Christian Schering, Holger Meyer and Andreas Heuer.
- *Towards Workload Shift Detection and Prediction for Autonomic Databases*; Marc Holze and Norbert Ritter.
- *Personal Digital Library: PIM through a 5S Perspective*; Yi Ma, Edward A. Fox and Marcos André Gonçalves.
- *Document Retrieval for Question Answering: A Quantitative Evaluation of Text Preprocessing*; Gracinda Carvalho David Martins de Matos and Vitor Rocio.
- *Combining Resources with Confidence Measures for Cross Language Information Retrieval*; Youssef Kadri and Jian-Yun Nie.
- *Top-K Subgraph Matching in a Large Graph*; Lei Zhou, Lei Chen and Yansheng Lu.

- *Formalizing Ontology Reconciliation Techniques as a Basis for Meaningful Mediation in Service Related Tasks*; Patricio de Alencar Silva, Cláudia M. F. A. Ribeiro and Ulrich Schiel.
- *Possibility and Necessity Measures for Relevance Assessment*; Fatma Zohra Bessai Mechmache, Mohand Boughanem and Zaia Alimazighi.
- *Collaborative Framework for Indigenous Knowledge Management*; Theodora Mwebesa Twongyirwe Mondo, Venansius Baryamureeba and Ddembe Williams.
- *Webview Selection from User Access Patterns*; Samia Saidi, Yahya Slimani and Khedija Arour.

In the following section, we give an overview of the cutting edge research problems and solutions presented at this workshop.

2 Cutting Edge Research

The research presented at PIKM spanned a myriad of topics: natural language processing, data visualization, streaming data management, multilingual information processing and more. We divided the workshop into four potpourri sessions in order to sustain the interests of an audience with a multitude of backgrounds. Thus, each session instead of being focused on one topic was an assortment of various topics since many of the papers overlapped multiple tracks, as suitable for the CIKM conference. We present a brief overview of the papers, session by session, as referenced in the conference proceedings [2].

2.1. Session One

The first paper in the workshop was by Matthew Lease. This paper proposed novel techniques for the integration of existing information retrieval techniques with modern natural language processing approaches with an interesting application to conversational speech. It was proven that better use of the language processing features led to an improvement in text as well as speech retrieval.

Lina Lubyte's paper described the reuse of relational information sources for the purpose of semantic access to information. This paper outlined an ontological framework proposing heuristics based on schema design and normalization for wrapping the hidden semantics extracted from relational data sources using an automated approach. It was found that this approach indeed preserved database semantic constraints.

The paper by Winter and Drobnik was a unique integration of XML information retrieval and peer-to-peer networks. They designed a search engine for XML documents, enhancing current information retrieval methods by distributing documents and global indices over a peer-to-peer network. This architecture provided a storage space that was virtually unlimited and improved relevance detection using structural information. They used this generic architecture to build a more specific component-structured architecture for relevance computation in dynamic XML documents.

Demartini's paper approached the problem of enterprise search which builds over a regular web search in terms of the user perspective and the link structure. The search method proposed in this paper combined the use of techniques in user modeling, information retrieval and semantic web in order to enhance the state-of-the-art in the field of enterprise search.

2.2. Session Two

In this session the area of ensemble learning in artificial intelligence was addressed in the work of Analoui et al. for intrusion detection in networks. They proposed to use multiple classifiers to reduce false alarms and error rates in addition to classifiers trained on different feature sets to provide better results compared to individual classifiers. This hierarchical two-tier combination of classifiers based on ensemble learning was used for detecting intrusion in networks leading to the optimization of recognition capabilities and hence enhancement of performance.

Loubier et al. discussed the problem of visualization with respect to huge complex graphs for information representation in knowledge engineering. They developed a tool called VisuGraph to simplify and analyze such large graphs by using the Markov clustering algorithm and by time-slicing a global graph.

In another paper, El Sayed et al. approached an issue pertaining to similarity measures between terms in a corpus with reference to context. They proposed knowledge-based approach using taxonomy along with data mining techniques for deriving semantic similarity measures taking into account the targeted domain. This paper solved problems related to context-dependency and coverage in corpus terms. The proposed approach was found to be effective in computing the required semantic distances as corroborated by experimental evaluation.

Bellec and Kechadi's paper also dealt with a problem in data mining, more specifically, clustering. They introduced a clustering algorithm called CUFRES for fault identification in telecommunication networks. This was their original algorithm based on fuzzy representative events selection and was found to improve fault detection when compared to other clustering algorithms and also compared to alternative approaches such as false alarms. This was proved by evaluation with network data from Ericsson.

The paper by Komapore and Mothe addressed a linguistic problem dealing with query categorization based on probability. They proposed a fusion technique based on classifying queries from linguistic features that are automatically extracted and used it to rank documents with respect to relevance. Their approach was observed to provide better results than state-of-the-art methods.
David Milne's paper on knowledge based information retrieval described the development of a search engine called Koru that exploits Web 2.0 based techniques. This paper investigated a method to provide automated knowledge bases without requiring computers to replace human indexers. It demonstrated the use of Wikipedia to provide manually refined inexpensive knowledge bases catered to the topics, semantics and terms in document collections. The system in this paper aimed to be a practical tool to assist real users in information retrieval tasks.

2.3. Session Three

This session was opened by Anja Klein with a paper on data quality in sensors. This paper focused on the problem of assuring quality of service in streaming data coming from sensors by proposing a jumping-window-based approach for efficient information transfer and a flexible meta model to store and propagate data quality. It included a comprehensive analysis of common data processing operators with respect to their effect on data quality, provided a fruitful knowledge evaluation and hence reduced erroneous business decisions.

Wu et al. discussed a machine learning problem pertaining to online news classification. They proposed a classification framework known as Mutually Beneficial Learning. This involved an iterative two-step process of first discovering the local structures of a feature space in order to resist noisy samples and then applying a consecutive classification process to the result, until a stopping criterion is met. It was observed from their experiments that their framework gave better results than approaches such as Naïve Bayes and SVM even with noisy and partly labeled data.

The issue of temporal constraints in complex event processing was addressed by Walzer et al. They considered the popular Rete algorithm used in rule-based systems and proposed an extension to it in order to support temporal operators using interval time semantics. A description language was used to specify the patterns of interest. Applications of this work included supply chain management for RFID, systems monitoring and stock market analysis.

Schering et al. pursued the topic of semi-structured data management in a digital archive. They went beyond traditional XML techniques in managing hierarchical semi-structured data, addressing the problems posed in retrieval, manipulation and storage of such data due to a lack of efficient query languages. They focused on a

project called the Digital Wossidlo Archive dealing with a huge number of arbitrarily correlated data units and introduced an approach to find a solution for the concerned problems with respect to this information system.

Autonomic databases were approached in the work of Holze and Ritter in their paper on detecting and predicting workload shift. They gave a broad analysis of the parameters influencing the self-management of an autonomic database, pointing out that workload has a considerable impact on both physical design of data and the configuration of the database management system. They proposed a workload model for light-weight, continuous workload monitoring and analysis to be used in the identification and prediction of workload shifts, that need autonomic re-configuration of the database.

Personal information management research, more specifically a personal digital library, was presented in the work of Ma et al. Their proposed framework involved a formal definition components and functionalities for a minimal personal digital library taking into account issues such as information overload and fragmentation problems occurring with large data sets in digital formats. They described their results of implementing the receptor module and behavioural information relevance and discussed potential challenges anticipated with suggestions for possible solutions.

Carvalho et al. approached the problem of text pre-processing in question answering. They focused on the various options of preprocessing Portuguese text before feeding it to the information retrieval component, evaluating the effect on the performance in the specific context of question answering in order to make a sustained choice of options. They inferred the advantage of the basic pre-processing techniques: case folding and removal of punctuation marks. Another interesting observation was that stop word removal improved performance but Stemming and Lemmatization did not. Their work is likely to be useful in organizations such as the Cross Language Evaluation Forum.

2.4. Session Four

More work on multilingual research was found in this session. Kadri and Nie presented a paper on cross language information retrieval, more specifically, query translation. They proposed confidence measures to adjust the initial scores of translations and to create a weight of the same nature for translations with different resources. They tested their technique on two language collections and got much better results than existing methods such as translation by simple linear combination.

Zhou et al. addressed a problem of sub-graph matching within large graphs using a top-k approach. They computed a score function defined as the sum of the pair-wise similarity between a vertex in the sub-graph and its matching vertex in the main graph. Based on this they proposed an approach called the Ranked Matching algorithm for efficient querying over sub-graphs. Due to efficiency of its pruning strategy, their approach was found to be fast and outperformed state-of-the-art methods by many orders of magnitude. Its applications include querying in biological and social networks.

Issues related to ontology were discussed by Silva et al. in their work on mediation in service related tasks. They gave a formal description of ontology reconciliation techniques, e.g., merging, alignment and integration, along with an explanation of their scope within semantic web services architecture. Their work was a step towards overcoming the Tower of Babel effect in communication across different applications brought about by the use of varying ontology. Their reconciliation methods provided more meaningful results in service oriented mediation.

Relevance assessment measures were described in the work of Mechmache et al. in information retrieval on semi-structured documents. They propose a model based on a concept called possibilistic networks where the relationships “document-elements” and “elements-terms” are depicted using measures of possibility and necessity. User queries start a process of propagation to recover documents or parts of documents necessarily or at least possibly relevant. They present illustrative examples of their approach in the paper.

Mwebesa et al. propose a framework based on collaboration for the management of indigenous knowledge, which forms an important part of the history and culture of local communities. There is a need to learn from local communities to enrich knowledge management. Knowledge is stored in people's memories and activities. It is expressed and communicated orally which posses a serious threat to its preservation and development. Collaborative frameworks are therefore useful in solving such problems where the effectiveness of the knowledge depends on the nature of the interaction. It is found in their work that collaboration allows for a better pooling of resources and sharing of experiences on indigenous knowledge, among individuals as well as organizations.

The issue of web usage mining was addressed by Samia Saidi in the last paper presented at this workshop. In this paper, an approach was proposed for selecting web views to be materialized in order to optimize the response time of web queries. The web view selection was mainly based on the estimation of metrics requiring hard collects of multiple statistics and mining on an interesting set of views from realistic data, i.e., web log files. Web log files were parsed, analyzed and treated to give a set of views, based on frequent closed item-sets. This approach was found to offer satisfactory performance in the concerned systems.

Thus, we had four very interesting sessions in PIKM with a plethora of research ideas and their implementations. This represented challenging work in the areas of database management, information retrieval and knowledge management that are of interest to CIKM.

3 Summary and Conclusions

The authors of PIKM papers, in addition to presenting their dissertation proposals outlining solutions to specific problems, also discussed open issues pertaining to their areas. For example, an important issue in both natural language processing and semi-structured information retrieval is the inclusion of pragmatics, i.e., world knowledge to augment semantics which relates to domain-specific knowledge. Another issue, applying to several areas, is the evaluation of proposed research ideas with real as opposed to synthetic data sets, given that real data may be a lot harder to obtain, and taking into account criteria such as privacy and availability. Some of the papers did use real data while in many other papers this was offered as a suggestion. Other issues included terminology related aspects in papers addressing ontology, indexing strategies in the papers on database management and other similar considerations. Several suggestions offered by the attendees of the workshop were found useful by the Ph.D. candidates for ongoing work on their dissertation sub-problems. Much of the research discussed in PIKM presented a vision for the future.

In summarizing the contributions of this workshop, we can make the claim that PIKM 2007 was a grand success. It was a notable mention among the highlights of CIKM. The conference organizers and attendees highly applauded the event and invited us to propose the workshop again next year. The best paper award in PIKM 2007 as judged by the reviewers and attendees of the conference went to Matthew Lease for his much applauded work in the area of natural language processing.

We earnestly wish that this Ph.D. workshop continues to be a successful event year after year in the CIKM conference. Based on the suggestions offered in PIKM 2007, it is expected that the forthcoming workshops will be even better in terms of organization and presentation. We also hope that PIKM serves as an inspiration for many more events of a similar nature.

4 Acknowledgments

We express our gratitude towards the CIKM 2007 organizers who have been very helpful in publicizing the PIKM workshop. In particular, we thank the Workshops Chair, Professor Dongwon Lee from Penn State University, USA; and the Conference Chair, Professor Mario J. Silva, the Local Arrangements Chair, Professor Francisco Couto and the Proceedings Chair, Professor Andre Falcao all from Universidade de Lisboa, Portugal.

Moreover, we sincerely thank the Steering Committee and Program Committee of PIKM 2007 for their role in organizing this workshop. Their inputs in proposing PIKM and their expertise in conducting peer reviews

helped maintain a high standard for the workshop as suitable for the CIKM conference. We list the names of the PIKM committee members below along with their affiliations.

Steering Committee

- Chabane Djerba, University of Science and Technology of Lille, France
- Daniel Keim, University of Konstanz, Germany
- Helena Galhardas, INESC-ID and Technical University of Lisbon, Portugal
- Johannes Gehrke, Cornell University, USA

Program Committee

- Fatma Bouali, Lille 2 University, France
- Ernest Friedman-Hill, Sandia National Labs, USA
- Edward Fox, Virginia Tech, USA
- Sreenivas Gollapudi, Microsoft Search Labs, USA
- Jiawei Han, University of Illinois at Urbana-Champaign, USA
- Vagelis Hristidis, Florida International University, USA
- Giti Javidi, Virginia State University, USA
- Andreas Koeller, Oracle Corp., USA
- Shuhua Lai, Virginia State University, USA
- Pawan Lingras, St. Mary's University, Canada
- Bin Liu, Ajia Lighthorse Asset Management Inc., China
- Murali Mani, Worcester Polytechnic Institute, USA
- Amelie Marian, Rutgers University, USA
- Florent Masseglia, INRIA, France
- Shubhabrata Mukherjee, Virginia State University, USA
- Stephen North, AT&T Labs, USA
- Prasan Roy, Aster Data Systems, USA
- Carolina Ruiz, Worcester Polytechnic Institute, USA
- Jaideep Vaidya, Rutgers University, USA
- Li Xiong, Emory University, USA
- Helen Yang, Virginia State University, USA
- Mohammed Zaki, Rensselaer Polytechnic Institute, USA
- Zhongfei Zhang, Binghamton University, USA

5 References

[1] Mario J. Silva, Andre Falcao et al.: Proceedings of the ACM 16th Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6 – 9, 2007.

[2] Aparna S. Varde, Jian Pei: Proceedings of the First Ph.D. Workshop in CIKM, PIKM 2007, Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 9, 2007.