Dissertation Research Problems in Data Management and Related Areas

Gerard de Melo
IIIS
Tsinghua University
Beijing, China
gdm@demelo.org

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

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

ABSTRACT

Databases and related fields such as Information Retrieval, Data Mining and Knowledge Management offer many topics of interest for dissertation research. Specific areas include, for instance, big data, social networks, Web question answering and interactive knowledge discovery. In this article, we provide a summary and critique of research problems presented in these and related areas at a workshop on dissertation proposals and early doctoral research.

Keywords

Wikipedia, IR Evaluation, Exploratory Search, Query Processing, Recommender Systems, Rule Mining

1. INTRODUCTION

New research trends are often best observed in the research topics of doctoral candidates, who benefit from the experience of their advisors but add a fresh perspective. Doctoral consortia, or PhD workshops, have emerged as useful forums for dissemination of student research at an early stage in the course of a PhD. They serve the purpose of enabling young scholars to solicit feedback from world-renowned experts and to publish dissertation proposals and initial research results, which can be indicative of emerging challenges and directions in the community. The PhD Workshop in Information and Knowledge Management (PIKM) has been co-hosted with the ACM Conference on Information and Knowledge Management (CIKM) ever since 2007. PIKM 2014, the 7th Edition, was collocated with CIKM 2014 in Shanghai, China [1]. This article presents the outcomes of this workshop.

The PIKM 2014 workshop had a regular paper track and a short paper track, both with oral and poster presentations. This was in order to increase interaction between the presenters and the audience. A notable highlight of PIKM 2014 was a special track with invited talks and papers by more experienced researchers in addition to a keynote speaker, providing additional guidance and advice to early PhD students.

The keynote speaker was Iadh Ounis, faculty member at the University of Glasgow, UK, who spoke about creating and refining PhD Thesis Statements. Among other points, he forcefully argued that a thesis is meant to spark debate and should thus include statements that could potentially raise further questions. He emphasized that instead of including statements of fact, a thesis should include statements that arouse curiosity, thereby propelling readers to study the dissertation in detail and also inspiring future research in interesting sub-problems emerging from the dissertation. This talk was found extremely useful to PhD students who received practical advice for writing and polishing their dissertation.

The best paper award went to Arunav Mishra from Max Planck Institute for Informatics, Germany for his work on "Linking Today's Wikipedia and News from the Past". This is summarized in Section 3 and more details can be found in the PIKM proceedings [1]. In recent years, PIKM has also been announcing an award for the best reviewer to recognize outstanding contributions by a PC member. The best reviewer for PIKM 2014 was Fabian Suchanek from Télécom ParisTech, France. The program committee team consisted of 23 reviewers from across the globe, spanning 16 countries and 6 continents, with a healthy mix of academia and industry.

Considering these highlights of PIKM 2014, we now present a summary and critique of the research contributions in the forthcoming sections. Section 2 covers invited papers, the topics being social network recommendation methods, interactive mining for local and global association rules and knowledge base rule mining respectively. The slides for these invited talks are available online[1]. Section 3 focuses on regular papers in the areas of Wikipedia and news, evaluation methods, search with modeling and efficient query processing. Section 4 deals with short papers, the two

[1] http://iiis.tsinghua.edu.cn/~weblt/pikm2014/

themes being question answering and outlier detection. Further details on all of this research are available in the PIKM proceedings [1]. Finally, Section 5 describes conclusions and ongoing work.

2. RECOMMENDERS & RULE MINING

2.1 Recommendations in Social Networks

Richi Nayak's invited talk focused on the highly topical issue of recommendation in online dating portals [2]. Conventional recommendation engines work in one direction, recommending objects to users based on their interests. In social networks, however, the interest needs to be mutual, so a form of two-way recommendation is needed. This is specifically challenging in online dating platforms, where some users may enter very specific requirements, perhaps even an ideal "Prince Charming" that no real person in the database can live up to, while others just provide broad categories like "blonde hair" or a popular kind of music taste, which could match many thousands of candidate profiles.

Nayak, a faculty member at Queensland University of Technology, Australia, addresses this issue by selecting different recommendation strategies based on how people are using the platform, distinguishing highly active users from infrequent posters, for instance [2]. These strategies can account for patterns observed in user profile information as well as in user activity logs. As a preprocessing step, co-clustering is used to improve the scalability of the recommendation engine.

2.2 Interactive Mining

As an ABD candidate looking forward to his PhD, Abhishek Mukherji, from Samsung Research, USA (in joint work with Elke A. Rundensteiner and Matthew O. Ward from WPI, USA), discussed results on interfaces that enable association rule mining to be conducted in an interactive manner [3]. Association rules capture salient correlations between items in a data source, e.g. "people who buy dips (tend to) also buy chips". Rule mining has a long history and analysts frequently study such rules in order to improve their business.

In practice, however, this can be very tedious without the right tools, often due to dependencies between rules (e.g. one being a special case of another) resulting in countless near-duplicates and due to different levels of confidence and statistical support. Mukherji et al. proposed new techniques and user interfaces that make this process much easier for the analyst. So-called local patterns, which apply only to specific subsets of the data, are a particular focus in his work [3]. For instance, the analyst might be interested in salary trends that only appear in a particular geographic region and demographic. Efficient algorithms are necessary in order to be able to compute relevant rules in a short amount of time and facilitate interactive exploration without long waiting times. In his recent work at Samsung Research, Mukherji is applying similar techniques to mine interesting patterns of mobile device usage.

2.3 Rule Mining in Knowledge Bases

Luis Galárraga is a doctoral student at Télécom ParisTech, France, and has already published several top papers, including the Best Student Paper at WWW 2013 [4]. His research considers rule mining on collections of knowledge about the world. One might discover, e.g., a rule stating that a person is likely to live in the same city as their spouse. Such a rule can be interesting in itself, or could be used to fill the gaps when information is missing in a database. This is an important task because even the largest available knowledge bases are known to be very incomplete.

Galárraga's research proposes novel techniques to assess the confidence of rules in this setting, overcoming some of the problems of the traditional closed world assumption, according to which any knowledge not in the database is assumed to be false. This assumption cannot hold true in large open-domain knowledge bases. Galárraga thus proposes the alternative Partial Completeness Assumption. Moreover, he presents scalable techniques to find such rules in very large knowledge collections, yielding results on big popular knowledge bases such as YAGO2 and DBpedia in mere minutes. The same method can also be used to connect different knowledge sources, even when these connections are more complex than mere one-to-one alignments.

3. WIKIPEDIA, EVALUATION, SEARCH AND QUERYING

3.1 Wikipedia and News

To increase user satisfaction about the results of Information Retrieval systems, an interesting approach was proposed by A. Mishra. This aimed at combining different information sources to enrich knowledge about events. More specifically, it focused on Wikipedia and news articles, which provide different levels of description about events [1]. While Wikipedia excerpts describe events in an abstract form omitting details, news articles may describe events in an overly detailed form, missing the overall picture. Thus, the goal was to combine these two sources by creating a link from any Wikipedia excerpt to a matching set of news articles and vice versa. The proposed approach modeled the problem as an IR problem. It exploited two text collections, the first one being a collection of news articles and the second one a collection of Wikipedia excerpts. For the first corpus, the query was a Wikipedia excerpt and for the second one the query was a news article. The authors demonstrated that unrelated Wikipedia excerpts and news articles may use the same vocabulary, and thus a keyword-based retrieval strategy delivered only mediocre results. To overcome this, they developed a new strategy that added timestamps to both Wikipedia excerpts and news articles. These timestamps were used to compute a distribution of time expressions in the top-k documents and then re-rank the entire result list by boosting those that have similar time expressions. Future work on text mining and entity resolution was considered by the authors to improve the quality of the results.

3.2 Robust and Reusable Evaluation

The importance of understanding a user's information need to improve the quality of exploratory search was emphasized in the paper by K. Athukorala [1]. The main challenge of this research was that user knowledge and needs changed as the search progressed, which required adequate prediction of relevant results to evolving user intents. The authors approached this problem by making an exploratory study of the behavior of academics in searching information. This application captured the essence of exploratory search, since scientific searches often dealt with the discovery of unfamiliar topics. The authors developed a formal model to represent the state of exploration using observable aspects of user behavior, including viewed search results and clicking actions. This model was then used to predict the relevance of search results to current user interests and knowledge. Further steps were considered to improve the prediction power of such a model by exploiting other implicit interaction data e.g., read-time, click-time, scroll length, and gaze distribution over results.

3.3 Exploratory Search through Modeling

The author K. Hui focused on the evaluation of Information Retrieval systems [1]. Currently, the evaluation of such systems is performed through manual assessment, where the result documents are labeled to indicate their degree of relevance for the query. These labels, however, are associated to the entire document and do not correspond to its content. Consequently, manual assessment can hardly be extended to unlabeled parts of the document collection. Moreover, it is very expensive and cannot be applied to large scale datasets. To address this problem, the author presented a new evaluation strategy for diversity and novelty of search results. The proposed approach connected the evaluation results to the content of documents. It generated, for each sub-topic, a ground truth language modeled from a set of sufficient labeled documents. Thus, evaluation results could be re-used to assess future information retrieval systems even when human labeling was not possible.

3.4 Efficient Query Processing

Uysal et al. addressed the problem of efficient query processing in Information Retrieval systems that performed similarity search of multimedia content [1]. For that purpose, the authors considered a distance measure known as the Earth Mover's Distance (EMD). This distance measure assesses image dissimilarity in terms of the minimum amount of work needed to transform one feature representation into another one. The main advantage of this distance measure was its strong expressiveness of perceptual similarity and its applicability to both feature histograms and signatures. A major impediment to using this distance measure, however, had been its exponential time complexity with respect to increasing numbers of representatives. The authors focused on how to reduce the complexity of EMD after presenting the main challenges related to efficient query processing on feature signatures. They proposed a new lower bound Independent Minimization for Signatures (IM-Sig) to the EMD on feature signatures. This lower bound was regarded as an efficient filter approximation approach combined with k-nearest neighbor queries. The authors presented extensive experiments showing highly efficient results of the proposed approach.

4. QUESTION ANSWERING AND OUTLIER DETECTION

4.1 Question Expansion in QA Services

This paper was presented by Kyoungman Bae and Youngjoong Ko from Dong-A University, Busan, South Korea. It detailed a question expanding method to classify questions for question-answering (QA) services [1]. Input questions are mostly written with just a small portion of text, and, due to this fact, may not always give sufficient details for good classification. The authors thus proposed to expand the questions as follows. They obtained question-answer pairs pertaining to an input question with a search engine and selected top relevant words for expansion. They then generated pseudo answers adding question-related words using translation probabilities from questions to answers. Their preliminary experiments indicated that QA services provided better answers with this question expansion method.

4.2 Outlier Detection in Subspaces

Researchers Zhana Bao and Wataru Kameyama from Waseda University in Tokyo, Japan presented a novel outlier detection method. The authors explained that current methods find prominent outliers but neglect certain kinds of hidden ones [1]. The authors instead proposed a two-stage inspection model to detect outliers in different subspaces. The first stage measured neighboring density in subspaces to discover low dimensional outliers. The second stage assessed the degree of deviation of neighbors in joint subspaces. The authors statistically analyzed the results, merging them into a single score for each item, and candidate outliers were output as top-scoring objects. This work was evaluated on both synthetic and real data sets and was proven to be better than existing methods.

5. CONCLUSIONS

We observe a continued trend for young researchers to investigate Data Management issues arising in more specific settings and domains. Examples include news retrieval, user modeling, data mining, knowledge bases, and online dating. This emphasizes the importance of multidisciplinary work spanning Data Management that has extended its horizons to many fields within as well as beyond Computer Science.

Much of the work presented at PIKM presents significant potential for future research as well. For example, social network mining for online dating can be further optimized to include criteria such as minimizing search time or reducing the number of unsuccessful hits. News retrieval can be further enhanced by mining data on current trends to find the hot topics that interest specific user communities and displaying these in search engines. Web personalization can be conducted based on user modeling, thus providing better service to users in various applications such as product marketing.

The PIKM workshop provides an excellent forum for presentation of research ideas in early doctoral work. This has been a highly successful event since 2007. The organizers try to introduce interesting aspects to this workshop year after year. For example, the poster track was introduced in 2008, best reviewer awards have been given in some of the recent PIKMs, and this year we had a track with invited papers that included a mix of recent and experienced researchers to motivate early PhD students in several areas. The presenters of the invited papers were in addition to the keynote speaker. The keynote track has been in PIKM for quite a few years now and we have many prominent speakers give us very exciting and inspiring talks on topics that are useful to PhD students, over and above presenting their own research for further inspiration.

We sincerely hope that PIKM continues to be an important highlight of CIKM every year. This workshop certainly encourages PhD students to present their dissertation proposals and early doctoral research. It serves the dual purpose of publishing their work and getting feedback from a worldwide audience. It also helps meeting fellow students and researchers for collaborative opportunities, job prospects and friendships. Finally, it provides a unique perspective on research topics that are likely to grow in importance in data management and related areas.

6. ACKNOWLEDGMENTS

The authors would like to thank the CIKM 2014 organizers for giving us the opportunity to chair this PhD workshop. We convey sincere thanks to the keynote speaker Iadh Ounis and the presenters of the invited papers, Richi Nayak, Abhishek Mukherji and Luis Galárraga.

We express our gratitude towards all the PC members of the PIKM 2014 workshop for their special efforts in reviewing papers and thereby guiding young researchers based on their expertise.

Gerard de Melo's research is supported by China 973 Program Grants 2011CBA00300, 2011CBA00301, and NSFC Grants 61033001, 61361136003, 61450110088.

Aparna Varde's participation in PIKM was supported by funds from the CSAM (College of Science and Math) Dean's Office at Montclair State University, NJ, USA.

7. REFERENCES

[1] Gerard de Melo, Mouna Kacimi, Aparna S. Varde (Eds.): Proceedings of the 7th Workshop on Ph.D Students, PIKM at CIKM 2014, Shanghai, China, November 3, 2014. ACM, ISBN 978-1-4503-1481.

[2] Lin Chen, Richi Nayak. Leveraging the network information for evaluating answer quality in a collaborative question answering portal. Social Network Analysis and Mining 2(3), pp. 197-215, Springer, 2012.

[3] Abhishek Mukherji, Elke A. Rundensteiner, and Matthew O. Ward. COLARM: Cost-based optimization for localized association rule mining. In Proceedings of EDBT, pages 181–192, 2014.

[4] Luis Galárraga, Christina Teflioudi, Katja Hose, Fabian Suchanek. Association Rule Mining under Incomplete Evidence in Ontological Knowledge Bases. In Proceedings of WWW, pages 413-422, 2013.