Guest Editors' Introduction: Special Section on Mining Large Uncertain and Probabilistic Databases

Reynold Cheng, *Member*, *IEEE*, Michael Chau, *Member*, *IEEE*,
Minos Garofalakis, *Member*, *IEEE*, and Jeffrey Xu Yu, *Member*, *IEEE*

Recent years have witnessed the emergence of novel database applications in various nontraditional domains, including location-based services, sensor networks, RFID systems, and biological and biometric databases. Traditionally, data mining has been widely used to reveal interesting patterns in the vast amounts of data generated by such applications. However, for most of these emerging domains, data is often riddled with uncertainty, arising, for instance, from inherent measurement inaccuracies, sampling and curation errors, and network latencies, or even from intentional blurring of the data (to preserve anonymity). Such forms of data uncertainty have to be handled carefully, or else the results of long and tedious data analyses could be inaccurate or even incorrect. In particular, it is important to collect and distill the knowledge from experts in developing mining and data processing methods that are uncertainty-aware. Recently, there has been active interest in the database community by treating uncertain data as a "first-class citizen," where the probability and statistical information of data are stored in the DBMS. Novel queries can be evaluated on these data to produce probabilistic results [1]. More recently, new mining algorithms that take into account data uncertainty, such as frequent pattern mining [2] and clustering [3], have also been proposed. Aggarwal and Yu [4] give a survey on the area of uncertain data querying and mining, where the key challenges include:

- models and structures for uncertain information in data mining and complex data analysis,
- association rule mining and clustering of uncertain data,
- machine learning in uncertain data,
- mining moving-object trajectories and biological data with noise,
- similarity matching of objects with uncertainty, and
- efficient mining and analysis of uncertain/probabilistic data streams.

This special section of the *IEEE Transactions on Knowledge and Data Engineering* features a collection of four papers, selected from 23 submissions, representing recent advances in the mining of uncertain databases. These works present new techniques for mining patterns, clustering, and ranking on uncertain data.

In applications like biological databases, graph data are often incomplete and imprecise. Mining uncertain graph data is semantically different from and computationally more challenging than mining exact graph data. The first paper, "Mining Frequent Subgraph Patterns from Uncertain Graph Data" by Zhaonian Zou, Jianzhong Li, Hong Gao, and Shuo Zhang, investigates the mining of frequent subgraph patterns from uncertain graph data. They propose an uncertain graph model, and formalize the subgraph mining problem, which is NP-hard. Thus, they develop an approximate and scalable algorithm.

The problem of clustering large uncertain location databases is investigated in "Clustering Uncertain Data Using Voronoi Diagrams and R-Tree Index," written by Ben Kao, Sau Dan Lee, Foris K.F. Lee, David W. Cheung, and Wai-Shing Ho. They show that the UK-means algorithm, which generalizes the k-means algorithm to handle uncertain objects, is inefficient due to a large amount of expensive expected distance calculation. They propose pruning techniques based on Voronoi diagrams to reduce the amount of expected distance calculation. These techniques are analytically proven to be more effective than the basic bounding-box-based technique. The authors also use an R-tree index to speed up the retrieval of the uncertain objects.

The third paper, "Scalable Probabilistic Similarity Ranking in Uncertain Databases" by Thomas Bernecker, Hans-Peter Kriegel, Nikos Mamoulis, Matthias Renz, and Andreas Zuefle, studies how to rank uncertain data according to their distances to a reference object. They propose a framework that incrementally computes, for each object instance and ranking position, the probability of the object falling at that ranking position. While existing approaches compute this probability distribution by using quadratic-complexity algorithms, the new algorithm requires linear time with the same memory requirements. They also show how the output of their method can be used to apply probabilistic top-k ranking for the objects according to different state-of-the-art definitions.

- *R. Cheng is with the Department of Computer Science, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: ckcheng@cs.hku.hk.*
- *M. Chau is with the School of Business, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: mchau@business.hku.hk.*
- *M. Garofalakis is with the Department of Electronic and Computer Engineering, Technical University of Crete, Greece. E-mail: minos@softnet.tuc.gr.*
- *J.X. Yu is with the Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: yu@se.cuhk.edu.hk.*

For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org.

1041-4347/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

In order to facilitate the querying and analysis of a large amount of uncertain data, new indexing techniques that are tailor-made for these data are highly desirable. In the fourth paper, "Effectively Indexing the Uncertain Space" by Ying Zhang, Xuemin Lin, Wenjie Zhang, Jianmin Wang, and Qianlu Lin, the authors propose a new R-tree-based inverted index called the URI-Tree. The new index efficiently supports a large variety of queries, such as range queries, similarity joins, and their size estimation, as well as top-k range query. Multidimensional uncertain objects with continuous and discrete distributions can also be handled.

In closing, we would like to thank the authors for their high-quality contributions to this special section and the referees for their generous help and valuable suggestions. We also like to thank Drs. Xindong Wu and Beng-Chin Ooi, the former and current Editor-in-Chief of *TKDE*, for giving us the opportunity to publish this special section.

Reynold Cheng
Michael Chau
Minos Garofalakis
Jeffrey Xu Yu
Guest Editors

REFERENCES

[1] R. Cheng, D. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data," *Proc. ACM SIGMOD*, 2003.

[2] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic Frequent Itemset Mining in Uncertain Databases," *Proc. 15th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2009.

[3] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K. Yip, "Efficient Clustering of Uncertain Data," *Proc. 22nd Int'l Conf. Data Eng.*, 2006.

[4] C. Aggarwal and P. Yu, "A Survey of Uncertain Data Algorithms and Applications," *IEEE Trans. Knowledge and Data Eng.*, vol. 21, no. 5, May 2009.

Reynold Cheng received the BEng degree in computer engineering and the MPhil degree in computer science and information systems from the University of Hong Kong (HKU) in 1998 and 2000, respectively, and the MSc and PhD degrees from the Department of Computer Science, Purdue University, in 2003 and 2005, respectively. He is an assistant professor in the Department of Computer Science at HKU. He was the recipient of the 2010 Research Output Prize in the Department of Computer Science at HKU. He is a keynote speaker in the First International Workshop on Quality of Context (QuaCon '09). From 2005 to 2008, he was an assistant professor in the Department of Computing at Hong Kong Polytechnic University, where he received two Performance Awards. He is a member of the IEEE, the ACM, ACM SIGMOD, and UPE. He has served on the program committees and review panels for leading database conferences and journals. He received an Outstanding Service Award at the CIKM 2009 conference. His research interests include database management, as well as querying and mining of uncertain data.

Michael Chau received the bachelor's degree in computer science and information systems from the University of Hong Kong in 1998, and the PhD degree in management information systems from the University of Arizona, Tucson, in 2003. He is currently an assistant professor in the School of Business at the University of Hong Kong. His current research interests include information retrieval, Web mining, data mining, knowledge management, electronic commerce, and security informatics. He has published more than 70 research articles in leading journals and conferences, including *Computer*, the *Journal of the America Society for Information Science and Technology*, the *Journal of the Association for Information Systems*, *Decision Support Systems*, and *Communications of the ACM*. He is a member of the IEEE, ACM, ASIST, and AIS and is the chairman of the Computational Intelligence Chapter of the IEEE (HK). He is the organizing chair or program chair of five conferences/workshops and a program committee member of more than 35 international conferences. More information can be found at http://www.business.hku.hk/~mchau/.

Minos Garofalakis received the diploma degree in computer engineering and informatics (School of Engineering Valedictorian) from the University of Patras, Greece, in 1992, and the MSc and PhD degrees in computer science from the University of Wisconsin-Madison in 1994 and 1998, respectively. He worked as a member of the technical staff at Bell Labs, Lucent Technologies in Murray Hill, New Jersey (1998-2005), as a senior researcher at Intel Research Berkeley in Berkeley, California (2005-2007), and as a principal research scientist at Yahoo! Research in Santa Clara, California (2007-2008). In parallel, he also held an adjunct associate professor position in the Electrical Engineering and Computer Sciences Department at the University of California, Berkeley (2006-2008). Since September 2008, he has been a professor of computer science in the Department of Electronic and Computer Engineering at the Technical University of Crete, and the director of the Software Technology and Network Applications Lab (SoftNet). He currently serves as an associate editor for the *ACM Transactions on Database Systems* and the *IEEE Transactions on Knowledge and Data Engineering*, and as an editorial board member for *Foundations and Trends in Databases* and the *Proceedings of the VLDB Endowment*. He was the Core Database Technology PC Chair for VLDB '07, and an associate editor for the *IEEE Data Engineering Bulletin* (2004-2006). His research interests include database systems, centralized and distributed data streams, data synopses and approximate query processing, probabilistic and uncertain databases, network-data management, XML/text databases, and data mining. His work has resulted in 33 US patent filings (15 patents issued) for companies such as Lucent, Yahoo!, and AT&T. Google Scholar gives more than 5,000 citations to his work, and an h-index value of 37. He is a senior member of the ACM, a member of the IEEE, and the recipient of the 2009 IEEE ICDE Best Paper Award, the Bell Labs President's Gold Award (2004), and the Bell Labs Teamwork Award (2003).

Jeffrey Xu Yu received the BE, ME and PhD degrees in computer science from the University of Tsukuba, Japan, in 1985, 1987, and 1990, respectively. Dr. Yu held teaching positions at the Institute of Information Sciences and Electronics, University of Tsukuba, Japan, and the Department of Computer Science, The Australian National University. Currently, he is a professor in the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong. Dr. Yu's current main research interests includes graph database, XML database, data mining, Web technology, and query processing and query optimization. He has published more than 200 papers that have been published in the *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, the *VLDB Journal*, the *ACM Transactions on Database Systems*, SIGMOD, SIGKDD, VLDB, ICDE, and EDBT. He was an associate editor of *TKDE*, and is a *VLDB Journal* editorial board member. He is a member of the IEEE.