User Study – DQProv Explorer

Supplemental Material I – Capturing and Visualizing Provenance from Data Wrangling

Christian Bors, Theresia Gschwandtner, Silvia Miksch

September 9, 2019

1 Evaluation Structure

1.1 Introduction

- 5-10 Minutes of introduction into the field of Data Wrangling
- Introduction into OpenRefine transformations and filters
- Introduction into the Explorer prototype
  - Overview of quality metrics used within the prototype (completeness, validity, numeric plausibility)
  - Used encodings
  - General functions of the three different components
    - Quality Flow View (QF)
    - Issue Distribution View (ID)
    - Provenance Graph View (PG)
    - Comparison View/Mode (CV)
  - Interactions available in the components (I)
- Introduction into the used data set, with short overview of particular data columns.

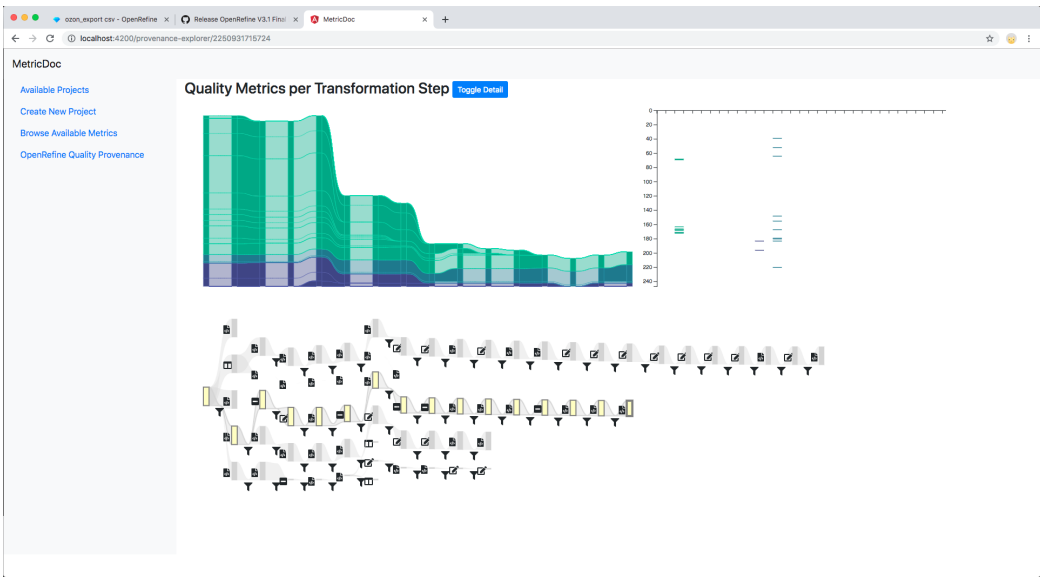

Figure 1: Overview of the *DQProv Explorer*. In the Provenance Graph View, branch 2 of 4 major branches is selected for analysis of quality development over time (in the quality flow view in the top left, and detailed analysis of remaining issues in the Issue Distribution View (top right).

1.2 Dataset

The dataset used during the user study is a slightly modified version (purposefully removed single cells) of the *TopGear* dataset obtained from the *R*-package *robustHD* [1] [1]. The data quality metrics employed alongside the dataset were deliberately chosen to be simplistic, so that participants need not to require further information on specific quality checks and validation schemata. The dataset exhibited quality issues in terms of validity (invalid data types), completeness (missing values), and plausibility (implausibly high/low values).

1.3 Tasks

The following describes the tasks given to the study participants, to see if the design of the prototype allowed conducting the tasks of confirming quality changes, discriminating between different changes in quality, validating if

[1] `https://www.rdocumentation.org/packages/robustHD/versions/0.5.1`

a dataset is usable in its current state, and understanding the sequence of transformations conducted by a different user.

$\mathbf{T}_{\text{act}}$ & $\mathbf{T}_{\text{pres}}$ - Look at the first state of the dataset and identify the column with the most issues (Column *'weight'*). Now look at the end node of one transformation branch and determine how quality evolved for this column. You can see multiple transformation branches: How different are the two branch end nodes in terms of quality, do similar issues remain? Can you find out what transformation/operation impacted the quality of this column the most?

$\mathbf{T}_{\text{meta}}$ - If only the dataset of the second branch was available for analysis, what columns would you use for analysis. If you look at the three different branches and compare remaining quality issues, which one would you choose for analysis, and for what type of analysis?

$\mathbf{T}_{\text{rec}}$ & $\mathbf{T}_{\text{rep}}$ - How did a sequence of actions influence the data? Going back to the Weight column, which of the branches would you use for analysis?

$\mathbf{T}_{\text{coll}}$ & $\mathbf{T}_{\text{meta}}$ - Can you determine the user's objective in the sequence of transformations shown in the branch at the bottom of the provenance graph?

1.4 Participant Expertise

Participants were asked about profession, and self-assessment of experiences in the fields of:

Data Wrangling, Data Profiling, or Data Cleansing

If Experienced, on what type of data

Information Visualization, Visual Analytics

2 Summarized Results

We have summarized the feedback from participants by views and interactions of *DQProv Explorer*. That way the usefulness of each part of the system could be assessed on its own, and how the interactions combined them. We also note how many participants used which view for which task.

2.1 Quality Flow View

This view was received well, 5 of 6 participants noted the usefulness for assessing the development of quality. The view was used in all tasks by all participants.
Two users initially had issues mapping the stacked bars to columns.

2.2 Provenance Graph View

The Provenance Graph View was used in all tasks by all users. But in terms of usability, the node size was deemed as not useful by 3 of 6 participants. 3 participants could not understand the encoding of data flow along the edges, hence they wondered why edges were seemingly bundled or disappeared (when the data was filtered to only very few values, the edge would become very thin). 1 participant noted to add a filter function to highlight nodes that affected particular columns, and adding a delta function to more effectively find changes in row size of the data. Participants explored the graph in different ways: while 2 participants iteratively navigated each node of the graph, the remaining participants were only interested in the end nodes (*"I would like to have all end nodes highlighted"*).

2.3 Issue Distribution View

4 of 6 Participants questioned the usefulness of the Issue Distribution View, using them as part to solve tasks $\mathbf{T}_{\text{meta}}$ (2 of 6), $\mathbf{T}_{\text{coll}}$ (2 of 6), which *"takes up whitespace"*, and *"I haven't used the detail view, and for the current selection it does not even give me useful information"*. We attribute these critical comments to the application scenario employed in our study design, and the assigned tasks not being specifically tailored to assessing error distribution in the dataset. 3 of 6 Participants criticized that the change of content in the difference view when switching to comparison mode is unclear and must be signaled accordingly.

2.4 Comparison Mode

The comparison mode was appreciated to compare branches, participants used them in tasks $\mathbf{T}_{\text{rec}}$& $\mathbf{T}_{\text{rep}}$ (6 of 6) and $\mathbf{T}_{\text{coll}}$& $\mathbf{T}_{\text{meta}}$ (3 of 6).
3 participants noted that the mirroring initially posed confusion. Even

though the participants were explicitly instructed about the mirroring, 2 participants still mixed up the branches during detailed inspection. It was noted to signal the mirroring more clearly (the colored nodes were not sufficiently indicative), and one participant suggested to mirror the Provenance Flow View vertically to compare the selected data revisions.

2.5 Interactions

Other critical feedback could be traced back to limited interaction possibilities, and we determined that some approaches pursued by participants during task execution would have required a more extensive set of interactions, such as metric selection to brush nodes affecting the metric in the Provenance Graph View, provenance graph node filtering, or highlighting techniques.

Evaluation – Participant 1

Participant

- Gender: male.
- Profession: MA Student.
- Expertise:
  - Data Wrangling, Data Profiling, or Data Cleansing: Yes, No, Yes. Advanced. Tools: LoD Refine[2] (OpenRefine[3]).
  - Data: Multiple data source harmonization.
  - Information Visualization: Entry level, data analysis plots and statistics plots.

Performance on $\mathbf{T_{act}}$ & $\mathbf{T_{pres}}$

- Could find the column with maximum error in the quality flow visualization (QF). But only assessed validity metric as maximum, even though also a second metric (completeness) signaled issues in this column.
- By selecting the last node of the top branch (PG), and observing the flow

[2] `https://sourceforge.net/projects/lodrefine/`
[3] `http://openrefine.org/`

of the metric developing over time (QF), he could find that one operation reduced quality.

Performance on $\mathbf{T}_{\mathbf{meta}}$

- Comparing the two top branches (PG) lead the participant to the conclusion that the top branch yielded more valid data, with the lower branch removing entries unnecessarily (I).

Performance on $\mathbf{T}_{\mathbf{rec}}$ & $\mathbf{T}_{\mathbf{rep}}$

- Looking at the second branch (CV, PG, I), quality was improved, but he found that this corresponds to changes of other problems as well (QF) (this is due to rows being removed, affecting the ratio of errors across all metrics.
- He would not use the dataset due to these transformations affecting all rows (CV, PG, I) (*[rows are being deleted]*).

Performance on $\mathbf{T}_{\mathbf{coll}}$ & $\mathbf{T}_{\mathbf{meta}}$

- The participant tried to focus on quality (QF) and try to comprehend what happened when multiple entries were edited but could not due to [*self-assessed*] missing info (I) ([*info is available on mouseover, but is limited that edit action was performed, but detailed information is missing*]).

Critical Feedback

- The participant could not see what impact an action had on the data (I), due to the edges (PG) not being clearly recognizable to him. The visual encoding of edge width corresponding to filtering rows of the dataset was not understood.
- More highlighting (I) was demanded, e.g., highlighting columns, searching for nodes that were changed in the provenance graph.
- The QF when comparing two branches should be scrollable (I).
- Raw data should be comparable on demand in the DV.
- differentiation in the CV between the two paths is unclear, a different linking should be employed to show the differences in metrics between the two end-nodes.

Positive Feedback

- The visualization of quality across data transformations (QF) was marked as very useful, especially when the scale of the dataset is larger.

Evaluation – Participant 2

Participant

- Gender: male.
- Profession: PhD Student.
- Expertise:
  - Data Wrangling, Data Profiling, or Data Cleansing: Yes, Yes, Yes. Expert, > 1 year. Tools: self developed tools.
  - Data: Text data, retrieval data.
  - Information Visualization: Expert, > 1 year.

Performance on $\mathbf{T_{act}}$ & $\mathbf{T_{pres}}$

- Could successfully determine the column *Weight* (QF), but immediately noted that height is sub-optimal encoding for lack of quality – he would rather prefer height maps to high quality.
- He noted the necessity for mouse-over trial-and-error (I) for finding the column with the highest number of issues.
- He did not understand alignment of nodes (PG) and QF bars at first, so searched for nodes that affected column *Weight* individually (PG, I), and noted that it could be beneficial to highlight nodes that affect column *Weight* on demand (I), to have insight how this column changed across all branches.

Performance on $\mathbf{T_{meta}}$

- Participant wants to see the changes rather than the overall quality development (PG). - Noted that the diff view (ID) is not very helpful in detecting the differences between two views. Alternatively the overall number of rows and quality issues could give better way of determining a difference. Also a

link into the data could help.
- By highlighting what nodes were affecting certain selected columns, exploration would be more easy, and to guide users towards relevant branches.

Performance on $\mathbf{T_{rec}}$ & $\mathbf{T_{rep}}$

- He used the provenance graph nodes to determine how many rows remain in the dataset for the top two branches of the provenance graph, determined that the second contained less data and that overall quality was not significantly lower, so preferred branch number 1 (from top).
- Within this process he noted that he would like to see all branches' endpoints highlighted. - Also it was noted that the diff did not help enough, because both the overall number of entries *and* the quality are key measures for high quality in the dataset.
- Validation would require inspecting the data – wants a link back into the data state.

Performance on $\mathbf{T_{coll}}$ & $\mathbf{T_{meta}}$

- Understood the transformation operations, in which a subset of the data was selected to conduct cleaning only on that data. But noted that the filter indication is not very expressive without the possibility to observe the content of the column.
- Single cell operations do not tell any information what happened – needs to be addressed to trace actions. - If overall error is increased, information without signaling the number of rows is rather ambiguous and needs to be determine in a separate step.
- The participant understood that the decision what transformation path to choose depends on the subsequent analysis, based on if high accuracy of the available data is favoured, or if more entries with imputed values are beneficial (e.g., for model building) [*it should be added that the used dataset provides data without significant outliers, but certain entries are incomplete*]

Observations

The user mainly utilized the quality flow visualization for determining changes in the data, and only used the provenance graph mouseover information if necessary.

Critical Feedback

- In the provenance graph, the filter analogy can be overlooked easily if filtering only yields a small number of entries. A delta of changes, or numeric values for total size and number of changes in the data (for each state of the dataset) would signal changes more effectively. - Mirroring the second quality flow visualization should be signalized.

Positive Feedback

- The prototype provides the ability to conduct collaborative cleaning by allowing users to see the branches that are created by different approaches.
- The quality flow visualization is very effective for signaling the overall quality.
- Linking of the prototype components is really smooth and helps with exploration.

Evaluation – Participant 3

Participant

- Gender: male.
- Profession: PhD Student.
- Expertise:
  - Data Wrangling, Data Profiling, or Data Cleansing: Yes, Yes, Yes. Entry to advanced level, < 1 year. Tools: scripting.
  - Data: Databases: relational/structured data.
  - Information Visualization: Expert, > 1 year.

Performance on $\mathbf{T_{act}}$ & $\mathbf{T_{pres}}$

- Easily found weight column by using mouseover (QFV), could determine the operation responsible for the change. However the filter operation was not clearly understood at first.

Performance on T_{meta}

- Participant is iteratively navigating nodes in comparison view, and trying to understand alternate path column changes.
- Could determine that second path solved quality issues at the same steps, but in a different way. But determined that the second branch is more beneficial for solving problems.
- Not clear that dark colored paths signal a change in metric (suggestion to use a different texture).

Performance on T_{rec} & T_{rep}

- The participant noticed that branch two reduced the data size while branch one retained the data and concluded that selecting between those branches came down to preference.
- First branch more complete, second branch removed data.
- To decide the user wants to know more information on what changed in the first branch to reconstruct (missing information what was changed in the single cell operations).
- If the dataset is unknown, and operations are provided in detail: The user preferred retaining information, under the assumption of knowing that issues were solved. This requires trust in the dataset and the user conducting the wrangling process

Performance on T_{coll} & T_{meta}

- Participant could find out that cars running fossil fuels were removed and found that quality degraded.
- He concluded correctly by comparing two branches that the ratios of problems increased due to the removal of more correct data, and retaining dirty ones.

Observations

The user used iterative selection of the PG nodes to exactly retrace changes done to the dataset. Hence, the participant could understand the used wrangling workflow quite effectively. Subsequently, the participant saw the need

for improving graph interactions, like a focus+context technique, or grouping operations. In contrast, the Issue Distribution View was not used at all. For him it was difficult to distinguish changes of quality in the quality flow visualization, color coding transitions like states added to this problem, could be addressed by employing a different coloring schema.

Critical Feedback

- differentiation in the comparison view between the two paths is unclear, a different linking should be employed to show the differences in metrics between the two end-nodes.
- Information on filters should be more intuitive and clear (range indicators, and condensing information)
- Connection to the dataset should allow more detailed analysis.
- Issue Distribution View adds to much white space, and does not resolve the question where the issues are, apart from position, but this is irrelevant for retrospective analysis.

Positive Feedback

- Adding signals to highlight nodes that have already been explored.
- The prototype allows for finding leaks and modifications more easily, if done in the tool.
- Use of icons for operations.
- Collaborative efforts can be explored.

Evaluation – Participant 4

Participant

- Gender: male.
- Profession: Post-doctoral researcher.
- Expertise:
  - Data Wrangling, Data Profiling, or Data Cleansing: Yes, Yes, No. Expert, < 1 year. Tools: Alteryx[4].

[4] `www.alteryx.com`

– Data: Text data, retrieval data.
– Information Visualization: Expert, > 1 year.

Performance on $\mathbf{T}_{\mathbf{act}}$ & $\mathbf{T}_{\mathbf{pres}}$

- Participant could easily find column *Weight*. But noticed that it disappeared, by attempting to click the transformations.

Performance on $\mathbf{T}_{\mathbf{meta}}$

- Participant valued branches with less operations to accomplish similar quality, but still determined branch one to be the best quality dataset, he also interpreted the completeness metric as the most worrying.

Performance on $\mathbf{T}_{\mathbf{rec}}$ & $\mathbf{T}_{\mathbf{rep}}$

- Inspection of individual changes in quality.
- Mostly focusing on filtering icons and mouse-over information, rather than filters and operations
- Participant preferred dropping columns (what's the least amount of columns to conduct an analysis on the entire dataset?)
- Trust in imputed values is only accepted if knowledge about who conducted the operations is available, otherwise dropping these entries is preferred.

Performance on $\mathbf{T}_{\mathbf{coll}}$ & $\mathbf{T}_{\mathbf{meta}}$

- Participant noticed worse amount of errors based on the row removal

Critical Feedback

- Participant suggested the ability to filter for changes in specific columns, to find transformations more quickly (T1)
- Single cell operations require more information.
- Operations icons should be encoded by a glyph.
- Quality flow should also encode information about number of rows/entries in the dataset.

Positive Feedback

- Quality flow was appreciated, but the participant suggested vertical mirroring instead of horizontal.
- Using a different set of metrics for determining a dataset's appropriateness for machine learning training.
- Usefulness is tied to the objective quality functions – the more expressive they are, the better the analysis can be.

Evaluation – Participant 5

Participant

- Gender: female.
- Profession: PhD student.
- Expertise:
  - Data Wrangling, Data Profiling, or Data Cleansing: Yes, Yes, No. Beginner.
  - Data: Scientific data, spatial data
  - Information Visualization: Expert, > 3 years.

Performance on $\mathbf{T_{act}}$ & $\mathbf{T_{pres}}$

- Participant wanted to use click interaction (QF, I) to find column *Weight*. But after all found out to use on demand mouseover information to find the column.
- Participant struggled to find context information to determine the corresponding transformation, alignment could not help adequately.

Performance on $\mathbf{T_{meta}}$

- The participant did valued lower quality over availability of the data. Upon asking the metrics were seen as trustworthy by the user.

Performance on T_{rec} & T_{rep}

- Upon inspection (CV) the participant expressed that the columns are rather unclear to her. It did not help her to comprehend what happened in the data (the miles per gallon column exhibited excessive amounts of implausible values), but she did not associate the change with the transformation (PG, QF).

Performance on T_{coll} & T_{meta}

- Participant could not associate the changes in quality to the nodes/edges in the provenance graph (PG).
- She did not find out for what purpose the sequence of actions/operations was executed, hence a distinction between the branches was not achieved by her (CV).
- The inspection only led to single insights, that certain actions caused a decrease in quality issues.

Observations

The participant did not try to explore all different modes of interaction, and hence also did not leverage them to determine the source of changes in quality or compare the differences between two selected quality flows.

Critical Feedback

- Difficult to see where data are filtered, suggested to use different encoding, only show filter information on demand.
- The number of data in a data state is not clearly visible, and rarely comparable.
- Data to ink ratio low.
- Change metric representation to resemble columns – vertical scaling of the visualization.

Positive Feedback

- Quality encoding makes sense intuitively.
- Comparison view works well.

Evaluation – Participant 6

Participant

- Gender: female.
- Profession: PhD student.
- Expertise:
  - Data Wrangling, Data Profiling, or Data Cleansing: Yes, Yes, Yes. Advanced, > 1 year. Tools: Excel.
  - Data: tabular data, relational data.
  - Information Visualization: Expert, > 3 years.

Performance on $\mathbf{T}_{\mathbf{act}}$ & $\mathbf{T}_{\mathbf{pres}}$

- Participant could easily find column *Weight*. Attempted to click the metric paths.
- She could find minimal changes in quality for the selected branch and could determine the operation responsible for the change, by iterating through all operations until she found the change in the column.

Performance on $\mathbf{T}_{\mathbf{meta}}$

- For comparison the participant disabled the detail view.
- Noted that differences were not significant (for the selected branches).
- Using the comparison mode, she decided for the branch with higher quality/lower amount of quality issues.
- Participant wanted to use node toggling to determine the changes in quality between two changes.

Performance on $\mathbf{T}_{\mathbf{rec}}$ & $\mathbf{T}_{\mathbf{rep}}$

- Would prefer the dataset with higher quality, when confronted with the branch that removed rows, she preferred the other, valuing data size as well.
- Did not trust the dataset enough to decide on a branch, without ability to look into the raw data.

Performance on T_{coll} & T_{meta}

- Participant noticed that operations had different implications, which came down to the observation that she used an iterative approach towards understanding the wrangling/cleansing process.
- Could distinguish and understand differences in operation types and their impact on quality.

Critical Feedback

- Demanded more interaction and linking abilities, in particular quality flow to provenance graph.
- Path highlighting was not sufficient to link the branches to the flow views.
- Legend missing.
- Graph structure changes during exploration, makes navigating harder.
- Demanded column labeling.

Positive Feedback

- Liked use of whitespace
- Participant stressed the importance of the Issue Distribution View, and would only provide a toggle to remove "empty" columns.
- Liked the use of color that make the elements distinguishable.

References

[1] Andreas Alfons. *Robust Methods for High-Dimensional Data*, 2016-01-08. TopGear Dataset. Robust methods for high-dimensional data, in particular linear model selection techniques based on least angle regression and sparse regression. (v0.5.1).