Letter from the Special Issue Editor

The simple keyword-based query interface has to a great extent contributed to the wide acceptance of the Internet and its proliferation of user-contributed contents. The interface allows users to query vast collections of information freely, and hence improves the usability of the technology. Over the past two decades, database systems have made great advances in terms of performance, scalability and fault tolerance. They can now process a huge number of concurrent complex queries efficiently over enormous and diverse data sets. Naturally, the next challenge is to improve the usability of database systems in terms of providing search and query interfaces more than structured query languages can, as well as query by examples and query by forms that prevail.

Coupled with the increasing volume of text-based data, the keyword-based query mechanism becomes a natural and effective means for users to interact with databases. However, outstanding issues remain to be addressed before we can have a paradigm shift that allows users to query a database meaningfully without having any knowledge about the underlying data repository. In this issue, we have eight papers that provide different aspects and insights of keyword-based search and retrieval methods. I thank the authors for their contributions.

One of the earlier approaches to keyword-based retrieval from structured data is to identify the connection of data items containing keywords. Recent work has attempted to go beyond such approaches, and Chakrabarti, Sarawagi and Sudarshan provide a survey of recent work on adding structure to keyword search.

The large collections of user-generated content in community systems enable construction of large knowledge bases, and these are typically represented in the RDF model. Elbassuoni et al. give an overview of recent and ongoing work on ranked retrieval of RDF data with keyword-augmented structured queries.

Strings are a common data type in many applications, ranging from relational databases, semi-structured and unstructured databases to scientific databases such as genome databases. Hadjieleftheriou and Srivastava discuss the issues of indexing techniques and algorithms based on inverted indexes and a variety of weighted set-based string similarity functions.

With the recent trend in generalizing prefix-based auto-completion, Li and Li provide an overview of the information-access mechanism, where the system attempts to find answers to queries as users type in their keywords. They discuss various technical challenges in terms of interactive search speed, and open problems that remain.

The keywords entered by users may imply different information needs of different users, and this causes ambiguity during query processing. Liu and Chen discuss several post-processing methods for keyword-based retrieval on structured data with the objective of making the results more meaningful to users.

Keyword-based retrieval is well studied in the context of information retrieval. Webber examines the evolving practices and resources for effectiveness evaluation of keyword search over relational databases, compares them with longer-standing full-text evaluation methodologies in information retrieval, and offers some suggestions for future development.

Duplicates are common in databases due to abbreviation and typographical errors, and these create the nearly duplicate records problem. Yang et al. describes a system called RSEARCH for identifying nearly duplicate records so that meaningful results may be generated efficiently.

Yu, Qin and Chang present a survey on recent developments in keyword-based search over relational databases that focus on identifying primitive structures as answers and finding top-k answers efficiently. The focus is on proposals that support keyword search over RDBMS using SQL, and those viewing a relational database as a directed graph.

Beng Chin Ooi
National University of Singapore