New Research Directions in Knowledge Discovery and Allied Spheres

Anisoara Nica
Research and Development
Sybase, An SAP Company
Waterloo, ON, Canada
anisoara.nica@sybase.com

Fabian M. Suchanek
Otto Hahn Research Group "Ontologies"
Max Planck Institute for Informatics
Saarbrücken, Germany
fabian@suchanek.name

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

ABSTRACT

The realm of knowledge discovery extends across several allied spheres today. It encompasses database management areas such as data warehousing and schema versioning; information retrieval areas such as Web semantics and topic detection; and core data mining areas, e.g., knowledge based systems, uncertainty management, and time-series mining. This becomes particularly evident in the topics that Ph.D. students choose for their dissertation. As the grass roots of research, Ph.D. dissertations point out new avenues of research, and provide fresh viewpoints on combinations of known fields. In this article we overview some recently proposed developments in the domain of knowledge discovery and its related spheres. Our article is based on the topics presented at the doctoral workshop of the ACM Conference on Information and Knowledge Management, CIKM 2011.

Keywords

Ranking, Text Mining, Extreme Web, ETL, Pattern Recognition, Resource Monitoring, Version Control, KNN, Semantic Web, Main Memory Database, Data Warehousing, Database Analytics

1. INTRODUCTION

Knowledge discovery is an interdisciplinary field of research, which encompasses diverse areas such as data mining, database management, information retrieval, and information extraction. This inspires doctoral candidates to pursue research in and across these related disciplines, with the core contributions of their dissertation being in one or more of these areas. In this article, we review some of the directions that the researchers of tomorrow pursue. We provide a report of the research challenges addressed by students in the Ph.D. workshop PIKM 2011. This workshop was held at the ACM Conference on Information and Knowledge Management, CIKM 2011. The CIKM conference and the attached workshop encompass the tracks of data mining, databases and information retrieval, thus providing an excellent venue for dissertation proposals and early doctoral work in and across different spheres of knowledge discovery. This workshop was the fourth of its kind after three successful PIKM workshops in 2007 [15, 16], 2008 [11, 14] and 2010 [9, 10]. The PIKM 2011 [8] attracted submissions from several countries around the globe. After a review by a PC comprising 19 experts from academia and industry worldwide, 9 full papers were selected for oral presentation and 4 short papers for poster presentation. The program was divided into 4 sessions: data mining and knowledge management; databases; information retrieval; and a poster session with short papers in all tracks.

The first highlight of the PIKM 2011 was a keynote talk on "Extreme Web Data Integration" by Prof. Dr. Felix Naumann from the Hasso Plattner Institute in Potsdam, Germany [8]. This talk addressed the integration and querying of data from the Semantic Web at large scale, i.e., from vast sources such as DBpedia, Freebase, public domain government data, scientific data, and media data such as books and albums. It discussed the challenges related to the heterogeneity of Web data (even inside the Semantic Web), common ontology development, and multiple record linkage. It also highlighted the problems of Web data integration in general, such as identification of good quality sources, structured data creation, standardization-related cleaning, entity matching, and data fusion.

We now furnish a review comprising a summary and critique of the dissertation proposals presented at this workshop, discussing new directions of research in data mining and related areas. We follow the thematic structure of the workshop with 3 topic areas: knowledge discovery, database research, and information retrieval. The knowledge discovery issues surveyed in this article include areas as diverse as pattern recognition in time-series, resource monitoring with knowledge-based models, version control under uncertainty, and random walk k-nearest-neighbors (k-NN) for classification. The database research problems presented here entail aggregation for in-memory databases, evolving extract-transform-load (E-ETL) frameworks, schema and data versioning, and automatic regulatory compliance support. The information retrieval themes involve paradigms such as user interaction with polyrepresentation, ranking with entity relationship (ER) graphs, online conversation mining, sub-topical document structure, and cost optimization in test collections.

The workshop also issues a best paper award to the most exciting dissertation proposal, as determined by the PC of the workshop. This year's award went to the proposal "Ranking Objects by Following Paths in Entity-Relationship Graphs" [6], in the Information Retrieval track.

The rest of this article is organized as follows. Sections 2, 3, and 4 discuss the different tracks of PIKM, i.e., knowledge discovery, database research, and information retrieval, respectively. In Section 5, we summarize the hot topics of current research, and compare them with the topics of the previous PIKM workshops.

2. KNOWLEDGE DISCOVERY

The topics surveyed here are those with main contributions in data mining and knowledge discovery, although some of them overlap with the other two thematic tracks, namely, databases and information retrieval.

2.1 Pattern Recognition in Evolving Data

Many devices today, such as mobile phones, modern vehicular equipment and smart home monitors, contain integrated sensors.

These sensors need to capture information with reference to context (e.g., abnormal motor behavior in vehicles) in order to enable the device to adapt to change and cater to users. This entails dealing with temporal data that is evolving in the respective environment. Spiegel et al. [13] addressed this problem. They proposed efficient methods to recognize contextual patterns in continuously evolving temporal data. Their approach incorporated a machine learning paradigm with a three step process: extracting features to construct robust models capturing important data characteristics; determining homogenous intervals and point of change by segmentation; and grouping the time series segments into their respective subpopulation by clustering and classification. This problem was considered interesting by the audience especially due to its application in smart phones where the proposed approach is useful in detecting patterns such as changing product prices to provide better responses to queries.

2.2 Resource Monitoring with KM Models

Abele et al. [1] focused on a knowledge engineering problem in the manufacturing domain. More specifically, their problem was on computerizing the monitoring of resource consumption in complex industrial production plants, a task that usually involves tremendous time and manual effort. They proposed a semi-automated method for monitoring through knowledge based modeling with sensors, reasoners, annotations and rule engines, easily adaptable to changes. Their modeling approaches included object-oriented models with UML and domain models in the Semantic Web, with specific use of a data format for plant engineering called AutomationML. Advantages of their monitoring approach were reduction in manual effort, saving of time and resources, application independence and flexibility. Their research was particularly appreciated due to the manner in which they dealt with a domain-specific problem in an application independent manner by proposing knowledge based models that adequately encompassed ontology and logic through formalisms.

2.3 Version Control under Uncertainty

Collaborative work on documents poses the problem of version control in the presence of uncertainty. Ba et al. [2] tackled this issue with much attention to XML documents. They proposed a version-control paradigm based on XML data integration to evaluate data uncertainty and automatically perform conflict resolution. They outlined the main features of versioning systems and defined a version-space formalism to aid in data collaboration and in the management of uncertainty. In their proposed solution, they incorporated aspects such as XML differencing with respect to trees, delta models for operations like insertions, moves and updates, directed acyclic graphs of version derivations and construction of probabilistic XML documents that minimize uncertainty. This work was found highly appealing due to its contributions to data mining, databases and IR since it addressed the important problem of uncertainty in knowledge discovery, proposed models based on database theoretical concepts and applied these within the context of XML in information retrieval.

2.4 Random Walk k-NN for Classification

The challenging problem of multi-label classification formed the focus of the work by Xia et al. [19]. In this problem, an instance belonged to more than one class as predicted by the classifier and the number of potential class labels could be exponential, posing a challenge in predicting the target class. The authors proposed an approach to solve this problem by exploiting the benefits of k nearest neighbors and random walks. They first constructed a link graph based on k-NN, and then executed a random walk on the graph, so as to obtain a probability distribution of class label sets for the target instance. Further, they determined a classification threshold based on minimizing the Hamming Loss that computed the average binary classification error. Although this work did not show any significant experimentation, and thus drew criticism from the audience, it provided a complexity analysis with true and predicted class labels that was found acceptable. The authors also provided good motivation for their work by addressing real-world applications such as functional genomics, semantic scene classification and text categorization.

3. DATABASE RESEARCH PROBLEMS

The work presented at PIKM 2011 in the area of database research includes topics in database analytics, main memory databases, evolution of resources using ETL, and schema and data versioning.

3.1 Aggregation Strategies for Columnar In-memory Database Systems

Some of the hottest topics in current database research include in-memory database systems, columnar database systems, self-managing and self-adapting database systems in the presence of mixed workloads. These topics, and, in general, issues related to large in-memory database systems were discussed by Müller and Plattner in [7]. The main goal of the proposed work is to design an adaptive engine for aggregation materialization for database analytics. The paper addresses problems related to aggregation materialization identifying relevant cost factors for deciding when the materialization is cost efficient, and also what characteristics a cost model used by this type of adaptive engines must have. The authors provided an excellent motivation for this research by discussing recent trends in database usage where online transactional processing and online analytical processing are all reunified in one single database.

3.2 Managing Evolving ETL Processing

Traditional data warehouse systems employ an ETL process that integrates the external data sources into the data warehouse. One of the hardest problems in ETL research is the frequent changes in the schema of the external data sources. Wojciechowski [18] took up the challenge of automating the process of adapting and evolving the ETL process itself to the changes in the structural schemas of the external data sources. The proposed E-ETL framework brings together, in a unique way, techniques from materialized view adaptation, schema and data evolution, versioning of schema and data [17], and multiversion in data warehouse systems. The author presented the current implementation of the E-ETL system and future work on developing an appropriate language for defining ETL operations, as well as extensions to the current set of changes at the external data sources that E-ETL can detect and adapt to.

3.1 Schema and Data Versioning Systems

The topic of schema and data versioning systems, which was briefly referenced in [18] in the context of evolvable ETL processing, was the main topic of the paper by Wall and Angryk [17]. The authors investigate two different approaches to schema and data versioning: one approach is based on storing the minimal set of changes from the base schema and data to each branch, a sandbox, representing a particular deviation from the base schema The queries run against the branch are mapped accordingly to the base schema and its data, as well as to the branch schema and data to correctly reflect the set of changes in the branch. Another approach is to create copies of the modified tables into the branch, and propagate changes done in the base schema and data to these copies. The most interesting parts of this work are related to investigation into the qualitative and quantitative differences between the two techniques. The authors described the design of the ScaDaVer system meant to be used in assessing and testing different approaches to schema and data versioning. The topic and the motivation for this research are well established in the database area with new importance being brought to by the SaaS systems where multiple instances of the database can be operational in the same time.

3.2 Automatic Regulatory Compliance

A unique research on using semantic web technologies to support the management of the compliance systems was presented by Sapkota et al. in [12]. Compliance Management systems, although using computer-assisted processes, still lack one of the fundamental requirements, which is the automation towards updating the system when the regulations change. The paper proposed a Semantic Web methodology to automate these processes. The proposed techniques include automatic extraction and modeling of regulatory information, and mapping regulations to organizational internal processes. The authors describe the implementation of the proposed methodology which has been started using the Pharmaceutical industry as a case study, and is being applied to the Eudralex European regulation for good manufacturing practice in the pharmaceutical industry.

4. INFORMATION RETRIEVAL THEMES

The third thematic track of the PIKM workshop was concerned with topics in information retrieval. It comprised 2 poster papers and 3 full papers, including the best paper award winner (see Section 4.5).

4.1 User Interaction with Polyrepresentation

Zellhoefer et al. [20] introduce a new interactive information retrieval model. They present a prototypical GUI-based system that allows users to interactively define and narrow down the documents they are interested in. The novelty of the proposed approach lies in the inspiration from the cognitively motivated principle of polyrepresentation. The principle's core hypothesis is that a document is defined by different representations such as low-level features, textual content, and the user's context. Eventually, these representations can be used to form a cognitive overlap of features in which highly relevant documents are likely to be contained. In their work, the authors link this principle to a quantum logic-based retrieval model, which enhances its appeal. The work also addresses the issue of information need drifts, i.e., of adjustments of the information needs of the user. This gives the work a refreshingly practical angle.

4.2 Online Conversation Mining

Inches et al. [5] address the problem of data mining in online instant messaging services, twitter messages and blogs. The texts in these new areas of the social Web exhibit unique properties. In particular, the documents are short, user generated, and noisy. The authors investigate two different but related aspects of the content of these colloquial messages: the topic identification and the author identification tasks. They develop a framework in which social-Web specific features, such as emoticons, abbreviations, shoutings, and burstiness are systematically extracted from the documents. These features are used to build up a model of topic representation and author representation. The paper at the workshop described work in progress, with the author identification, and the synthesis of the features still to be explored.

4.3 Sub-topical document structure

In text segmentation, a document is decomposed into constituent subtopics. Ganguly et al. [3] analyze how text segmentation can help for information retrieval. This can happen, for example, by computing the score of a document as a combination of the retrieval scores of its constituent segments, or by exploiting the proximity of query terms in documents for ad-hoc search. Text segmentation can also help for question answering (QA), where retrieved passages from multiple documents are aggregated and presented as a single document to a searcher. Text segmentation can also help segmenting the query, if the query is a long piece of text. This is particularly important for patent prior art search tasks, where the query is an entire patent.

4.4 Cost Optimization in Test Collections

Information retrieval systems are usually evaluated by measuring the relevance of the retrieved documents on a test collection. This relevance has to be assessed manually. Since this is usually a costly process, Hosseini et al [4] consider the problem of optimally allocating human resources to construct the relevance judgments. In their setting, there is a large set of test queries, for each of which a large number of documents need to be judged, even though the available budget only permits to judge a subset of them. In their work, the authors propose a framework that treats the problem as an optimization problem. The authors design optimization functions and side constraints that ensure not only that the resources are allocated efficiently, but also that new, yet unseen systems can be evaluated with the previous relevance judgments. It also takes into account uncertainty that is due to human errors. This way, the proposal aims to tackle holistically a problem that is of principal importance in the area of information retrieval in general.

4.5 Ranking with ER graphs

The area of Information Retrieval is no longer restricted to text documents. It can equally well be applied to entity-relationship graphs or RDF knowledge bases. Kahng et al. [6] explore this idea by looking at paths in entity-relationship graphs. They put forward two ideas: First, an entity-relationship graph can represent not just factual information, but also other, additional, heterogeneous information. This allows treating tasks that have traditionally been seen as orthogonal within one single model. Second, the paper proposes to take into account the schema of the graph, and in particular the labels along the edges of paths. The paper shows that this allows treating tasks that have traditionally been seen as different, such as information retrieval and item recommendation, in the same model. This contribution earned the work the best paper award of the PIKM 2011.

5. CONCLUSIONS

In this article, we have looked at the area of knowledge discovery from the viewpoint of Ph.D. students. We have presented some promising research proposals, which treat not just the area of knowledge discovery but also the neighboring fields of database management and information retrieval.

PIKM 2011 was the fourth Ph.D. workshop in a series of such workshops. Looking back, we see that PIKM 2007 concentrated on topics such as fuzzy clustering, linguistic categorization, online classification, rule-based processing and collaborative knowledge management frameworks. In 2008, the focus was on social networking, text mining and speech information retrieval. In 2010, the research areas tilted towards security, quality and ranking, getting more interdisciplinary. In PIKM 2011, three new topics emerged: mining in social media [5], mining in the Semantic Web [6], and main memory databases [7]. We see this as a proof of the growing attraction of these domains. In addition, one theme that caught particular attention across multiple areas was the evolution of resources over time, be it in the area of ETL processing [18], in the area of XML [2], or in database versioning [17]. We also see an overarching topic of process management in general, in the sense of resource monitoring [1] and regulatory compliance [12], as well as in the sense of human cost optimization when producing IR test collections [4].

After four successful Ph.D. workshops in Information and Knowledge Management, the PIKMs in 2007, 2008, 2010 and 2011, we hope to continue these events in future conferences. We believe that such workshops benefit not just Ph.D. students, but also the research community as a whole, since Ph.D. thesis proposals point out new research avenues and provide fresh viewpoints from the researchers of tomorrow.

6. REFERENCES

[1] L. Abele, M. Kleinsteuber, and H. Thorbjoern. Resource monitoring in industrial production with knowledge-based models and rules. *PIKM 2011*, pp. 35 – 43.

[2] M. L. Ba, T. Abdessalem, and P. Senellart. Towards a version control model with uncertain data. *PIKM 2011*, pp. 43 – 50.

[3] D. Ganguly, J. Leveling, and G. J. Jones. Utilizing sub-topical structure of documents for information retrieval. *PIKM 2011*, pp. 75 – 78.

[4] M. Hosseini, I. Cox, and N. Milic-Frayling. Optimizing the cost of information retrieval test collections. *PIKM 2011*, pp. 79 – 82.

[5] G. Inches and F. Crestani. Online conversation mining for author characterization and topic identification. *PIKM 2011*, pp. 19 – 26.

[6] M. Kahng, S. Lee, and S.-G. Lee. Ranking objects by following paths in entity-relationship graphs. *PIKM 2011*, pp. 11 – 18.

[7] S. Müller and H. Plattner. Aggregation strategies for columnar in-memory databases in a mixed workload. *PIKM 2011*, pp. 51-57.

[8] A. Nica and F. M. Suchanek, editors. *Proceedings of the 4th Ph.D. Workshop in CIKM, PIKM 2011, 20th ACM Conference on Information and Knowledge Management, ACM CIKM 2011*, Glasgow, UK.

[9] A. Nica, F. M. Suchanek, and A. S. Varde. Emerging multidisciplinary research across database management systems. *SIGMOD Record*, 39(3):33–36, 2010.

[10] A. Nica and A. S. Varde, editors. *Proceedings of the Third Ph.D. Workshop in CIKM, PIKM 2010, 19th ACM Conference on Information and Knowledge Management, ACM CIKM 2010*, Toronto, Canada.

[11] P. Roy and A. S. Varde, editors. *Proceedings of the Second Ph.D. Workshop in CIKM, PIKM 2008, 18th ACM Conference on Information and Knowledge Management, ACM CIKM 2008*, Napa Valley, CA.

[12] K. Sapkota, A. Aldea, D. A. Duce, M. Younas, and R. Bãnares-Alćantara. Towards semantic methodologies for automatic regulatory compliance support. *PIKM 2011*, pp. 83 – 86.

[13] S. Spiegel, B.-J. Jain, E. W. D. Luca, and S. Albayrak. Pattern recognition in multivariate time series. *PIKM 2011*, pp. 27-33.

[14] A. S. Varde. Challenging research issues in data mining, databases and information retrieval. *SIGKDD Explorations*, 11(1):49–52, 2009.

[15] A. S. Varde and J. Pei. *Proceedings of the First Ph.D. Workshop in CIKM, PIKM 2007, Sixteenth ACM Conference on Information and Knowledge Management, ACM CIKM 2007*, Lisbon, Portugal.

[16] A. S. Varde and J. Pei. Advances in information and knowledge management. *SIGIR Forum*, 42(1):29–35, 2008.

[17] B. Wall and R. Angryk. Minimal data sets vs. synchronized data copies in a schema and data versioning system. *PIKM 2011*, pp. 67 – 73.

[18] A. Wojciechowski. E-ETL: Framework for managing evolving ETL processes. *PIKM 2011*, pp. 59-65.

[19] X. Xia, X. Yang, S. Li, C. Wu, and L. Zhou. RW.KNN: A proposed random walk KNN algorithm for multi-label classification. *PIKM 2011*, pp. 87 – 90.

[20] D. Zellhoefer and I. Schmitt. A user interaction model based on the principle of polyrepresentation. *PIKM 2011*, pp. 3 – 10.

About the authors:

Anisoara Nica holds a Ph.D. in Computer Science and Engineering from the University of Michigan, Ann Arbor, USA, 1999, with the dissertation in the areas of information integration systems and data warehousing. She is currently Distinguished Engineer in the SQL Anywhere Research and Development team of Sybase, An SAP Company, in Waterloo, Canada. Her research interests and expertise are focused on database management systems, in particular query processing and query optimization, data warehousing, distributed, mobile, and parallel databases. Her work experience includes Lawrence Berkeley National Laboratory and International Computer Science Institute in Berkeley. Dr. Nica holds eight patents and has several other patents pending. She has published more than 25 research articles, and she reviews for NSERC, ACM SIGMOD, IEEE ICDE, ACM CIKM.

Fabian Suchanek is the leader of the Otto Hahn Research Group "Ontologies" at the Max Planck Institute for Informatics in Germany. In his Ph.D. thesis, Fabian developed inter alia the YAGO-Ontology, earning him a honorable mention of the SIGMOD dissertation award. His interests include information extraction, automated reasoning, and ontologies in general. Fabian has published around 30 scientific articles, and he reviews for conferences and journals such as ACL, EDBT, WSDM, WWW, TKDE, TODS, AIJ, JWS, and AAAI.

Aparna Varde, Computer Science Faculty at Montclair State University (New Jersey), has a Ph.D. from WPI (Massachusetts) with a dissertation in the data mining area. Her interests include scientific data mining and text mining with projects in green IT, nanoscale analysis, markup languages, terminology evolution and collocation, overlapping AI & DB areas. She has 45 publications and 2 trademarks. Her students are supported by grants from PSE&G, NSF, Roche & Merck. She has served as a panelist for NSF; journal reviewer for TKDE, VLDBJ, DKE, KAIS and DMKD; and PC member at ICDM, SDM and EDBT conferences.