Letter from the Special Issue Editor

The field of data management has gone a long way since the age of relational databases where data is well defined, managed by a centralized system, and accessed through methods based on well-founded semantics. In today's data management world, we are facing two major challenges: the diversity and the heterogeneity of the data, and the ad-hoc methods for accessing and manipulating such data. Even within a single organization, data could be so varied and applications so diverse that no centralized system is able to cover all needs of data management. In the public domain, challenges are often more broad and complex, with data ranging from Wikipedia to human genome and applications ranging from question answering to disease prediction.

In this issue, we survey several typical scenarios of data engineering in the post relational databases age. The data we look into include enterprise data, social network data, Wikipeida data, etc., and the applications we cover range from simply cataloging the data to building natural language interfaces for the data. While there is not going to be a one-size-fits-all solution to the many challenges that come with the heterogeneity of the data, we argue there are still several priorities to follow in the new age of data engineering. First, cataloging heterogeneous datasets to make them readily available to the user is the first requirement of data management. Second, data is characterised by richer and denser relationships they exhibit, and it is such relationships that drive many applications. Third, effective retrieval is essential for data to provide value. While we may not have a set of operators defined on well-founded semantics that can serve any need, investigating novel retrieval methods is important. Finally, data management is no longer just a system issue. Modeling and understanding the semantics of the data (particularly text data) holds the key to a new generation of intelligent applications.

We start with creating awareness for enterprise data. Enterprise data has seen such an explosive growth in recent years that most enterprise users are not necessarily aware of the availability or the semantics of their own data. Halevy et. al. discuss the situation at Google. They handle the scale and heterogeneity challenge of Google's internal data and introduce techniques to extract metadata from them, so as to facilitate the understanding and use of such data internally at Google. Microsoft's Kaushik et. al. focus on the massive amount of data amassed by web search engines. They describe services (e.g., synonymy service and web table service) created out of Bing's search data that benefit a variety of Microsoft's products including Office and Cortana.

Rich relationships inside graph-like data are the driving force of many applications, and such data posts great challenges to data management. Huang et. al. cast their attention on social networks. They survey several state-of-the-art community models based on dense subgraphs, and investigate social circles, which are a special kind of communities formed by friends in 1-hop neighborhood network for a particular user. Another important graph data is the knowledge graph. Weikum et. al. focus on advances the knowledge harvesting community has made in turning internet content, with its wealth of latent-value but noisy text and data sources, into crisp "machine knowledge" that can power intelligent applications.

New data calls for novel retrieval methods. Zhai introduces a game-theoretic formulation of the text retrieval problem to optimize user experience in search The key idea is to model text retrieval as a process of a search engine and a user playing a cooperative game, with a shared goal of satisfying the users information need (or more generally helping the user complete a task) while minimizing the users effort and the operation cost of the retrieval system. Lu et. al. propose a neural network architecture for answering natural language questions against databases. It achieves this by finding distributed representations of queries and knowledge base tables.

As we see in both Weikum et. al.'s work on knowledge harvesting and Lu et. al.'s work on neural natural language QA, the biggest challenge in data engineering is no longer just a system challenge, instead, how to model and understand the data is often the key to the success of applications. Tao et. al. study the problem of *phrase-based summarization* of a set of documents of interest. The authors introduce a phrase ranking measure to leverage the relation between subsets of documents. Atzori et. al. describe a smart system that allows people enter natural language questions and then translates them into SPARQL queries executed on DBpedia.

Haixun Wang
Facebook Inc.