Data Profiling

Ziawasch Abedjan
MIT CSAIL
abedjan@csail.mit.edu

Lukasz Golab
University of Waterloo, Canada
lgolab@uwaterloo.ca

Felix Naumann
Hasso Plattner Institute, Potsdam
felix.naumann@hpi.de

***Abstract*—One of the crucial requirements before consuming datasets for any application is to understand the dataset at hand and its metadata. The process of metadata discovery is known as data profiling. Profiling activities range from ad-hoc approaches, such as eye-balling random subsets of the data or formulating aggregation queries, to systematic inference of structural information and statistics of a dataset using dedicated profiling tools. In this tutorial, we highlight the importance of data profiling as part of any data-related use-case, and discuss the area of data profiling by classifying data profiling tasks and reviewing the state-of-the-art data profiling systems and techniques. In particular, we discuss hard problems in data profiling, such as algorithms for dependency discovery and profiling algorithms for dynamic data and streams. We conclude with directions for future research in the area of data profiling. This tutorial is based on our survey on profiling relational data [1].**

I. Introduction

We can safely assume 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. More advanced techniques 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 the 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 compute involve multiple columns, such as inclusion dependencies or functional dependencies.

Traditional use cases of data profiling include data exploration, data cleansing, and data integration. Statistics about data are also useful in query optimization. Additionally, domain-specific use cases have emerged in scientific data management and big data analytics. In particular, "big data", with their high volume, high velocity, and high variety [2], are data that cannot be managed with traditional techniques. Fetching, storing, querying, and integrating big data is expensive, despite many modern technologies. Thus, data profiling gains a new importance. For instance, before exposing an infrastructure to the Twitter firehose, it might be worthwhile to find out the properties of the data one is receiving; before downloading significant parts of the linked data cloud, some prior sense of the integration effort is needed; before augmenting a warehouse with text mining results, an understanding of data quality is required. In general, many big data and related data science scenarios call for data mining and machine learning techniques to explore and mine data. Again, data profiling is an important preparatory task to determine which data to mine, how to import data into various tools, and how to interpret the results [3].

Leading researchers have noted that "*if we just have a bunch of datasets 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*[...]." [4] Data profiling addresses precisely this problem and faces three challenges:

1) Managing the input
2) Performing the computation
3) Managing the output

Apart from typical data formatting issues, the first challenge includes the problem of specifying the expected outcome, i.e., determining which profiling tasks to execute on which parts of the data. In fact, many tools require a precise specification of what to inspect. Other approaches are more open and perform a wider range of tasks, discovering all metadata automatically.

Data profiling tools and algorithms have tackled these challenges in different ways. Many rely on the capabilities of the underlying DBMS, as many profiling tasks can be expressed as SQL queries. Others have developed innovative ways to handle the individual challenges, for instance using indexing schemes, parallel processing, and reusing intermediate results. Several methods have been proposed that deliver only approximate results for various profiling tasks, for instance by profiling samples. Finally, users may be asked to narrow down the discovery process to certain columns or tables. For instance, there are tools that verify inclusion dependencies on user-suggested pairs of columns, but cannot automatically check inclusion between all pairs of columns or column sets.

This tutorial is based on our survey on data profiling [1]. We discuss data profiling use cases, present a classification of data profiling tasks (summarized in Figure 1), and focus on the second challenge from the above list (performing the computation), which has received a great deal of attention in research and practice. The computational complexity of data profiling algorithms depends on the number or rows, with a sort being a typical operation, but also on the number of columns. Many tasks need to inspect all column combinations, i.e., they are exponential in the number of columns. In addition, the scalability of data profiling methods is important, as the ever-growing data volumes demand disk-based and distributed processing.

We also shed light on the third challenge of meaningfully interpreting the results of data profiling. Obviously, any discovered metadata refer only to the given data instance and cannot be used to derive schematic/semantic properties with certainty, such as value domains, primary keys, or foreign key

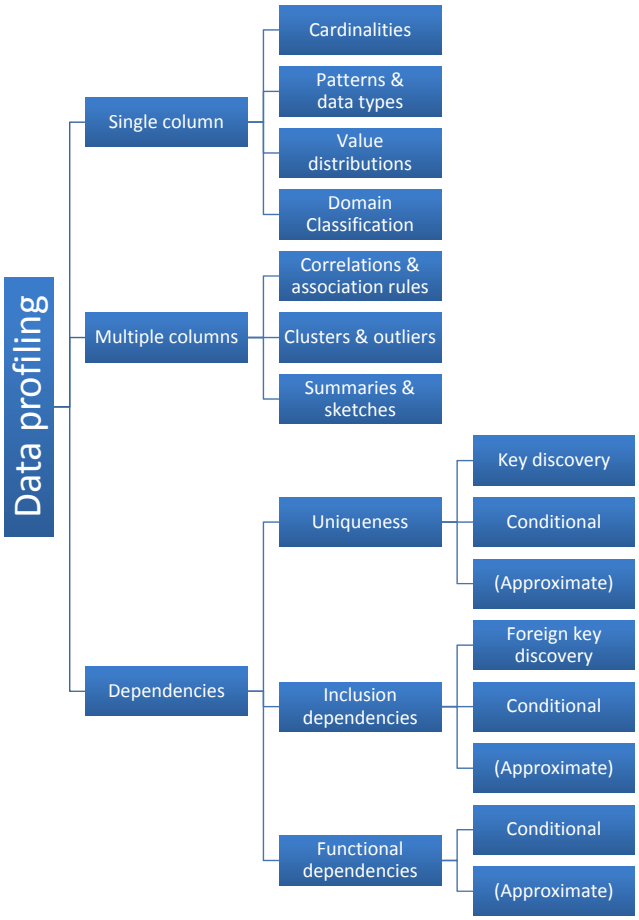

Fig. 1. A classification of traditional data profiling tasks [1].

relationships. Thus, profiling results need interpretation, which is usually performed by database and domain experts.

II. Tutorial Outline

In this section, we present the structure of this tutorial, outlining the scope and depth of each subtopic. The tutorial takes 90 minutes with 15 minutes dedicated to motivating applications and current data profiling systems, 1 hour to data profiling algorithms, and 15 minutes to open problems and future research directions.

A. Motivation and Current Profiling Systems

We motivate the problem of data profiling with real life scenarios from data integration, data exploration, and data management. In particular, we present a definition that separates data profiling as such from data mining and data exploration. Furthermore, we discuss the capability of state-of-the-art data profiling tools from industry and research [5]–[11]. Because data profiling is such an important capability for many data management tasks, there are various commercial data profiling tools. In many cases, they are a part of a data quality / data cleansing tool suite, to support the use case of profiling for frequent patterns or rules and then cleaning those records that violate them. In addition, most Extract-Transform-Load tools have some profiling capabilities.

In the research literature, data profiling tools are often embedded in data cleaning systems. For example, the Bellman [5] data quality browser supports column analysis (counting the number of rows, distinct values, and NULL values, finding the most frequently occurring values, etc.), and key detection (up to four columns). Furthermore, an interesting application of Bellman was to profile the evolution of a database using value distributions and correlations [12]: which tables change over time and in what ways (insertions, deletions, modifications), and which groups of tables tend to change in the same way. The Potters Wheel tool [6] also supports column analysis, in particular, detecting data types and syntactic structures/patterns. We discuss the pros and cons of these and other related systems, and identify easier and more challenging profiling tasks.

B. Classification of Profiling Tasks

We then categorize data profiling tasks as shown in Figure 1. We classified the tasks according to their dimensional complexity. Single column profiling refers to the analysis of values in a single column, and ranges from simple counts and aggregation functions to distribution analysis and discovery of patterns and data types. Multi-column profiling is the set of activities that can be applied to a single column but allows for the analysis of inter-value dependencies across columns, resulting in association rules, clustering and outlier detection. Finally, we describe the class of meta-data that constitute dependencies describing relationships among columns of a single table, such as keys and functional dependencies, and relationships across multiple tables, such as foreign keys and inclusion dependencies [1]. We address relevant concepts, such as partial dependencies and approximate solutions, before we discuss the challenging aspects of dependency discovery.

C. Single and Multi Column Analysis

We begin our discussion of data profiling algorithms with those for single and multi column analysis. We overview distribution and outlier analysis, data summaries, sketches and signatures, pattern discovery, characterizing missing and default values, as well as clustering and association analysis over multiple columns.

D. Dependency Discovery

We then zoom in on dependency discovery, and give an in-depth technical description of strategies that tackle the exponential complexity of dependency discovery tasks. An important concept to discuss here is the concept of "minimality" that reduces the result set of a dependency discovery task to only non-redundant dependencies. In particular, we discuss traditional apriori-based approaches [13]–[15] for pruning the search space of attribute combinations and present new algorithms [16]–[19] that significantly outperform traditional approaches through improved pruning techniques. Here we categorize existing algorithms into two major classes: row-based and column-based algorithms. Row-based algorithms process the set of candidate dependencies row by row and check which dependencies still hold. Column-based approaches generate candidate dependencies of a certain size (i.e., a certain number of columns), validate them through scanning the whole

database, and then generate a new candidate set by expanding current candidates with more columns.

We further explore the area of dependency discovery by revisiting variations of dependencies and algorithms. In particular, we address approximate algorithms that discover dependencies that may not hold on the entire dataset. Those are discovered using sampling [20] or summarization techniques [21]. The benefit of approximate solutions is that they often are significantly faster to compute than exact approaches, yet they can be used in many scenarios where inaccuracies are tolerated, e.g., in dirty datasets where glitches are expected. Furthermore, we address conditional dependency discovery, whose goal is to identify dependencies as well as conditions which specify subsets of the data where the given dependency holds [22], [23].

Finally, in addition to "classical" dependencies such as functional and inclusion dependencies, we discuss discovering other types of dependencies that are relevant to data profiling. These include denial constraints [24], differential dependencies [25], sequential dependencies [26], and temporal rules [27].

E. Open Problems and Directions for Future Research in Data Profiling

Recent trends in data management have led to new challenges but also new opportunities for data profiling. Under the *big data* umbrella, industry and research have turned their attention to data that they do not own or have not made use of yet. Much of the data that shall be used is of non-traditional type for data profiling, i.e., non-relational, non-structured (textual), and heterogeneous. Many existing profiling methods cannot adequately handle that kind of data: Either they do not scale well, or there simply are no methods yet. We discuss some of these trends and their implications toward data profiling. In particular, we address specific challenges in profiling dynamic data, i.e., data that changes and makes previously obtained meta-data obsolete or data that is consumed as a continuous stream. Furthermore, we address the peculiarities of data profiling when data resides in various non-relational formats, such as XML, RDF, or structure-less text documents.

Finally, a substantial challenge in the area of data profiling is the effective visualization of the generated meta-data. Effective visualization of meta-data requires not only effective visualization technologies but also methods and metrics for an appropriate selection of the meta-data at hand. While a data set might contain thousands of syntactically valid dependencies, it is important to identify the top-k most interesting dependencies that can be presented to the user. We hope that this tutorial encourages a stronger cooperation between the database and the visualization community.

III. Intended Audience

Data profiling touches many aspects of the very basics of data management and analytics. With its mix of use-cases and concrete methods we expect the tutorial to appeal to a large portion of the database community:

- *Researchers and data scientists* in the fields of data integration, data cleansing, data mining, data analytics find a consolidated view of various metadata discovery algorithms that either provide novel insights or help prepare datasets for subsequent tasks.
- *Students* can learn about this exciting research area, which still has many open problems and a wide field of potential use-cases.
- *Practitioners* developing and distributing any of the many products in the areas of data integration, data cleansing, data warehousing, and data analytics typically include some data profiling component that can significantly benefit from the research results presented in the tutorial.

This tutorial has not been given before by the speakers and is topically orthogonal to a previous tutorial on data exploration, which focused on data extraction and visualization [28] rather than computing statistics and metadata about data. We are not aware of any other comparable tutorials in the past.

IV. Speaker Biographies

Ziawasch Abedjan studied IT Systems Engineering at the Hasso Plattner Institute in Potsdam, Germany, where he received his bachelor's degree in 2008 and his master's degree in 2010. In 2014, he graduated as a member of the "HPI Research School on Service Oriented Computing" and joined the database group at MIT CSAIL as a postdoctoral associate. His research interests include data integration, data mining, and data cleansing.

Lukasz Golab is an Assistant Professor and Canada Research Chair at the University of Waterloo. Prior to joining Waterloo, he was a Senior Member of Research Staff at AT&T Labs. He holds a B.Sc. from the University of Toronto and a Ph.D. from the University of Waterloo. Lukasz's research interests include data stream management, data quality and data analytics. He has published over 50 articles and has given tutorials on data stream warehousing at SIGMOD 2013 and ICDE 2014.

Felix Naumann studied mathematics at the Technical University of Berlin and received his diploma in 1997. As a member of the graduate school "Distributed Information Systems" at Humboldt-University of Berlin, he finished his PhD thesis in 2000. In the following two years Felix Naumann worked at the IBM Almaden Research Center. From 2003-2006 he was an assistant professor at Humboldt-University. Since 2006 he is a full professor at the Hasso Plattner Institute heading the information systems group.

References

[1] Z. Abedjan, L. Golab, and F. Naumann, "Profiling relational data: a survey," *VLDB Journal*, vol. 24, no. 4, pp. 557–581, 2015.

[2] D. Laney, "3D data management: Controlling data volume, velocity and variety," Gartner, Tech. Rep., 2001.

[3] D. Pyle, *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

[4] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom, "Challenges and opportunities with Big Data," Computing Community Consortium, http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf, Tech. Rep., 2012.

[5] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk, "Mining database structure; or, how to build a data quality browser," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, 2002, pp. 240–251.

[6] V. Raman and J. M. Hellerstein, "Potters Wheel: An interactive data cleaning system," in *Proceedings of the International Conference on Very Large Databases (VLDB)*, 2001, pp. 381–390.

[7] L. Golab, H. Karloff, F. Korn, and D. Srivastava, "Data Auditor: Exploring data quality and semantics using pattern tableaux," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 3, no. 1-2, pp. 1641–1644, 2010.

[8] X. Chu, I. Ilyas, P. Papotti, and Y. Ye, "RuleMiner: Data quality rules discovery," in *Proceedings of the International Conference on Data Engineering (ICDE)*, 2014, pp. 1222–1225.

[9] J. M. Hellerstein, C. Ré, F. Schoppmann, D. Z. Wang, E. Fratkin, A. Gorajek, K. S. Ng, C. Welton, X. Feng, K. Li, and A. Kumar, "The MADlib analytics library or MAD skills, the SQL," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 5, no. 12, pp. 1700–1711, 2012.

[10] Z. Abedjan, T. Grütze, A. Jentzsch, and F. Naumann, "Profiling and mining RDF data with ProLOD++," in *Proceedings of the International Conference on Data Engineering (ICDE)*, 2014, pp. 1198–1201, demo.

[11] T. Papenbrock, T. Bergmann, M. Finke, J. Zwiener, and F. Naumann, "Data profiling with Metanome," in *Proceedings of the VLDB Endowment (PVLDB)*, vol. 8, no. 12, 2015.

[12] T. Dasu, T. Johnson, and A. Marathe, "Database exploration using database dynamics," *IEEE Data Engineering Bulletin*, vol. 29, no. 2, pp. 43–59, 2006.

[13] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen, "TANE: An efficient algorithm for discovering functional and approximate dependencies," *Computer Journal*, vol. 42, no. 2, pp. 100–111, 1999.

[14] Z. Abedjan and F. Naumann, "Advancing the discovery of unique column combinations," in *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, 2011, pp. 1565–1570.

[15] J. Bauckmann, U. Leser, F. Naumann, and V. Tietz, "Efficiently detecting inclusion dependencies," in *Proceedings of the International Conference on Data Engineering (ICDE)*, 2007, pp. 1448–1450.

[16] A. Heise, J.-A. Quiané-Ruiz, Z. Abedjan, A. Jentzsch, and F. Naumann, "Scalable Discovery of Unique Column Combinations," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 7, no. 4, pp. 301 – 312, 2013.

[17] Z. Abedjan, J.-A. Quiané-Ruiz, and F. Naumann, "Detecting unique column combinations on dynamic data," in *Proceedings of the International Conference on Data Engineering (ICDE)*, 2014, pp. 1036–1047.

[18] Z. Abedjan, P. Schulze, and F. Naumann, "DFD: Efficient functional dependency discovery," in *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, 2014, pp. 949–958.

[19] T. Papenbrock, S. Kruse, J.-A. Quiané-Ruiz, and F. Naumann, "Divide & conquer-based inclusion dependency discovery," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 8, no. 7, 2015.

[20] I. F. Ilyas, V. Markl, P. J. Haas, P. Brown, and A. Aboulnaga, "CORDS: Automatic discovery of correlations and soft functional dependencies," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, 2004, pp. 647–658.

[21] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, "Synopses for massive data: Samples, histograms, wavelets, sketches," *Foundations and Trends in Databases*, vol. 4, no. 13, pp. 1–294, 2011.

[22] J. Bauckmann, Z. Abedjan, H. Müller, U. Leser, and F. Naumann, "Discovering conditional inclusion dependencies," in *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, 2012, pp. 2094–2098.

[23] L. Bravo, W. Fan, and S. Ma, "Extending dependencies with conditions," in *Proceedings of the International Conference on Very Large Databases (VLDB)*, 2007, pp. 243–254.

[24] X. Chu, I. F. Ilyas, and P. Papotti, "Discovering denial constraints," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 6, no. 13, pp. 1498–1509, 2013.

[25] S. Song and L. Chen, "Differential dependencies: Reasoning and discovery," *ACM Transactions on Database Systems (TODS)*, vol. 36, no. 3, pp. 16:1–16:41, 2011.

[26] L. Golab, H. Karloff, F. Korn, A. Saha, and D. Srivastava, "Sequential dependencies," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 2, no. 1, pp. 574–585, 2009.

[27] Z. Abedjan, C. Akcora, M. Ouzzani, P. Papotti, and M. Stonebraker, "Temporal rules discovery for web data cleaning," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 9, no. 4, 2015.

[28] S. Idreos, O. Papaemmanouil, and S. Chaudhuri, "Overview of data exploration techniques," in *Proceedings of the International Conference on Management of Data (SIGMOD)*, 2015, pp. 277–281.