A Visual Data Profiling Tool for Data Preparation

Bjørn Marius Von Zernichow and Dumitru Roman
SINTEF Digital
Oslo, Norway
BjornMarius.vonZernichow@sintef.no
Dumitru.Roman@sintef.no

***Abstract*—In this paper, we propose a tool that implements visual data profiling capabilities for data preparation – an essential step in the process of linked data generation. Our tool features visual data profiling – a technique that identifies and visualizes potential data quality issues, relevant data cleaning functions, and an interactive spreadsheet table view. The proposed demonstration of the tool will focus on the use of visual data profiling in a scenario of cleaning and transforming tabular weather data – as a pre-processing step for linked data generation.**

Keywords-data preparation; visual data profiling; linked data; usability testing; interactive data cleaning and transformation.

I. INTRODUCTION

Tabular data has been an important type of source for generating linked data, but very often tabular data has quality issues that create challenges in generating linked data [1]. Examples of data quality issues include occurrences of missing values, outliers, inconsistencies, and noisy data [2]. Despite considerable recent research in the area of data quality, there are still opportunities for innovative solutions that can improve data quality and make cleaning and transformation processes more efficient [3][4]. Moreover, there is a need for better solutions and tools to assist users in the data preparation phase that usually comes before the linked data generation process. In this context, visual data profiling is a technique that can support the data preparation process. Visual data profiling systems are used to assess the data quality of datasets, and identify sources of quality issues such as missing and extreme values [2].

We propose a tool that implements visual data profiling capabilities in data preparation by taking as a baseline the Grafterizer framework [1][4][5], a framework for data cleaning and linked data generation – part of the DataGraft platform [4][6][7]. The profiling system checks a selection of a dataset (e.g., a column of values) against a rules matrix to display only possible and relevant charts and data cleaning functions. As an example, number based columns will enable functions (e.g., 'Replace empty cells with median value') and charts (e.g., boxplot) that are not allowed for string columns.

Furthermore, we performed a usability study with 24 users to evaluate usefulness and ease of use of the prototype, while areas of improvement were identified by four expert reviewers. A data preparation scenario was used to compare usefulness and ease of use of the existing version of Grafterizer with the proposed tool that implements visual data profiling capabilities. Drawbacks of the existing Grafterizer framework include usability issues such as a steep learning curve and a complex, less intuitive user interface. To address Grafterizer's usability challenges, we implemented a proof of concept tool that features visual data profiling capabilities to ease the process of data preparation, and improve data quality.

The remainder of this paper is organized as follows. Related work is discussed in Section II, while the implementation of the software prototype is presented in Section III. The demonstration of the tool is outlined in Section IV. Finally, Section V summarizes this paper and outlines avenues for future work.

II. RELATED WORK

Currently, there exists no framework for tabular data cleaning, transformation and linked data generation that targets both data developers and non-developers [1]. Profiler [8], Data Wrangler [11], Trifacta [12] and Talend [13] are examples of systems for data quality analysis that include data mining and anomaly detection techniques in addition to visualizations of relevant data summaries that can be used to evaluate data quality issues. Our data preparation tool is inspired by Profiler, Trifacta, Talend, and the implementation of a spreadsheet table for direct manipulation of data [14][15]. Still, none of these tools for visual data profiling in data preparation target linked data generation. The above-mentioned tools lack specific capabilities that are needed in a linked data transformation process, e.g., the annotation of data with URIs (Uniform Resource Identifiers), and mapping of data into a linked data format that conforms to a specific ontology and data model. Our proposed tool features visual data profiling capabilities that are aimed to be easily integrated in a linked data generation pipeline by replacing the existing version of Grafterizer.

 ISBN: 978-1-61208-603-3

III. SOFTWARE PROTOTYPE

Based on the drawbacks that were identified in the existing version of Grafterizer, the improved tool should provide a) Visual data profiling capabilities, b) Recommendations for relevant data cleaning and transformation functionality, c) A pipeline that reflects the applied data preparation steps, and d) A solution that is useful in data scientists' work activities, and easy to use.

The visual data profiling approach was implemented in a software prototype featuring 14 data cleaning and transformation functions. The implemented data cleaning and transformation process involves the following activities [8]–[10]:

1. **Discovery**: Based on the visual data profiling charts and statistical feedback, the user explores the content and structure of the dataset to understand the quality of the data.
2. **Data preparation:** The user applies relevant data preparation functions to the dataset to clean and transform the data.
3. **Validation**: The visual data profiling charts are used to validate that the data is cleaned and transformed according to the intended quality and structure.

The tool consists of five interacting components:

1. The file import (Fig. 1-1): Parsing of the tabular dataset (i.e., in CSV (Comma Separated Values) format).
2. The tabular view (Fig. 1-2): The data can be manipulated directly in the tabular view which features spreadsheet functionality such as 'copy/paste' or 'insert column'.
3. The visual data profiling view (Fig. 1-3): When the user clicks a column in the table, the visual data profiling view returns relevant information about missing values and data distribution of the values in that column.
4. The suggested data cleaning functions (Fig. 1-4): Based on the assessment of data quality, the user selects one of the suggested, relevant data cleaning and transformation functions to improve data quality.
5. The steps pipeline (Fig. 1-5): Finally, the applied data preparation steps are added to a steps pipeline that reflects the data cleaning history.

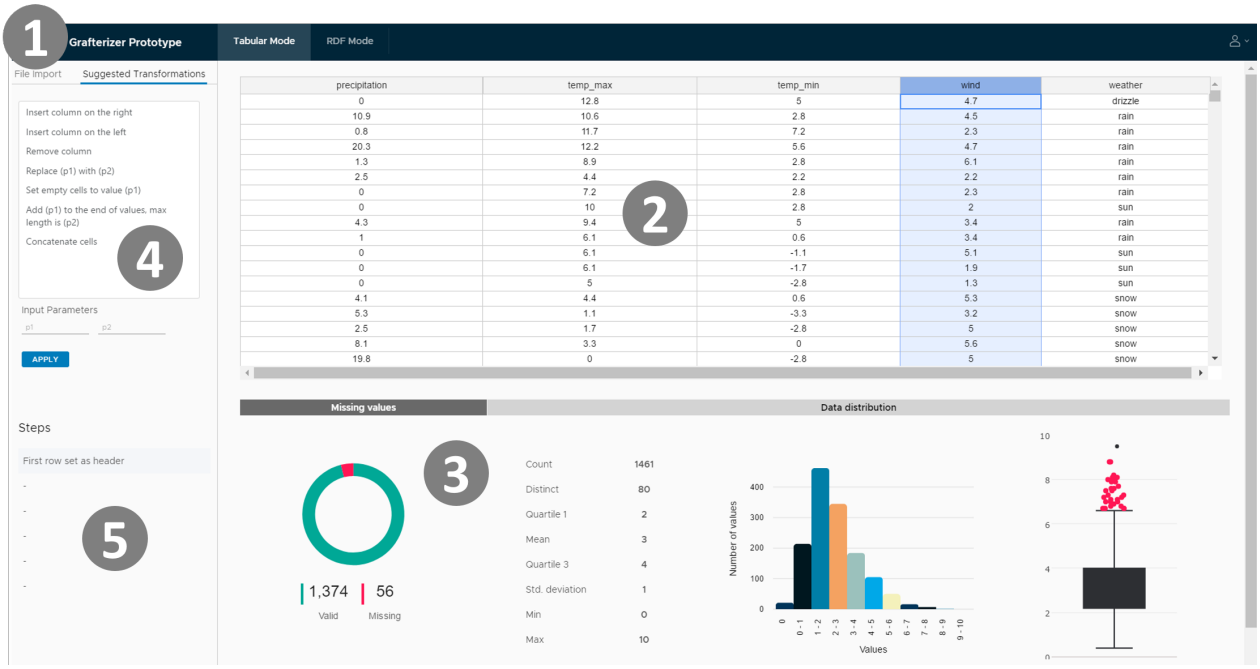

Figure 1. User interface of the visual data profiling tool

IV. DEMONSTRATION OUTLINE

The visual data profiling prototype will be demonstrated in a scenario that cleans and transforms tabular weather data [16] (precipitation, minimum and maximum temperatures, wind speed, and weather condition). The iterative cleaning and transformation process assisted by the visual data profiling system will include identifying and correcting missing values, and applying transformation functions to the dataset to prepare it for linked data mapping. The demonstration will include all three activities of the visual data profiling process described in Section III, and a typical scenario will include the following steps:

 ISBN: 978-1-61208-603-3

1. The weather dataset is imported [16].
2. The user selects one of the columns, e.g., 'wind', directly in the table view.
3. The data quality, i.e., missing values and data distribution in this case, is assessed by the visual data profiling system. The user reads from the leftmost chart that there are 56 missing values in the 'wind' column.
4. Based on the information about missing values, the user selects a relevant data cleaning function (i.e., 'Replace missing values with a defined value'). In this context, the missing values will be replaced by the median value of all values in the 'wind' column. Alternatively, the function can be selected by right clicking the table view.
5. The rightmost boxplot chart is used to find the median value (i.e., 15.6), and the user applies the data cleaning function to replace all missing values of the column.
6. The user will once more use the profiling charts to assess the current number of missing values. The function has been successfully applied, and the leftmost chart updates to reflect zero missing values.
7. Steps 2 – 6 are repeated to continue improving the quality of the dataset.
8. The resulting dataset is imported to DataGraft and transformed to linked data.

The open source code of the visual data profiling tool is currently available at GitHub [17].

V. CONCLUSION AND FUTURE WORK

This paper proposed a visual data profiling tool that implements visual data profiling capabilities and statistical feedback, recommendations for data cleaning operations, and an interactive spreadsheet table view. The proposed capabilities can improve data quality, and reduce time spent on cleaning and transforming data. Furthermore, the visual data profiling tool has been evaluated in terms of usability, and found to be perceived useful and easy to use [5].

Future work will focus on developing a framework that simplifies the technical user specification in a domain specific language. This will be achieved by implementing a visual recommender system for data profiling, and a semi-automated data preparation approach to guide the user through an incremental process of cleaning and transforming data.

ACKNOWLEDGEMENTS

The work in this paper is partly supported by the EC funded projects proDataMarket (Grant number: 644497), euBusinessGraph (Grant number: 732003), and EW-Shopp (Grant number: 732590).

REFERENCES

[1] D. Sukhobok et al., “Tabular Data Cleaning and Linked Data Generation with Grafterizer,” ESWC (Satellite Events), pp. 134–139, 2016.

[2] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[3] J. M. Hellerstein, “Quantitative Data Cleaning for Large Databases,” United Nations Economic Commission for Europe (UNECE), Feb. 2008.

[4] D. Roman et al., “DataGraft: One-Stop-Shop for Open Data Management,” To appear in the Semantic Web Journal (SWJ) – Interoperability, Usability, Applicability (published and printed by IOS Press, ISSN: 1570-0844, DOI: 10.3233/SW-170263), 2017.

[5] B. M. V. Zernichow and D. Roman, “Usability of Visual Data Profiling in Data Cleaning and Transformation,” To appear in the proceedings of ODBASE 2017 - The 16th International Conference on Ontologies, DataBases, and Applications of Semantics, Springer, 24-25 October 2017, Rhodes, Greece, in press.

[6] D. Roman et al., “Datagraft: Simplifying open data publishing,” in ESWC (Satellite Events), pp. 101–106, 2016.

[7] D. Roman et al., “DataGraft: A Platform for Open Data Publishing,” In the Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. (LIME/SemDev@ESWC), 2016.

[8] 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, New York, NY, USA, pp. 547–554, 2012.

[9] J. Heer, J. M. Hellerstein, and S. Kandel, “Predictive Interaction for Data Transformation.,” in CIDR, 2015.

[10] S. Chen, “Six Core Data Wrangling Activities eBook,” Trifacta, 2015.

[11] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, “Wrangler: Interactive visual specification of data transformation scripts,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3363–3372, 2011.

[12] M. Heinsman, “Data Wrangling | Prepare Raw & Diverse Data Faster,” Trifacta. [Online]. Available: https://www.trifacta.com/. [Accessed: 2017.09.28].

[13] “Talend Data Preparation: Self-Service Data Prep for Analytics,” Talend Real-Time Open Source Data Integration Software.

[14] E. Bakke and D. R. Karger, “Expressive query construction through direct manipulation of nested relational results,” in Proceedings of the 2016 International Conference on Management of Data, pp. 1377–1392, 2016.

[15] “Microsoft Excel 2016 Spreadsheet Software.” [Online]. Available: https://products.office.com/en/excel. [Accessed: 2017.09.28].

[16] GitHub - vega-datasets: Common repository for datasets used by Vega-related projects. Vega, 2017.

[17] GitHub - data-fixer: Tool for tabular data cleaning, preparation and transformation. DataGraft, 2017.

 ISBN: 978-1-61208-603-3