Eurographics Conference on Visualization (EuroVis) (2015)
R. Maciejewski and F. Marton (Editors)

QualityFlow: Provenance Generation from Data Quality

C. Bors[1], T. Gschwandtner[1], and S. Miksch[1]

[1]Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria

Abstract

If properly recorded, provenance information of a data set reveals the story about its origins, which hands it passed through, and how the data has been modified. Moreover, each data operation may have considerable impact on the quality of the data set. This information is of great value for anyone who needs to decide if the quality of a data set is sufficient for further processing. However, current approaches in data quality assessment feature only a limited amount of provenance information. We present the interactive QualityFlow visualization that provides the history of operations on a data set and their influence on respective data quality metrics to support sense-making. QualityFlow allows users to explore this information and share their insights with collaborators.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces And Presentation]: User Interfaces—User-centered design H.5.3 [Information Interfaces And Presentation]: Group and Organization Interfaces—Computer-supported cooperative work

1. Introduction

A prerequisite of any kind of data analysis is to make sure that the data's quality is sufficient to produce meaningful results. Handling Data Quality (DQ) is an iterative task that requires user expertise and domain knowledge of the given data [KPHH11, KPP*12]. Due to context-specific domain characteristics, analysts develop comprehensive knowledge of quality problems inherent to the data domain that can be useful to colleagues. Current DQ assessment environments widely disregard collaboration; analysts can merely retrace other colleagues' actions by examining transformation operations, given that this information is actually available.

Collaboration services are incorporated into Information Visualization applications to perceive the rationale behind users' sense-making during the analytic process [LWPL11]. Understanding the process of how the data was collected and transformed is a key component of understanding inherent and implied information. We present a conceptual design that:

(1) Exploits information from data pre-processing steps and from DQ metrics to provide provenance information.
(2) Combines these types of information to support the user in investigating the history and the quality of the data.
(3) Supports collaboration by providing a visual documentation of the transformation history and annotation means.

2. Related Work

Data provenance information is utilized to resolve conflicting data and to determine reliability based on lineage [BKWC01, SPG05]. Provenance generation is focused on generating implicit knowledge from actions and creating historical data. Attfield et al. [AHW10] suggest using visualization prototypes to support analysts hypothesize on the provenance data. In DQ management an approach to measuring DQ is computing DQ metrics on specific data characteristics [Red12], aiming to find structural or measurement errors by means of computation. Kandel et al. [KHP*11] argue that interactive and visual systems could facilitate DQ assessment as well as data verification. As a common way of propagating insights to collaborators or analysts, annotations are used to give additional information [LLC08]. Only little research directive [AHW10] has been put into adapting the concepts of generating provenance to support the sense-making process in DQ assessment. To this end, we propose a visualization approach to augment DQ assessment with provenance and collaboration support.

3. *QualityFlow* - Visualization

Our design of provenance representation is centered around interdependencies between transformation operations and DQ metrics. Such operations change properties of the data, and thus, may change the quality of the data. A combined

representation is required that allows investigating (1) the history of transformation operations, (2) DQ and its change over time, and the effect of transformation operations on different aspects of DQ. Regarding accessibility, we aim at providing a visualization that is easily accessible to data analysis experts from various domains. Thus, we focus on incorporating familiar design elements into our conceptual design, such as bar charts, a structure resembling tables, and node-link diagrams.

3.1. Provenance from Transformation Operations and Quality Metrics

In general, analysts looks for information on how data has been created and modified, and how characteristics and properties are affected by this process. Different data transformation operations, regardless of complexity, have different impacts on the data and require a mechanism to make such consequences explicit. DQ metrics serve as such an indicator on how the operations affect the data. In the context of our approach, a DQ metric is a *normalized* statistical measure that *indicates an established quality aspect* of the data.

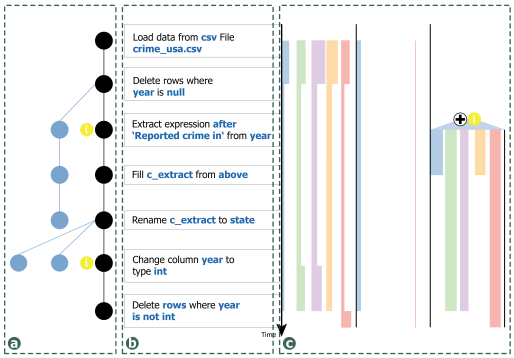

Figure 1: *Interactive visual exploration environment: (a) provenance graph of user actions, (b) the transformation log, (c) DQ metrics view resembling a tabular data structure.*

3.2. Design Rationales

We based the conceptual design of *QualityFlow* on Miksch and Aigner's [MA14] *design triangle*, mapping analysts as users, provenance information as data, and DQ assessment as task. We combine a tabular view with a DQ metrics view representing metrics for each data table column, and a provenance graph view mapping each action to a node. Each metric is indicated by a colored vertical line for each column. These lines are vertically aligned with the transformation actions given in the transformation log view. This visually links both provenance information types and helps to assess the impact of each transformation step on the different aspects of the data (indicated by a change in the width of the line of the respective metric). Although time is usually represented on a horizontal axis from left to right, we decided for a vertical time axis from top to bottom in accordance with conventional representations of transformation logs, where recent transformations are appended at the bottom.

Another consideration in our design regards the actual representation of DQ metrics. The user is not interested in data that already features high quality, but rather the lack of quality. Hence, we emphasize on this lack of quality and visualize the inverted quality measures, i.e., the amount of present quality problems. Based on their impact on the table structure, we distinguish between regular and structural transformation operations. To convey structural column changes in our visualization, we introduce a horizontal transition element that indicates the structural consequences on the table. Besides the visual representation, we further support traceability and collaboration by providing means for annotations: both, user-authored and automatically generated annotations (more information about the visualization and interaction can be found in the supplementary material).

4. Discussion and Future Work

Our conceptual visualization design directs the user's attention to comprehensive information about the transformation operations executed on a data set rather than providing information in raw data form. Conspicuous actions can be determined more efficiently, while regular activities can be quickly skimmed through to confirm validity. Making sense of collaborators' actions is a non-trivial task that depends on particular circumstances, like the data domain, the users' experience, and best practices that can be facilitated by annotations. Supplementary information can be conveyed from one user to the other through annotations along with the respective data. Heer and Agrawala [HA07] summarized design considerations for collaboration in visual applications, some of which are integrated in our approach.

DQ metrics are very specific in their application domain and can be highly useful but also misleading. Future work will involve providing functionality to customize and adapt metrics for better mapping data domain characteristics and resolving ambiguous interpretations. Growing data—with a high number of columns or rows—pose a problem to this type of visualization that needs to be addressed in the ongoing development process by either condensing DQ metrics of multiple columns or contextually optimizing metrics display, as well as optimizing the generation of metrics. While profiling and cleansing data, exploration may lead to dead ends from where backtracking is required. We propose logging users' activities during DQ assessment to obtain information about their learning processes.

Acknowledgments This work is part of the the Laura Bassi Centre of Expertise. CVAST is funded by the Austrian Federal Ministry of Economy, Family and Youth (project number: 822746).

References

[AHW10] ATTFIELD S. J., HARA S. K., WONG B. L. W.: Sensemaking in Visual Analytics: Processes and Challenges. In *2010: International Symposium on Visual Analytics Science and Technology* (Bordeaux, France, 2010), Kohlhammer J., Keim D., (Eds.), Eurographics Association, pp. 1–6. URL: http://diglib.eg.org/EG/DL/PE/EuroVAST/EuroVAST10/001-006.pdf, doi:10.2312/PE/EuroVAST/EuroVAST10/001-006. 1

[BKWC01] BUNEMAN P., KHANNA S., WANG-CHIEW T.: Why and Where: A Characterization of Data Provenance. In *Database Theory - ICDT 2001*, Van den Bussche J., Vianu V., (Eds.), vol. 1973 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2001, pp. 316–330. URL: http://dx.doi.org/10.1007/3-540-44503-X_20. 1

[HA07] HEER J., AGRAWALA M.: Design Considerations for Collaborative Visual Analytics. In *Symposium on Visual Analytics Science and Technology, 2007. VAST 2007* (Oct. 2007), pp. 171–178. doi:10.1109/VAST.2007.4389011. 2

[KHP*11] KANDEL S., HEER J., PLAISANT C., KENNEDY J., VAN HAM F., RICHE N. H., WEAVER C., LEE B., BRODBECK D., BUONO P.: Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization Journal 10*, 4 (2011), 271–288. URL: http://ivi.sagepub.com/content/10/4/271.short. 1

[KPHH11] KANDEL S., PAEPCKE A., HELLERSTEIN J., HEER J.: Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2011), '11, ACM, pp. 3363–3372. URL: http://doi.acm.org/10.1145/1978942.1979444, doi:10.1145/1978942.1979444. 1

[KPP*12] KANDEL S., PARIKH R., PAEPCKE A., HELLERSTEIN J. M., HEER J.: 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, 2012), '12, ACM, pp. 547–554. URL: http://doi.acm.org/10.1145/2254556.2254659, doi:10.1145/2254556.2254659. 1

[LLC08] LI Q., LABRINIDIS A., CHRYSANTHIS P.: User-Centric Annotation Management for Biological Data. In *Provenance and Annotation of Data and Processes*, Freire J., Koop D., Moreau L., (Eds.), vol. 5272 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2008, pp. 54–61. URL: http://dx.doi.org/10.1007/978-3-540-89965-5_7. 1

[LWPL11] LU J., WEN Z., PAN S., LAI J.: Analytic Trails: Supporting Provenance, Collaboration, and Reuse for Visual Data Analysis by Business Users. In *Proc. of the 13th IFIP TC 13 Int. Conf. on HCI - Vol. IV* (Berlin, Heidelberg, 2011), '11, pp. 256–273. URL: http://dl.acm.org/citation.cfm?id=2042283.2042311. 1

[MA14] MIKSCH S., AIGNER W.: A Matter of Time: Applying a Data-Users-Tasks Design Triangle to Visual Analytics of Time-Oriented Data. *Computers & Graphics, Special Section on Visual Analytics 38* (2014), 286–290. URL: http://www.ifs.tuwien.ac.at/silvia/pub/publications/miksch_cag_design-triangle-2014.pdf, doi:10.1016/j.cag.2013.11.002. 2

[Red12] REDMAN T. C.: Data Quality Management Past, Present, and Future: Towards a Management System for Data. In *Handbook of Data Quality*, Sadiq S., (Ed.). Springer Berlin Heidelberg, 2012, pp. 15–40. URL: http://link.springer.com/chapter/10.1007/978-3-642-36257-6_2. 1

[SPG05] SIMMHAN Y. L., PLALE B., GANNON D.: A Survey of Data Provenance in e-Science. *SIGMOD Rec. 34*, 3 (Sept. 2005), 31–36. URL: http://doi.acm.org/10.1145/1084805.1084812, doi:10.1145/1084805.1084812. 1