Guest Editors Introduction: Special Section on Keyword Search on Structured Data

Surajit Chaudhuri, Yi Chen, and Jeffrey Xu Yu

With the prevalence of Web search engines, keyword search has become the most popular way for users to retrieve information from text documents. On the other hand, there is an enormous amount of valuable information stored in structured form (relational or semistructured) in Internet, intranet, and enterprise databases. To query such data sources, users traditionally depended on specialized applications because for most users it is difficult to use structured or semistructured query languages. In recent years, enterprise search has gained popularity where a keyword-based search model is used for intranet data sources. However, in most of these systems, the structured data objects that can be retrieved via keyword search have to be predefined.

The database research community has been focusing on developing some of the key technology that holds the promise of generalizing the reach of keyword search over structured and semistructured data beyond the state of the practice in commercial enterprise search engines. Some of the problems that have received attention include the task of automatically assembling a data object on the fly in response to a keyword search query over structured or semistructured data, designing an appropriate ranking function, and supporting top-k retrieval efficiently for the ranking functions. This special section of the *IEEE Transactions on Knowledge and Data Engineering* (*TKDE*) features a collection of four papers, selected from 16 submissions, representing recent advances in keyword search on structured data. These works present novel techniques for searching relational databases, text-rich databases, as well as XML data.

The first paper, "SPARK2: Top-*k* Keyword Query in Relational Databases" by Yi Luo, Wei Wang, Xuemin Lin, Xiaofang Zhou, Jianmin Wang, ang Keqiu Li addresses the effectiveness and efficiency challenges of keyword search on relational databases. The authors propose a new ranking method that adapts the state-of-the-art IR ranking principles for keyword search over structured data. However, in generating top-k ranked results efficiently, the nonmonotonic nature of this ranking function renders known top-k query processing techniques inapplicable. To address the challenge, the authors propose a set of efficient top-k query processing algorithms for this ranking method that minimize database probing by leveraging novel score upper bounding functions.

In the second paper, "Finding Top-*k* Answers in Keyword Search over Relational Databases Using Tuple Units," Jianhua Feng, Guoliang Li, and Jianyong Wang use indexes to record joined tuples (named as *tuple units*) in the databases. In contrast to existing work where a query result is a single tuple unit, this paper allows multiple related tuple units to be leveraged to answer a keyword query to improve search quality. To enhance the performance, the authors propose two indexes that capture relationships between different tuple units, and then develop new ranking techniques and algorithms to progressively find the top-k query results.

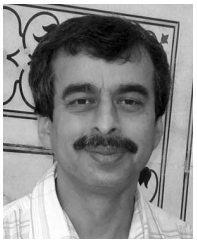
The third paper is "Efficient Keyword-Based Search for Top-K Cells in Text Cube" by Bolin Ding, Bo Zhao, Cindy Xide Lin, Jiawei Han, Chengxiang Zhai, Ashok Srivastava, and Nikunj C. Oza. It focuses on the scenario where the repository contains both structured and text data. Specifically, it studies the problem of keyword search in *text cube*, built on a multidimensional text database where each row is associated with a document and several structured dimensions. Unlike existing work where an individual document or a (joined) tuple is a query result, this work considers a cell as a query result. Given a keyword query, the goal of this paper is to find the top-k most relevant cells. The authors develop an IR-style relevance model for ranking cells, and then propose efficient algorithms to address the computational challenge due to the large number of cells in a text cube.

The final paper in this special section, "Returning Clustered Results for Keyword Search on XML Documents" by Xiping Liu, Changxuan Wan, and Lei Chen, presents a new semantics for answering keyword queries on XML data and techniques to generate clustered search results. The authors propose an efficient algorithm that clusters results on-the-fly by first generating cluster labels and then clustered results. Furthermore, they propose a technique that constructs a cluster hierarchy that is interpretable and provides a general-to-specific view of the results.

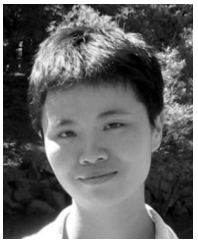
We would like to thank all of the authors who submitted papers to this special section for their high-quality contributions. We also thank the referees for their generous help and valuable suggestions. We are grateful to Professor Beng-Chin Ooi, the Editor-in-Chief of *TKDE*, for his strong support for this special section.

- *S. Chaudhuri is with Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399. E-mail: surajitc@microsoft.com.*
- *Y. Chen is with the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287-8809. E-mail: yi@asu.edu.*
- *J.X. Yu is with the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, 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.



Surajit Chaudhuri received the PhD degree from Stanford University and the BTech from the Indian Institute of Technology, Kharagpur. He is a principal researcher as well as the research area manager overseeing data management research activities at Microsoft Research, Redmond, Washington. His areas of interest include self-tuning technology for databases, query optimization, data cleaning, enterprise search, and multitenant database systems. His research projects have led to technology that have been incorporated in Microsoft SQL Server product. He is an ACM Fellow, and recipient of the ACM SIGMOD Edgar F. Codd Innovations Award and the ACM SIGMOD Contributions Award. He is currently a member of the VLDB Endowment Board and the ACM SIGMOD Advisory Board. He was the program committee chair of ACM SIGMOD 2006, and a program committee cochair of ACM SIGKDD 1999 as well as the ACM Symposium on Cloud Computing (SOCC) in 2010. He has served on the editorial boards of the *ACM Transactions on Database Systems* and the *IEEE Transactions on Knowledge and Data Engineering*.



Yi Chen received the PhD degree in computer science from the University of Pennsylvania in 2005 and the BS degree from Central South University, China, in 1999. Currently, she is an associate professor of computer science in the School of Computing, Informatics, and Decision Systems Engineering at Arizona State University. Her current research focuses on supporting keyword search on structured and semi-structured data, workflow management, social network, information integration, and information extraction. She is a recipient of a CAREER Award from the US National Science Foundation (2009), an IBM Faculty Award (2010), and a best researcher award in Computer Science and Engineering at Arizona State University (2011). She is a general cochair for SIGMOD 2012.

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 in 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. His main research interests include keyword search in databases, graph database, XML database, Web technology, and query processing and query optimization. Dr. Yu was an associate editor of the *IEEE Transactions on Knowledge and Data Engineering* (2004-2008), the information director of ACM SIGMOD (2007-2011), and is a VLDB Journal editorial board member.