The Future of Data Integration

Renée J. Miller
University of Toronto
miller@cs.toronto.edu

ABSTRACT

The value of data explodes when it is integrated. In this talk, I present some important innovations in data integration over the last two decades. These include data exchange [1], which provides a foundation for reasoning about the correctness of transformed data, and the use of declarative mappings in integration [2]. I discuss how data mining has been used to facilitate data integration, including constraint discovery [3], mapping discovery [4], and in schema discovery to combat database decay and facilitate integration [5,6]. I present some important new data integration challenges that arise in data science. These include the use of mining for query and visualization recommendation over massive data lakes [7] and data set search, finding datasets of interest at interactive speeds [8].

CCS Concepts/ACM Classifiers

• Data integration; data exchange; information integration; data mining; data cleaning

Author Keywords

Data science

BIOGRAPHY

Renée J. Miller is a fellow of the Royal Society of Canada and a Professor of Computer Science. She has been named a fellow of the ACM and Bell Canada Chair of Information Systems. She received the US Presidential Early Career Award for Scientists and Engineers (PECASE), the highest honor bestowed by the United States government on outstanding scientists and engineers beginning their careers. She received an NSF CAREER Award, the Premier's Research Excellence Award, and an IBM Faculty Award. She and her co-authors received the ICDT Test-of-Time Award for their influential 2003 paper establishing the foundations of data exchange. She has served on the Board of Trustees of the VLDB Endowment and as President of the Endowment. Her research is funded by NSERC, NSF, IBM, SAP, and Bell Canada among others. She received her PhD in Computer Science from the University of Wisconsin, Madison and Bachelor's degrees in Mathematics and in Cognitive Science from MIT.

REFERENCES

[1] R. Fagin, Ph. G. Kolaitis, R. J. Miller, L. Popa. Data Exchange: Semantics and Query Answering. Theoretical Computer Science, 336(1):89-124, May 2005.

[2] R. Fagin, L. M. Haas, M. A. Hernandez, R. J. Miller, L. Popa, Y. Velegrakis. Clio: Schema Mapping Creation and Data Exchange. Conceptual Modelling: Foundations & Applications, 198-236, 2009.

[3] F. Chiang and R. J. Miller, Discovering Data Quality Rules. PVLDB 1(1):1166-1177, 2008.

[4] A. Kimmig, A. Memory, R. J. Miller, L. Getoor. A Collective Probabilistic Approach to Schema Mapping Discovery. IEEE ICDE, 921-932, 2017.

[5] R. J. Miller and P. Andritsos. Schema Discovery. IEEE Data Engineering Bulletin, 26(3):40-45, 2003.

[6] P. Andritsos, R. J. Miller, P. Tsaparas. Information-Theoretic Tools for Mining Database Structure from Large Data Sets. ACM SIGMOD, 33(2):731-742, 2004.

[7] E. Kandogan, M. Roth, P. M. Schwarz, J. Hui, I. G. Terrizzano, C. Christodoulakis, R. J. Milller. LabBook: Metadata-Driven Social Collaborative Data Analysis. IEEE Big Data, 431-440, 2015.

[8] E. Zhu, F. Nargesian, K. Q. Pu, R. J. Miller. LSH Ensemble: Internet-Scale Domain Search. PVLDB, 9(12):1185-1196 2016.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD'17, August 13-17, 2017, Halifax, NS, Canada.

© 2017 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-4887-4/17/08.
DOI: http://dx.doi.org/10.1145/3097983.3105809