Databases, Information Retrieval and Knowledge Management: Exploring Paths and Crossing Bridges

Mouna Kacimi
Free University
of Bozen-Bolzano, Italy
Mouna.Kacimi@unibz.it

Fabian M. Suchanek
Max Planck Institute
for Informatics, Germany
suchanek@mpi-inf.mpg.de

Aparna Varde
Montclair State University
Montclair, NJ, USA
vardea@mail.montclair.edu

ABSTRACT

The International Conference on Information Retrieval and Knowledge Management (CIKM) brings together three avenues of data-oriented research, namely, Database Management, Information Retrieval and Knowledge Management. The confluence of these avenues becomes evident also in the PhD theses of doctoral students: Stream processing makes use of knowledge representation techniques, linked data is emerging as a research topic that bridges information retrieval and knowledge representation, and new forms of querying draw on techniques from both information retrieval and databases. In this paper, we survey new PhD theses at the meeting point of the three research avenues. Our survey is based on the 5th PhD workshop at the ACM CIKM conference. The topics include themes as diverse as link prediction, source code querying, and video stream processing.

1. INTRODUCTION

PIKM 2012 was the 5th workshop in a series of PhD workshops collocated with the International Conference on Information and Knowledge Management (CIKM). The goal of this workshop series is twofold. First, it gives doctoral students an opportunity to present their work on a global stage. This allows them to receive feedback from reviewers, from fellow students and from the general CIKM audience. Second, we believe that the CIKM research community, too, benefits from such a workshop: PhD theses are the grassroots of research. They point out new research avenues and provide fresh viewpoints from the researchers of tomorrow. Previous workshops have been held in 2007 [15], 2008 [14], 2010 [12], and 2011 [11].

This year's PIKM workshop attracted 19 submissions, of which 10 were accepted as full papers and 4 further submissions as poster papers. The PIKM has always had a poster session in order to promote interaction among student presenters and attendees. The topics of this year's papers included business process modeling, recommendation using linked data, collaborative Web search, link prediction, and querying scientific data. The PIKM honors the best submission to the workshop with a "Best Paper Award". The winner of the award is determined by the best reviews, after discussion in the program committee and the steering committee. This year's award went to the paper "When Big Data Leads to Lost Data" by V.M. Megler and David Maier [9].

Our experience with the preceding workshops allows us to carefully select our reviewers for the PIKM. We invite the reviewers who have provided reviews of outstanding quality for the last workshops. Thereby we can steadily raise the calibre of the reviews, and make them more helpful for the students. To further motivate reviewers to deliver high quality reviews, this year's PIKM introduced the "Best Reviewer Award" for the reviewer with the most helpful reviews. With this prize, we want to encourage reviewers to provide rich in-depth reviews, and reward the most engaged reviewer for his effort. This year, the honor went to Gerard de Melo (ICSC Berkeley, California, USA). The runners-up for the best reviewer award were Georgiana Ifrim (4C, Cork, Ireland), and Pierre Bourhis (Oxford University, UK). Also, we sincerely appreciate the efforts of our entire team of 21 reviewers comprising experts from academia and industry across the globe.

In the last years, the PIKM has always been able to attract a senior keynote speaker. This year's speaker was Ingmar Weber from Yahoo Research Barcelona/Spain. In a highly entertaining talk, Dr. Weber gave "Advice for Young Jedi Knights and PhD Students". The talk looked at the task of pursuing a PhD with both a humorous and a serious eye. The serious eye gave practical advice

on how to choose an advisor and a thesis topic, and looked at different successful approaches to doing research and to publishing. The humorous eye looked at whether education can contribute to personal happiness, the positive influence of chocolate on studying behavior and the effect of too much time spent on Facebook. Dr. Weber deserves our sincere thanks for making this talk a true highlight of the workshop.

The submissions to this PhD workshop spanned different areas of research. Even though much of the work presented at PIKM crosses multiple CIKM tracks, we broadly group the papers by their main area corresponding to each CIKM track, pointing out links to the other areas. The main areas are Database Systems, Knowledge Management, and Information Retrieval.

2. DATABASE SYSTEMS

Research in database systems explores advances in technology that supports data management, as well as the new problems that arise from such advances. The work leading to PIKM's best paper award [9] falls precisely in this category. It explores the popular area of cloud computing while addressing the issue of not finding exactly relevant information amidst the huge volume of data on the cloud. The paper is aptly titled "When big data leads to lost data". It proposes an approach to find relevant information from big scientific datasets by exploiting text retrieval techniques (such as interactive searching by asynchronous scanning for feature extraction) in the context of numeric data. This approach also uses information from a metadata catalog for feature extraction, scoring and ranking to enhance performance. The paper provides experimentation to convince the audience that the proposed approach is a feasible solution to the lost data problem. The primary area of this work is databases, but it bridges to the IR track due to its emphasis on techniques popular in text search. It briefly touches upon mining aspects by its interactive querying and ranking processes.

Other work in the database track of PIKM includes external source code querying, unified scientific data processing and data stream event detection. Garcia-Alvarado et al. [6] deal with the problem of querying multiple external program files that reference metadata, which requires asserting impacts of changes between the programs and databases. Since existing solutions to this problem are found to be highly complex, they propose algorithms analogous to keyword searches to analyze references between external programs and database schemas such that dependencies are preserved and changes in a source are reflected in another. Chamanara et al. [1] address the realm of scientific data management by proposing a query language called SciQL to perform scientific data processing and management in a unified manner. It gives scientists the means to express queries on data refinement, transformation, visualization and other tasks in a common format irrespective of physical data sources. Thus, scientists can use one language to interact with various sources such as plain text, Excel spreadsheets, RDBMS and MapReduce systems, making it easier to manage their research. The work of Weiler et al. [16], presented as a poster, focuses on streaming data, and, more specifically, on detecting events in real-time in data streams of high volumes. To achieve this, they consider HPC (high performance computing) coupled with continuous querying. They process queries over the data streams and fast queries over the respective historical data to detect events, classify them and rank them.

3. KNOWLEDGE MANAGEMENT

Knowledge management approaches deal with the extraction of information from data. They cover two main avenues of research, knowledge discovery and knowledge representation. Knowledge discovery is used in many real world applications to understand the meaning of data and predict its implicit properties. For example, Drzadzewski et. al. [3] help users grasp the topics of a new document collection. To this end, they propose a system that explores and analyzes clusters of documents using Online Analytical Processing (OLAP), an effective strategy for extracting and analysing views of data. The approach provides efficient and accurate techniques for creating and aggregating clusters, finding representative documents, finding relationships between clusters, and determining their strength. Another example for making implicit information explicit is given by the work of Xu et. al. [17]. The authors propose a feature selection strategy for link prediction in networks. The proposed algorithm is based on the discriminative abilities and the correlations between pairs of features. The best features would maximize the total discriminability score while minimizing the total correlation scores. Link prediction is central to many real world networks such as social media platforms.

Knowledge representation approaches, in contrast, deal with capturing the different aspects of data. One of the most important aspects is dynamicity, because it challenges the accuracy of models

learned from old data. Examples include stream data and dynamic business processes. Florez et. al., [4] address the problem of dynamicity in the context of video labeling. They propose an unsupervised framework based on topic modeling to represent the different activities of a video scene, and an algorithm to label such activities in previously unseen movies. This method can, e.g., detect and predict dangerous behavior of a car at a given place. Schütz et. al., [13] address the dynamic aspects of data in process modeling. They extend multilevel modeling approaches that capture the dependencies of processes belonging to different hierarchical levels within a company. The approach models life cycles using UML state machines and associates a life cycle model to each class for each level in the hierarchy.

4. INFORMATION RETRIEVAL

Information Retrieval is concerned with finding relevant Web documents for a user query. Several papers at the workshop investigate new dimensions of the field. Johansson et al. [7] propose to analyze queries not just per se, but in their temporal dimension and interaction. In winter, e.g, a search for "coughing" is more likely to indicate that the user is interested in remedies against a cold. In summer, the query is more likely to indicate interest in an allergy. Queries can also be studied in their temporal relation to each other. Queries about a vacation by car, e.g., are likely to be followed by queries about parking lots. Knäusel et al. [8] put forward the idea of studying what parts of a Web page a user is interested in. The authors present a user study that analyzes what parts of a Web page users read after having issued a query. The hypothesis is that we can aid the user by proactively selecting the interesting part of the Web page.

With the rise of the Web 2.0, Web search, and interaction on the Web in general, becomes more social. In this spirit, Correa et al. [2] analyze how user communities develop on Twitter. The authors analyzed thematic and social communities, based on the hashtags and @-tags that users employ in their tweets. This yields a social graph of users and their interests. Galán-García et al. present a more commercially oriented work [5]. The authors aim to find relevant advertisements for social network users, based on their chats. This approach has the advantage of customizing the ads to the real-time interests of users. The task is challenging because instant messages are usually in colloquial, often even faulty natural language. Meymandpour and Davis [10] present a research work that spans the Web 2.0 with the newly emerging Semantic Web. The goal is to extend the realm of social recommendation to the Semantic Web. Given that some users like some entities, how can we find other entities they like? The authors develop similarity metrics that use both content features and graph features for this purpose. Yue et al. [18] explore the setting of collaborative search in general, where multiple users work together to solve an information retrieval task. The authors conduct user experiments on the search platform CollabSearch, and use Hidden Markov Models to model how users find information in this environment.

5. CONCLUSION

The PIKM 2012 workshop showed us a wide variety of doctoral dissertation topics in the areas of databases, knowledge management and information retrieval. In PIKM 2011, the upcoming hot topic was research on linked data and social networks. While these topics were still prevalent at the PIKM 2012, the workshop focused more on Web data management and mining, especially considering the use of the cloud. Among cloud-based areas, the favorite domain was still social networks. In particular, various areas pertinent to search seemed to attract attention, both within social networks and otherwise, e.g., temporal query modeling, personalized advertisements, user preferences and video streams. In fact, enhancing search results by exploring technological advances database management, mining, and IR appeared to attract PhD students from all CIKM tracks. "Search" thus seemed to be the bridge that connected almost all the PIKM 2012 papers along each path.

6. REFERENCES

[1] Javad Chamanara and Birgitta König-Ries: SciQL: A Query Language for Unified Scientific Data Processing and Management. In PIKM 2012

[2] Denzil Correa, Ashish Sureka and Mayank Pundir: iTop: Interaction Based Topic Centric Community Discovery on Twitter. In PIKM 2012

[3] Grzegorz Drzadzewski and Frank Tompa: Exploring and Analyzing Documents with Online Analytical Processing. In PIKM 2012

[4] Omar U Florez and Curtis Dyreson: Is That Scene Dangerous? Transferring Knowledge Over a Video Stream. In PIKM 2012

[5] Patxi Galán-García, Carlos Laorden and Pablo G. Bringas: Towards a More Efficient

and Personalized Advertisement Content in On-line Social Networks. In PIKM 2012
[6] Carlos Garcia-Alvarado and Carlos Ordonez: Querying External Source Code Files of Programs Connecting to a Relational Database. In PIKM 2012
[7] Fredrik Johansson, Tobias Färdig, Vinay Jethava and Svetoslav Marinov: Intent-Aware Temporal Query Modeling for Keyword Suggestion. In PIKM 2012
[8] Hanna Knäusl and Bernd Ludwig: Assessing the Relationship between Context, User Preferences, and Content in Search Behavior. In PIKM 2012
[9] V.M. Megler and David Maier: When Big Data Leads to Lost Data. In PIKM 2012
[10] Rouzbeh Meymandpour and Joseph Davis: Recommendation Using Linked Data. In PIKM 2012
[11] Anisoara Nica, Fabian M. Suchanek, Aparna S. Varde: New Research Directions in Knowledge Discovery and Allied Spheres. Submitted to ACM SIGKDD Explorations 2012
[12] Anisoara Nica, Fabian M. Suchanek, Aparna S. Varde: Emerging multidisciplinary research across database management systems. SIGMOD Record 39(3): 33-36, 2010
[13] Christoph Schütz, Michael Schrefl and Lois Delcambre: Multilevel Business Process Modeling: Motivation, Approach, Design Issues and Applications. In PIKM 2012
[14] Aparna S. Varde: Challenging research issues in data mining, databases and information retrieval. In SIGKDD Explorations 11(1): 49-52, 2009
[15] Aparna Varde and Jian Pei: Advances in Information and Knowledge Management. In SIGIR Forum Journal, 42(1): 29-35, 2008
[16] Andreas Weiler, Svetlana Mansmann and Marc Scholl: Towards an Advanced System for Real-Time Event Detection in High Volume Data Streams. In PIKM 2012
[17] Ye Xu and Dan Rockmore: Feature Selection for Link Prediction. In PIKM 2012
[18] Zhen Yue, Shuguang Han, Jiepu Jiang, and Daqing He: Search Tactics as Means of Examining Search Processes in Collaborative Exploratory Web Search. In PIKM 2012