Data Profiling as a Process

Bridging the Gap Between Academia and Practitioners

Fabian Schomm
European Research Center for Information Systems (ERCIS)
Leonardo-Campus 3
Münster, Germany
fabian.schomm@wi.uni-muenster.de

ABSTRACT

Dealing with data necessitates understanding data, which can be facilitated by analyzing its metadata. This metadata is often not readily available, and thus, needs to be extracted and collected. This activity is called data profiling and is well established and researched in the literature. However, what is missing is a structured description of a holistic data profiling process, that puts together all the individual pieces and guides a user from the problem to the solution.

This paper describes such a process and the results and insights that have been achieved so far.

Keywords

Data Profiling, Data Exploration, Metadata Extraction

1. INTRODUCTION

When working with data, one often faces situations in which new or unknown datasets appear, which need to be processed to achieve a specific goal. Due to the unknown characteristics of the data and its schema, it is unclear how the processing should be carried out, which makes it necessary to inspect and analyze the data first. In the literature, this inspection is usually referred to as *data profiling*, which has been defined as "the set of activities and processes to determine the metadata about a given dataset" [3]. Examples of metadata about data include aggregated counts, correlations, value distributions, or functional dependencies. The extraction of these metadata has been the focus of various research activities in the past, and numerous efficient algorithms for their discovery have been developed [3].

Still, it has been our observation that a comprehensive data profiling is performed very rarely by practitioners in real-world scenarios. Rather, it seems to be common practice to inspect an unknown dataset in a manual fashion, by opening it, e.g., in a spreadsheet tool, and simply skimming through it. This ad hoc approach, which has derogatorily been called "data eyeballing" [3] or "data gazing" [8], is not only time- and cost-intensive, but also highly dependent on individual skill, prone to errors or inconsistencies, and it does not scale at all to larger datasets that contain more than a handful of values.

There is a number of reasons for the missing adoption of elaborate data profiling techniques. The most important factor is time: When deadlines have to be met, there is often little room for assessing the data carefully. Instead, starting as soon as possible and producing results immediately is perceived as more important. However, this can quickly lead to costly backtracking when certain assumptions about the data (e.g., its quality) turn out to be wrong.

Another factor for the negligence of data profiling is missing knowledge. Many people do not know enough about it or how it is done, even if they are trained professionals. Additionally, there seems to be little understanding of the potential benefits, which leads to a reluctance to learn.

This paper describes an ongoing research effort that has the goal to bridge this gap between sophisticated data profiling techniques and algorithms described in the literature on one side, and the often simplistic and crude data inspection approaches encountered in practice on the other side. In order to reach this goal, the development of a process model is proposed, which acts as a practitioner's guide to the application of data profiling and highlights its benefits. This process model should list necessary steps for data profiling, describe the involved components (i.e., algorithms, tools, people) and their interactions, and demonstrate the overall usefulness of a structured approach to data profiling. Further research is structured along a research agenda that loosely follows the suggestions of the design science approach [13]. It consists of the following steps:

- Identify and narrow down the problem by gathering use cases and examples that could potentially benefit from data profiling
- Define the objectives and features of the to-be-developed solution
- Perform a literature review to research and collect data profiling methods, techniques, and tools, as well as further work in related fields
- Design the data profiling process and complementary artifacts to guide practitioners in their data profiling efforts
- Demonstrate how the solution can be applied in real-world scenarios

28^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 24.05.2016 - 27.05.2016, Nörten-Hardenberg, Germany.

- Evaluate and validate the applicability and performance of the result

Throughout this work, the term *data object* will commonly be used. A data object is the most general abstraction of a dataset. Such an object can be any kind of data collection, set, database, flat file, or even multiples of these. This generality is needed in order to remove any preconceptions about the syntax or semantics of the data in question. In particular, a data object does not have to adhere to the relational data model, but could also be semistructured (e.g., XML, JSON) or even unstructured.

The remainder of this work is structured as follows: Section 2 lists the most important use cases in which data profiling should be one of the first steps. In Section 3, the distinction between intrinsic and extrinsic information is introduced, which is important to get a complete picture of what a data profile is about and what it entails. After that, Section 4 describes selected tools and techniques to explore the work that has already been done in this field. A first draft of the process model is shown and explained in Section 5. Finally, in Section 6 it is described how and evaluation and validation of the produced result could look like.

2. USE CASES FOR DATA PROFILING

The extraction of metadata from a data object is helpful whenever that data object needs to be processed in some way. As such, there are many different use cases in which data profiling can be applied. Abedjan et al. have provided an initial overview over use cases and the way in which data profiling tasks applies to them [3]. A concise summary of these use cases is given in this section.

2.1 Data Quality Assessment

Assessing the quality of a given data object is usually done by defining quality metrics, calculating their values and interpreting the results [12]. The definition of these metrics often involves the same results that also occur in a data profiling run, e.g., completeness or accuracy. This overlap makes it a natural fit to perform data quality assessment with the help of a data profiling tool. One example is *Profiler* [5], which utilizes visualization techniques and data mining methods to allow fast assessment of tabular data.

2.2 Data Cleansing

To clean a data object, it is necessary to first figure out where it is dirty. To do so, data profiling can help by supplying information about instances of dirty data, such as outliers, missing values, or skewed distributions. After the data has been cleaned, another profiling run can be executed to verify that the cleaning efforts were successful.

2.3 Data Integration

One core challenge in data integration is to combine multiple heterogeneous data sources into one unified mediated schema. The heterogeneity of these sources necessitates an individual handling of every source, e.g., in the form of customized ETL processes. This necessitates knowledge about the structure and content of each source. Gathering this knowledge can be sped up significantly through the usage of data profiling. Additionally, profiling multiple sources at once allows the detection of overlaps or duplicates, which can facilitate the integration task.

2.4 Data Migration

In order to migrate a data object, it is very helpful to have certain key characteristics available. For example, the physical size of the data is important for making sure that the target destination has enough free space available. Furthermore, when transferring a data object from one database to another, it should be ascertained that the target database offers support for the required schema, e.g., regarding data types and column sizes. These kinds of information can be gathered through data profiling.

2.5 Query Optimization

A query to a database usually consists of multiple operations that are ordered in a hierarchical access plan. The order of these operations heavily influences the time it takes to answer the query. With the knowledge of the size of involved data objects, a query optimizer can re-arrange the operations in such a way that the result stays the same while the execution time improves. Research on the usage of profiling techniques for query optimization goes back as far as 1988 [7]. Nowadays, an end-user never worries about the optimization of his queries, because every modern DBMS comes with finely tuned optimization techniques. These optimizations are based on data profiles, which are usually not revealed. It would be interesting to investigate whether these internal profiles could be exposed and used in different contexts.

3. INTRINSIC VS. EXTRINSIC INFORMATION

Before stepping into the development of a process, a proper definition of what exactly a data profile consists of is needed. Previous work in this area mainly addresses the quantifiable and algorithmically determinable properties of data, such as various counts, value distributions or dependencies. This point of view however does not include everything that can be learned about a data object, and hence, does not fully describe a data profile.

This leads us to the distinction between what we call *intrinsic* and *extrinsic* information. Intrinsic information is about characteristics of the data that is inherent to the raw data values themselves, such as value distributions or correlations. We call these kinds of information intrinsic, because they can be extracted and derived directly from the data, without any outside knowledge. These extraction activities are what Naumann denotes as “data profiling tasks” [9], and they have been the focus of many research activities in the past.

After gathering all intrinsic information, there is still more metadata that cannot be discovered from the data values alone, such as the data provenance, or the interpretation mode of special values (empty strings, `NULL` values). These types of metadata are what we call *extrinsic* to differentiate them from their intrinsic counterparts, and to emphasize the fact that they can only be learned from sources outside of the data values themselves.

Specifying and classifying the various kinds of extrinsic information is an ongoing research effort at our group. It is especially challenging to gather extrinsic information, because per definition, they cannot be derived from only the data. As such, it is necessary to search for other sources that are able to provide the needed information. One approach is to contact the data owner or the data creator, if they are known. It could be assumed that they possess the information in question and are willing to disclose it. Further possible sources are documentations or transformation logs about the original data object.

Intrinsic and extrinsic information complement each other and collectively form the profile of a data object. However, gathering and inspecting such a data profile are not the only approaches that can be taken by a data profiler. A more direct way is to visualize the data object in a graphical way to immediately identify its overall structure and possible anomalies. Of course, such a visualization should still assume no prior knowledge about the data object, because getting that knowledge is the goal in the first place. We call these kinds of visualization *raw*, because they address the raw data values, and not the processed and calculated result of a data analysis procedure. The major requirement for raw visualization techniques is therefore the ability to be executed with little to no configuration required. Although raw visualizations can be considered intrinsic, we intentionally choose to treat them separately, because the results and purposes are different. An example of such a technique will be described in the next section.

4. SELECTED DATA PROFILING TOOLS AND TECHNIQUES

There are many different data profiling tools, both standalone programs as well as techniques integrated into bigger software suites. Here we present two examples of such tools, one from academia and the other from the software market. A more detailed survey is given in [3]. Additionally, this section describes CityPlot as an example of a visualization technique that could prove useful in a profiling context. Further visualization techniques are studied in the area of *visual data exploration* [6], and data profiling could benefit greatly from incorporating more ideas from this field.

4.1 Metanome

Metanome is a platform that can be used to perform data profiling and automatically discover metadata [11]. It has been developed by the chair of Prof. Felix Naumann at the Hasso-Plattner-Institut in Potsdam, Germany. Metanome offers a range of state-of-the-art algorithms to perform traditional data profiling tasks and display their results. It is conceived as a modular platform that allows an easy integration of self-developed or third-party algorithms, and thus, allows comparisons and benchmarks. Metanome is available for free under the Apache license at the project website [2].

4.2 Talend Open Studio

Talend Open Studio (TOS) is a software suite that covers many data-related activities. One component is TOS Data Quality, which offers many data profiling features in a convenient user interface [1]. TOS Data Quality is a prime example of a free and easy-to-use data profiling tool that is able to cover a lot of different use cases. All components of TOS are distributed under the Apache license.

4.3 CityPlot

CityPlot is an algorithm that provides a combined view of database structure and contents by replacing data values with colored rectangles [4]. The color of a rectangle represents the data type. Thus, it is very easy for the human eye to assess a data object at a quick glance. For instance, in the example shown in Figure 1, it is immediately visible that a small number of columns contains numerical values (the red ones), while the majority contains string values (the blue ones). This can be used by a data profiler as a first indication about the contents of the data object. Additionally, CityPlot allows to assess the overall data completeness by looking at the percentage of white space, which represents empty cells, and also identify the columns or rows where the most data is missing.

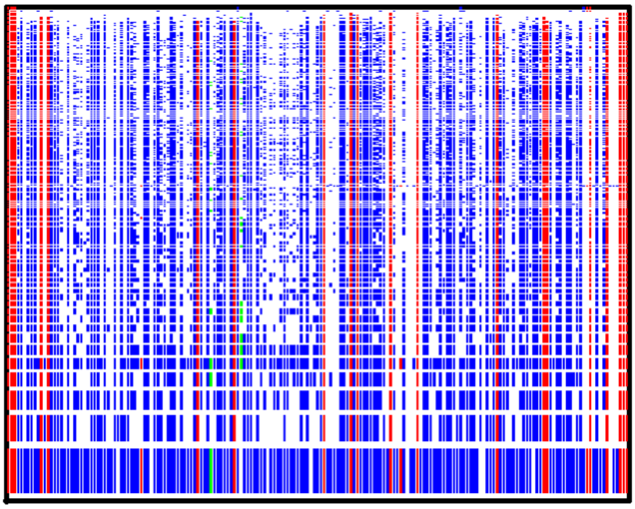

Figure 1: Raw visualization of an exemplary dataset using CityPlot. Cells in red represent numerical values, green are categorical values, and blue indicates strings.

5. PROCESS DRAFT

A first version of the process model has been developed, which can be seen in Figure 2.

The process starts with three input objects: The data object is the central piece about which more information is required. It can be of any form or shape, e.g., a spreadsheet, a database, or a flat file. It is assumed that this data object is in some way new or unknown, because otherwise (if the object were known) there would be little need to profile it in the first place. Unknown data objects are encountered frequently, for example when starting work on a new project, or adding additional data from new sources.

The second input is the task description, which is assumed to be informal and not machine-readable. It contains information about what needs to be done with the data and how an end result should look like. Task descriptions directly relate to the use case, and examples include “integrate this data object into our database” or “clean up any quality deficits”.

The last input object is initial information about the data object in question. This captures everything that is already known, like previous profiling results, schema information, or documentations. The initial information can be empty if there is no such previous knowledge. Other researchers have a rather pessimistic view on previously known information: Olson writes that “any available metadata [...] is either

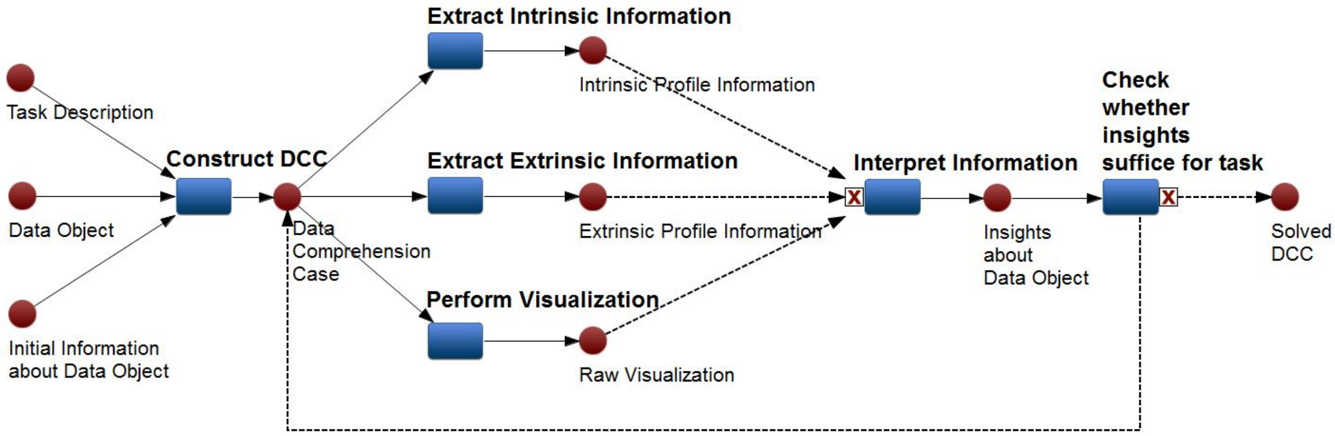
Extract Intrinsic Information
Task Description
Intrinsic Profile Information
Check whether insights suffice for task
Construct DCC
Extract Extrinsic Information
Interpret Information
Data Object
Data Comprehension Case
Extrinsic Profile Information
Insights about Data Object
Solved DCC
Perform Visualization
Initial Information about Data Object
Raw Visualization

Figure 2: Preliminary process modell to solve data comprehension cases

wrong or incomplete" [10, p. 122]. However, he further argues that available metadata is still useful, as it can provide a basic structure and a good starting point for a profiling activity.

All input objects are used to construct a so called *Data Comprehension Case* (DCC). Thus, a DCC is characterized by the need to comprehend a given data object in order to solve a given task. The DCC is the focus of the process, and solving it is it's ultimate goal. Based on the distinction introduced in Section 3, a DCC can be approached from three main directions:

Intrinsic information can be extracted from the data object using traditional data profiling tools and algorithms, and ranges from simple counts (e.g., nulls, duplicates) over statistical variables (e.g., histograms, patterns) to multi-column constructs (e.g., correlations, dependencies).

Extrinsic information is metadata that cannot be derived from the data object itself. How exactly these kinds of metadata can be gathered is a current research question at our group. First ideas to tackle this challenge are surveys or interview as a structured approach for knowledge extraction from relevant people, i.e., the data owner or the data creator.

Raw visualizations allow a user to directly assess a data object in a graphical way. The key difference that sets this approach apart from data eyeballing is the fact that visualizations leverage colors, shapes, and the human cognitive ability to process images faster than text, to directly reveal the inner structure of the data object.

Each of these steps can be performed individually. After the information has been gathered, it needs to be interpreted and transformed into insights by a human, which is not an easy task for untrained users. The process should provide assistance for this step, e.g., in the form of an interpretation guide. Such a guide should explain the meaning and definition of the various metadata, how they are related, and what value ranges are usually expected. The reason this guide would be useful is that we have experienced that people who are new to data profiling get lost easily in the overflow of information that a profiling tool provides. A little guidance and pointers on what to look out for can go a long way here.

In the last step of the process, it is checked whether the gained insights are sufficient for the execution of the task. This is the case when most uncertainties have been removed and the user feels confident enough. In case that check fails, further information can be extracted by looping back to the beginning of the DCC. This loop can be repeated as often as necessary in order to gather all necessary information, so that the DCC is solved, the data object is comprehended by the user sufficiently, and the initially set task can be executed.

The absence of an explicit failure state as an end point is intentional. It is assumed, that any DCC can be solved if it is explored thoroughly enough. Should this assumption turn out to be inappropriate, the process needs to be adapted. Another key feature is that it is not required to gather every bit of information possible before proceeding. Instead, the process ends as soon as the minimum amount of required information has been extracted. This ensures that no unnecessary work is done, and that the profiling process is efficient and result-oriented.

6. EVALUATION AND VALIDATION

In order to evaluate the process model and verify its usefulness, it needs to be tested. An experiment will be set up that tests the process model in the following way: First, a concrete use case is needed that would benefit from the application of data profiling. For example, the implementation of an ETL process for the integration of two unknown datasets could be used. This task requires that a mediated schema is designed into which the sources can be integrated. Second, test subjects are needed that execute the task. These test subjects should have basic skills required for the completion of the task. For example, Master students from our Information Systems programme could be recruited. The test subjects are then divided into two groups. One of the groups is provided with the data profiling process and corresponding tools, while the other group acts as a control group and receives no additional material. Both groups get the task description and the data. It is then measured how long it takes everybody to complete the task, and how good the individual results are. If the data profiling process is good, the first group should perform much better than the control group.

Depending on the number of test subjects available, this experiment could be even more diversified. For example, the size of the to-be-integrated datasets could vary from only a few rows to thousands of rows. It would then be possible to hypothesize, that an extensive data profiling approach only shows its benefits if the dataset exceeds a certain size, while small datasets can be processed just fine with manual ad hoc methods.

After the experiment, a short survey should be issued that interrogates the participants about the perceived complexity of the task and whether or not data profiling was helpful in solving it. The gathered feedback could then be used to further refine and improve upon the model.

7. SUMMARY

This paper described our current state of research regarding the application of data profiling as described in academia to real-world use cases of practitioners. The approach mainly consists of the development of a process model which guides a user through the various steps of data profiling and how they should be applied. This process model is complemented by an interpretation guide that facilitates the handling of results by the end user.

Additionally, we established the notion of extrinsic information as a complement to classical data profiling tasks, i.e., intrinsic information. This leads to a more complete picture of what the profile of a data object is and how it can be created.

There is of course much work that still needs to be done as indicated by the research agenda in Section 1. The process model has been described from a very broad perspective and needs to be refined to a more granular level. Many details about how individual steps should be executed are still missing and need to be filled in.

8. REFERENCES

[1] *Data Profiling - Talend*, 2016. URL: `https://www.talend.com/resource/data-profiling.html` [cited 31 March 2016].

[2] *Metanome - Data Profiling - Hasso-Plattner-Institut*, 2016. URL: `http://hpi.de/naumann/projects/data-profiling-and-analytics/metanome-data-profiling.html` [cited 31 March 2016].

[3] Z. Abedjan, L. Golab, and F. Naumann. Profiling relational data: a survey. *The VLDB Journal - The International Journal on Very Large Data Bases*, 24(4):557–581, 2015.

[4] M. Dugas and G. Vossen. CityPlot: Colored ER Diagrams to Visualize Structure and Contents of Databases. *Datenbank-Spektrum*, 12(3):215–218, 2012.

[5] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer. Profiler: integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 547–554. ACM, 2012.

[6] D. A. Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, 2001.

[7] M. V. Mannino, P. Chu, and T. Sager. Statistical Profile Estimation in Database Systems. *ACM Comput. Surv.*, 20(3):191–221, Sept. 1988. URL: `http://doi.acm.org/10.1145/62061.62063`, `doi:10.1145/62061.62063`.

[8] A. Maydanchik. *Data quality assessment.* Technics publications, 2007.

[9] F. Naumann. Data Profiling Revisited. *SIGMOD Rec.*, 42(4):40–49, Feb. 2014. URL: `http://doi.acm.org/10.1145/2590989.2590995`, `doi:10.1145/2590989.2590995`.

[10] J. E. Olson. *Data quality: the accuracy dimension.* Morgan Kaufmann, 2003.

[11] T. Papenbrock, T. Bergmann, M. Finke, J. Zwiener, and F. Naumann. Data Profiling with Metanome. *Proc. VLDB Endow.*, 8(12):1860–1863, Aug. 2015. URL: `http://dx.doi.org/10.14778/2824032.2824086`, `doi:10.14778/2824032.2824086`.

[12] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[13] R. H. von Alan, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.