Letter from the Special Issue Editor

Over the past years there has been a growing recognition of the importance of provenance in data management. Besides it traditional use for query results explanations, new applications of provenance range from query optimization, to distributed data management, human-machine interaction, and experiments reproducibility. In this issue, we have a slate of very interesting articles discussing the different roles of provenance in information management in a variety of domains.

We start with “Provenance for non-experts”, Daniel Deutch, Nave Frost and Amir Gilad. The paper considers the flourish of data-intensive systems that are geared towards direct use by non-experts, such as Natural Language question answering systems and query-by-example frameworks. It highlights the importance of incorporating provenance in building such user interfaces.

The second paper, “Provenance and the Different Flavors of Reproducibility”, Juliana Freire and Fernando Seabra Chirigati, considers the important problem of experiments reproducibility. It provides an overview of the different types of provenance and how they influence reproducibility. The goal here is to help researchers find the most appropriate provenance capture for their experiment, based on the level of reproducibility they need to attain.

The next paper “Data Citation: A New Provenance Challenge”, Abdu Alawini, Susan Davidson, Gianmaria Silvello, Val Tannen and Yinjun Wu, proposes a provenance-based novel framework of the citation of query results. The proposed solution is to specify citations for a small set of frequent queries citation views and then use these views to construct a citation for general queries.

The fourth paper, “Provenance Analysis for Missing Answers and Integrity Repairs”, Jane Xu, Waley Zhang, Abdussalam Alawini, and Val Tannen, points that prior approaches for provenance used positive provenance and were thus not directly usable for explaining missing answers or failure of integrity constraints. The paper addressee this shortcoming by offering provenance-based explanations via (minimal) repairs, applicable for debugging, repairing, and cleaning databases.

In “GProM - a Swiss Army Knife for Your Provenance Needs”, Bahareh Arab, Su Feng, Boris Glavic, Seokki Lee, Xing Niu and Qitian Zeng, provided an overview of GProM, a novel generic provenance middleware for relational databases. The system supports diverse provenance and annotation management tasks through query instrumentation.

Next, “Supporting Data Provenance in DISC Systems”, Matteo Interlandi and Tyson Condie, uses data provenance as a key building block to provide debugging support for data processing pipelines. Specifically, the paper reports experience in building Titian: a data provenance system targeting the Apache Spark framework.

Finally, we conclude with “Data Center Diagnostics with Network Provenance”, Ang Chen, Chen Chen, Lay Kuan Loh, Yang Wu, Andreas Haeberlen, Limin Jia, Boon Thau Loo and Wenchao Zhou. Diagnosing problems in data centers are a challenging problem due to their complexity and heterogeneity. The promising approach described in the paper leverages provenance, which provides the fundamental functionality that is needed for performing fault diagnosis and debugging.

I hope that you enjoy the issue as much as I enjoyed putting it together!

Tova Milo
Tel Aviv University