Arachnid: Generalized Visual Data Cleaning

Conder L. Shou, Amita Shukla
Columbia University
New York, NY
{cs3544,ajs2349}@columbia.edu

ACM Reference Format:

Conder L. Shou, Amita Shukla. 2019. Arachnid: Generalized Visual Data Cleaning. In *2019 International Conference on Management of Data (SIGMOD '19)*, June 30–July 5, 2019, Amsterdam, Netherlands. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3299869.3300108>

1 INTRODUCTION

Data cleaning is an inherently exploratory and visual process. Visualizations help analysts spot errors or surprising patterns/relationships that would otherwise go unnoticed. Once identified, the user will want to execute transformations in an attempt to correct these errors. However, alternating between the contexts of transforming and visualizing the data can be tedious. As a result, there is a need for data to be visualized and cleaned *on the fly*. Current solutions address this issue by allowing users to specify data cleaning transformations through a limited set of interactions that can *directly manipulate* visualizations pre-defined by the cleaning system itself. These visualizations are either generated as a table (OpenRefine [1], Wrangler [6], Microsoft Excel [4]), or as bar graphs (Tableau Prep [5]). By mapping a narrow set of mouse-based interactions to data cleaning specifications, these systems allow users to intuitively and quickly transform data within a constrained set of use cases. However, in order to gain a comprehensive visual understanding of the data and to quickly clean it for additional use cases, analysts often need to generate multiple visualizations from a set of common types and interactively execute data transformations beyond those supported by current cleaning software. Thus, we propose Arachnid as a novel system that builds upon existing work in generalized selection[2] and direct manipulation to introduce a model for translating mouse-based interactions on common types of user-defined visualizations into an enhanced set of data cleaning transformations.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMOD '19, June 30–July 5, 2019, Amsterdam, Netherlands

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5643-5/19/06.

<https://doi.org/10.1145/3299869.3300108>

2 MOTIVATING USE CASE

We begin with a motivating use case. A journalist is exploring how the demographics of passengers on luxury cruises has changed over time and wonders if the Titanic incident correlates with any changes. She finds a dataset of Titanic passengers taken from Kaggle[3] and notices some passengers have missing ages. The journalist wants to estimate the missing values so that she can properly visualize and present the information to her readers. We chose this task as an example because imputing null values is a common data cleaning operation for developers interested in machine learning classification and is a more advanced task for curious non-technical users. A portion of the dataset is shown:

ss	Name	Sex	Age	Sib
1	Crosby, Mrs. Edward Gifford (Catherine Elizabeth Halstead)	female	64	
1	Straus, Mr. Isidor	male	67	
1	Cavendish, Mrs. Tyrell William (Julia Florence Siegel)	female	76	
3	Ilieff, Mr. Ylio	male		
1	Fliegerheim, Mrs. Alfred (Antoinette)	female		
3	Serman, Mr. Ellis	male		
3	Johnston, Mrs. Andrew G (Elizabeth Lily" Watson)"	female		
2	Laroche, Miss. Louise	female	1	
2	Wells, Master. Ralph Lester	male	2	

Figure 1: Example records of the Titanic dataset from Kaggle.

Generally, there are two methods for imputing null values. Users can pick some statistical trait of the current column, i.e. the mean, mode, or median of Age, or attempt to discover a closer approximate value from other dimensions of the data. The open-ended nature of the latter, however, also lends itself to a more exploratory and ad-hoc data cleaning approach, and can often seem to necessitate more technical involvement from the user. One reasonable approach here is to fill in the missing ages of the passengers with the average age of passengers sharing the same honorific.

3 APPROACH AND CONTRIBUTIONS

We show in this section how the key contributions of Arachnid as a generalized visual data cleaning system address this use case in relation to the state of the art.

Comprehensive Visual Selection: The ad-hoc nature of current data cleaning processes requires users to be able to comprehensively select any piece of visualized data considered erroneous. Current visual data cleaning approaches allow users to immediately do so for either tables or rigidly defined bar graphs. Tableau Prep offers basic click-based selections within both and can only generate bar graphs for a single column of data at a time. These approaches prevent

Figure 2: Clicking and dragging from the top of the bar sets the targeted selection to the corresponding value (A). Dragging the top of the bar to selected fields in the table also replaces their values (B).

Figure 3: Interactive selection of data records

the user from freely exploring and selecting data based on anomalies she notices during visualization.

Alternatively, Arachnid offers an intuitive mapping from the data to the visual encoding channel in order to allow users to directly interact with a more comprehensive set of visualized value or record(s). It does so by first categorizing each component of a visualized element as a value type. For example, the bar the user interacts with in Figure 2 is divided into three components: the *cap*, *body*, and *label*. The cap represents a numerical value based on the the dependent axis. The body and label text represent groups of either numeric or string cell values from the set of records filtered by the label of the bar. These values are mapped from the column computed within either the dependent or independent axis, respectively. Because of this abstraction, Arachnid can support multifaceted selection of data within bar graphs between any two columns the user chooses.

Arachnid also supports a similar abstraction for other common types of visualizations, including scatter plots, box plots, and tables, and can be extended to other types in future work. We describe how this abstraction allows users to flexibly transform visualized data in the next section.

Generalized Interactions as Cleaning Specifications:

Users need to directly manipulate visualizations in order to quickly execute desired transformations. However, state of the art approaches have limited visual-interactive support. With Tableau Prep, the user can split passenger names for honorifics with a sequence of menu clicks and drag bars representing counts of each misspelled honorific on top of one another for quick deduplication. However, to impute missing values, the user must iterate between current systems. First, she reloads data into visualization software like

Figure 4: Users can write custom code transformations and apply them interactively on to other grouped values.

Tableau Desktop to explore the relationship between the honorifics and ages of passengers. Then, she loads the data into another system like OpenRefine, re-selects groups of passengers by their honorifics, cross-checks with previous visualizations, and executes another sequence of clicks and text entry to impute missing age values.

Building on its framework for visual selection, Arachnid instead allows users to freely modify values based on their semantics. In the case of bar graphs, each component can be modified or used to modify other values as long as involved components have the same type. The cap of a bar (which represents a numerical value) can be modified by directional dragging or by dragging another numerical value on top of it (i.e. a point on a scatter plot). Similarly, the group of numeric values represented by the body of the bar can be modified by drag-and-dropping other numeric values. Like Tableau Prep, the textual label can also be dragged on to another for deduplication. While this means that Arachnid supports the same interactions as current approaches for splitting strings and deduplicating misspellings, exploring the data and directly modifying values becomes an immediate process.

After generating and deduplicating honorifics, we are left with the interface shown in Figure 2. As seen in Figure 2A, once the user confirms there is a reasonable number of complete records with corresponding honorifics as records with missing values, she can drag the *cap* of one bar to the *body* of another to fill in the respective cells. She can also quickly select null records from the raw data table (bottom) with the interaction specified in Figure 3 and choose the records themselves as the drag target, as shown in Figure 2B. Dragging and hovering values over the targetted records also immediately provides a preview of the transformation in the raw data table and the selected records pane (upper right).

These interactions and abstraction of visual components generalize to scripted transformations as well. Arachnid builds upon the interactions supported by Excel to allow the user to write scripts in place for any visual selection of values, records, or columns. To do so, she can click on the visual components and press the $+f(x)$ button. After, she can interactively apply her logic on other visual elements of the same semantic type by selecting and dragging on to them, as shown in Figure 4.

REFERENCES

- [1] Google. 2010. Open Refine.
- [2] Jeffrey Heer, Maneesh Agrawala, and Wesley Willett. 2008. Generalized Selection via Interactive Query Relaxation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 959–968. <https://doi.org/10.1145/1357054.1357203>
- [3] Kaggle. [n. d.]. Titanic: Machine Learning from Disaster. <https://www.kaggle.com/c/titanic>
- [4] Microsoft. 2016. Microsoft Excel.
- [5] Tableau. 2018. Tableau Prep.
- [6] Trifacta. 2015. Trifacta Wrangler.