Discovering Data Transformations in Web Resources (Abstract)

Ziawasch Abedjan[1] John Morcos[2] Ihab F. Ilyas[2]
Mourad Ouzzani[3] Paolo Papotti[4] Michael Stonebraker[5]

[1] TU Berlin [2] University of Waterloo [3] Qatar Computing Research Institute
[4] Arizona State University [5] MIT CSAIL
abedjan@tu-berlin.de {jmorcos,ilyas}@uwaterloo.ca
mouzzani@qf.org.qa ppapotti@asu.edu stonebraker@csail.mit.edu

Abstract. In data integration, data curation, and other data analysis tasks, users spend a considerable amount of time converting data from one representation to another. For example US dates to European dates or airport codes to city names. In practice, data scientists have to code most of the transformation tasks manually, search for the appropriate dictionaries, and involve domain experts. In a previous vision paper, we presented the initial design of `DataXFormer`, a system that uses web resources to assist in transformation discovery [1]. Specifically, `DataXFormer` discovers possible transformations from web tables and web forms and involves human feedback where appropriate. We demonstrated the system at SIGMOD 2015 and deployed an open version of the system, which helped us to increase our initial workload from 50 to 120 transformations [3]. At the same time we extended `DataXFormer` with new algorithms to find

1. transformations that entail multiple columns of input data,
2. indirect transformations that are compositions of other transformations,
3. transformations that are not functions but rather relationships, and
4. transformations from a knowledge base of public data.

We report on experiments with the collection of 120 transformation tasks, and show our enhanced system automatically covers 101 of them by using openly available web resources [2].

References

1. Z. Abedjan, J. Morcos, M. Gubanov, I. F. Ilyas, M. Stonebraker, P. Papotti, and M. Ouzzani. DataXFormer: Leveraging the web for semantic transformations. In *CIDR*, 2015.
2. Z. Abedjan, J. Morcos, I. F. Ilyas, P. Papotti, M. Ouzzani, and M. Stonebraker. DataXFormer: A robust data transformation system. In *ICDE*, 2016.
3. J. Morcos, Z. Abedjan, I. F. Ilyas, M. Stonebraker, P. Papotti, and M. Ouzzani. DataXFormer: An interactive data transformation tool. In *SIGMOD*, 2015.