Data Profiling in the Relational World

Felix Naumann

Hasso Plattner Institute
University of Potsdam, Germany
felix.naumann@hpi.de

We can be confident that most computer or data scientists have engaged in the activity of data profiling, at least by "eye-balling" spreadsheets, database tables, XML files, etc., aptly called data gazing [7]. More advanced techniques to extract metadata may have been used, such as keyword-searching in datasets, writing structured queries, or even using dedicated data profiling tools. Data profiling is the set of activities and processes to determine metadata about a given dataset. Among the simpler results are per-column statistics, such as the number of null values and distinct values in a column, its data type, or the most frequent patterns of its data values. Metadata that are more difficult to discover involve multiple columns, such as inclusion, functional and order dependencies or denial constraints [2].

With the emergence and collection of ever more structured datasets from diverse sources, as manifested for instance in data lakes, the ability to manage, understand and analyze such data is increasingly difficult but equally important: "*If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain...*" [3].

Traditional uses for metadata discovered by data profiling algorithms include data exploration, data cleansing, and data integration. For instance, a discovered (approximate) dependency can be elevated to a business rule with the aim of ridding the data of all its violations [5]. Statistics about data are commonly used for database query optimization. Yet, a significant obstacle to data profiling, especially to discover dependencies, is the inherent complexity of the problems. For instance, the number of potential key candidates, i.e., subsets of table columns that contain only unique value combinations, is exponential in the number of columns. And validating each candidate requires a scan of the entire dataset. As a consequence, a plethora of algorithms has been developed tackling the many individual data profiling problems [1].

Data profiling remains an exciting field of research, with many open challenges extending well beyond the analysis of a static, relational table. Among the open problems are efficient profiling of dynamic data, trading off efficiency and accuracy of profiling algorithms, discovery of more complex types of (semantic) constraints, and of course combining research ideas and directions from the field of relational data profiling with those geared towards data of other data models, such as graph data [4,6].

References

1. Abedjan, Z., Golab, L., Naumann, F.: Profiling relational data: a survey. VLDB Journal **24**(4), 557–581 (2015)
2. Abedjan, Z., Golab, L., Naumann, F., Papenbrock, T.: Data Profiling, Synthesis Lectures on Data Management, vol. 10. Morgan & Claypool Publishers (nov 2018)
3. Agrawal, D., et al.: Challenges and opportunities with Big Data. Tech. rep., Computing Community Consortium, `http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf` (2012)
4. Ellefi, M.B., Bellahsene, Z., Breslin, J.G., Demidova, E., Dietze, S., Szymanski, J., Todorov, K.: RDF dataset profiling - a survey of features, methods, vocabularies and applications. Semantic Web **9**(5), 677–705 (2018)
5. Ilyas, I.F., Chu, X.: Data Cleaning. Association for Computing Machinery, New York, NY, United States (2019)
6. Kruse, S., Jentzsch, A., Papenbrock, T., Kaoudi, Z., Quiané-Ruiz, J., Naumann, F.: RDFind: Scalable conditional inclusion dependency discovery in RDF datasets. In: Proceedings of the International Conference on Management of Data (SIGMOD). pp. 953–967 (2016)
7. Maydanchik, A.: Data Quality Assessement. Technics Publications, New Jersey (2007)