Eurographics Conference on Visualization (EuroVis), Posters Track (2016)
T. Isenberg and F. Sadlo (Guest Editors)

TaCo: Comparative Visualization of Large Tabular Data

R. Hourieh[1], H. Stitz[1], N. Gehlenborg[2], and M. Streit[1]

[1]Johannes Kepler University Linz, Austria
[2]Harvard Medical School, United States of America

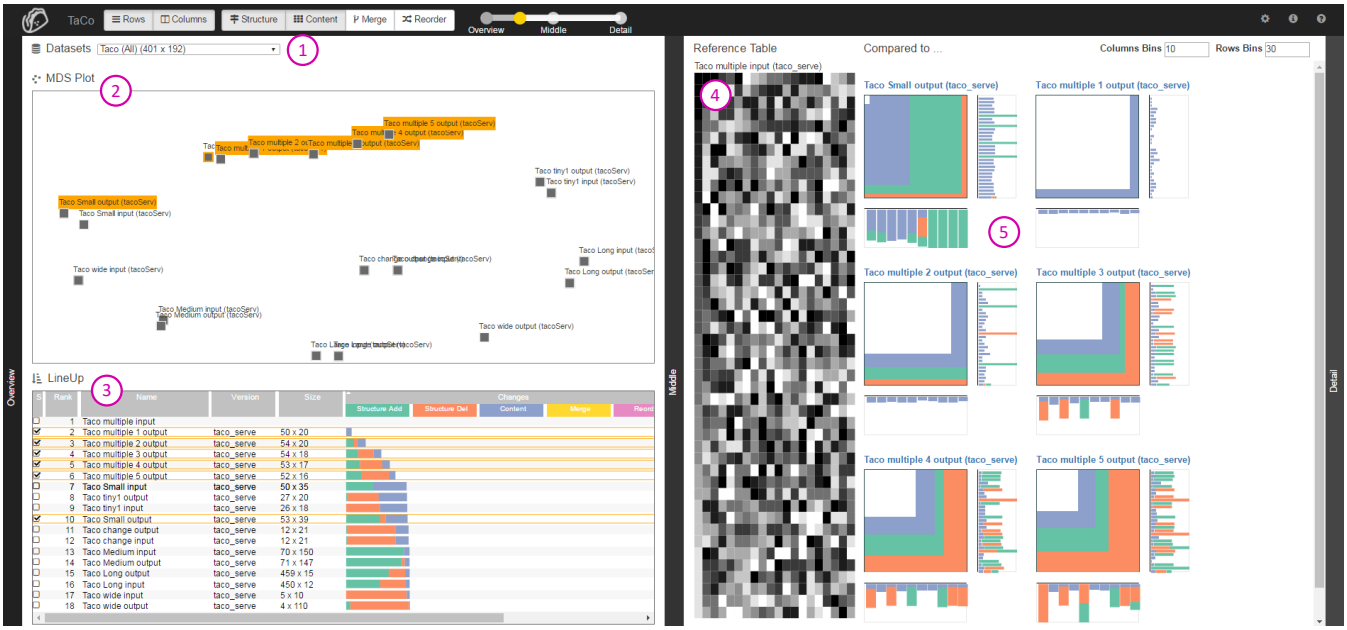

Figure 1: *The multi-view interface of TaCo showing multiple instances of an artificially generated table. The overview on the left side (1) lets the user select a collection of tables that are plotted using (2) Multidimensional Scaling (MDS) based on the calculated similarity among the tables. (3) The user compares one selected reference table to all other tables in LineUp. The middle view on the right side shows (4) the reference table as a heatmap and (5) the aggregated differences to a selected group of tables for both row and column changes.*

Abstract

Tabular data plays a vital role in many different domains. In the course of a project, changes to the structure and content of tables can result in multiple instances of a table. A challenging task when working with such derived tables is to understand what exactly has changed from one version to another. Traditional comparison tools assist users in inspecting differences between multiple table instances, however, the resulting visualizations are often hard to interpret or do not scale to large tables with thousands of rows and columns. To address these challenges, we developed TaCo, an interactive comparison tool that effectively visualizes the differences between multiple tables at various levels of granularity: (1) the aggregated differences between all table instances, (2) the differences between one table compared to all others, and (3) the detailed differences between two instances.

1. Introduction

Understanding tabular data is an essential task in many domains, such as accounting, biology, and computer science. An important task when making sense of such data is to investigate the difference between multiple instances of a table, for example, to detect modifications in monthly payroll tables or to observe differences in multiple biological experiments. In this work we present *TaCo* (***Ta**ble **Co**mparison*), a visual comparison tool that calculates the

DOI: 10.2312/eurp.20161149

difference between tabular data and provides a novel interactive visualization to encode the difference. The current version of *TaCo* compares only homogeneous tables comprising the same data type and semantic in all columns (or rows).

Together with biomedical data analysts we elicited a series of tasks to be supported by an effective comparative visualization:

T I: Identify the type of changes as one of the four types: *structural changes* for added or removed rows/columns, *content changes* for modifications in a cell value, *reordering changes* for repositioned rows/columns, and *merge changes* for combining multiple rows/columns together to yield only one row/column.

T II: Compare two or multiple tables at various levels of granularity: (a) compare multiple tables to each other ($N : N$), (b) compare a reference table to all other tables ($1 : N$), and (c) compare two tables ($1 : 1$).

T III: Compare tables with regard to their dimensions to achieve *row-wise*, *column-wise*, and *cell-wise* table comparison.

2. Related Work

Comparing large tabular data requires a two-part solution: (1) calculating the difference between tables and (2) visualizing the difference in an effective and scalable way. Several table comparison tools exist (e.g., DiffKit [Pan16] and Daff [Fit16]), but most of them are limited to particular file formats (e.g., ExcelCompare [San16]) or relational database tables (e.g., AQT [Car16]). These tools usually generate textual representation of the difference with basic color encoding which does not scale to large tables with thousands of rows or columns. Furthermore, existing tabular comparative visualizations are usually task dependent, for example, for networks analysis [ABHR*13, ZLD*15] or database query comparison [EST08]. Other visualizations lack the ability to perform simultaneous row-wise, column-wise, and cell-wise comparison of tables [LSP*10, BDF*14].

3. *TaCo* Visualization Approach

As a prerequisite for visualizing the difference between multiple table instances, we first calculate the pair-wise difference between all instances.

TaCo follows a four-stage visualization approach (see Figure 2) that allows users to reduce the number of compared tables from stage to stage, while increasing the details shown.

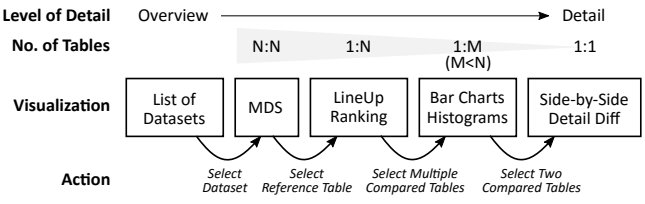

Figure 2: *Four-stage comparative visualization approach.*

The user starts by comparing N table instances in a *Multidimensional Scaling* (MDS) plot that positions instances with high similarity close to each other. After selecting a reference table, the

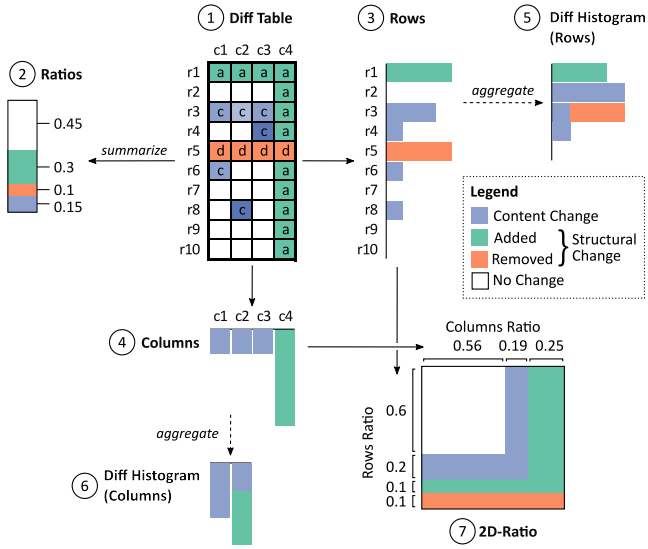

Figure 3: *The difference between two tables is visualized as (1) a difference table. Changes are summarized on a per cell basis and visualized as a (2) ratios bar plot. The difference table can be aggregated for (3) row and (4) column directions separately. (5,6) Further aggregation for one direction is shown as a histogram. Summarizing changes for rows and columns results as a (7) 2D-ratio visualization.*

differences to all others are visualized as a ranked table using the LineUp [GLG*13] technique (see Figure 1(3)). In the ranking visualization, each stacked bar corresponds to one table and each bar segment represents one type of change. By changing the width of the change type columns, the user can adjust the weights of the change types. The user can then again select multiple tables (rows) in the ranking visualization for which *TaCo* will visualize the aggregated difference, as illustrated in Figure 3. *TaCo* shows a one dimensional histogram that encodes the different changes for each dimension of the table (rows and columns) together with a 2D ratio plot (see Figure 3(7)) that summarizes the changes in both directions as ratios (see Figure 1(5)). The last and most detailed part of the interface provides a one-to-one comparison between two selected tables by presenting full heatmaps for both instances together with a union difference heatmap, which encodes the four possible types of change identified in Section 1. All interaction stages are shown in further detail using large tables in the accompanying video.

4. Future Work

We found *TaCo* to be effective for comparing multiple instances of homogeneous data tables. This motivates us to extend our approach to heterogeneous tables, to better integrate the temporal relation among different table versions, and to support additional change types, such as replacement and rename operations.

5. Acknowledgements

This work was supported by the Austrian Science Fund (FWF P27975-NBL) and the State of Upper Austria (FFG 851460).

References

[ABHR*13] ALPER B., BACH B., HENRY RICHE N., ISENBERG T., FEKETE J.-D.: Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2013), CHI '13, ACM, pp. 483–492. 2

[BDF*14] BEHRISCH M., DAVEY J., FISCHER F., THONNARD O., SCHRECK T., KEIM D., KOHLHAMMER J.: Visual Analysis of Sets of Heterogeneous Matrices Using Projection-Based Distance Functions and Semantic Zoom. In *Computer Graphics Forum* (2014), vol. 33, Wiley Online Library, pp. 411–420. 2

[Car16] CARDETT ASSOCIATES LTD.: AQT Data Compare. http://querytool.com/tourdcomp.html, 2016. Accessed: 2016-04-07. 2

[EST08] ELMQVIST N., STASKO J., TSIGAS P.: DataMeadow: a visual canvas for analysis of large-scale multivariate data. *Information Visualization 7*, 1 (2008), 18–33. 2

[Fit16] FITZPATRICK P.: daff. http://paulfitz.github.io/daff/, 2016. Accessed: 2016-04-07. 2

[GLG*13] GRATZL S., LEX A., GEHLENBORG N., PFISTER H., STREIT M.: LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13) 19*, 12 (2013), 2277–2286. 2

[LSP*10] LEX A., STREIT M., PARTL C., KASHOFER K., SCHMALSTIEG D.: Comparative Analysis of Multidimensional, Quantitative Data. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '10) 16*, 6 (2010), 1027–1035. 2

[Pan16] PANICO J.: DiffKit. http://www.diffkit.org, 2016. Accessed: 2016-04-07. 2

[San16] SANCHAY: ExcelCompare. https://github.com/na-ka-na/ExcelCompare, 2016. Accessed: 2016-04-07. 2

[ZLD*15] ZHAO J., LIU Z., DONTCHEVA M., HERTZMANN A., WILSON A.: MatrixWave: Visual Comparison of Event Sequence Data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 259–268. 2