Senior Honors Thesis Proposal

Tiffany Lee (yeukyul)

March 21, 2017

Abstract

Clean, high-quality data are vital to any statistical analysis and modeling. Most research in data cleaning is dedicated to discovering pattern and spotting anamolies, using an algorithmic or mathematical approach. Yet, few are about making data cleaning more accessible to human users; as a result, data cleaning remains a complicated cognitive task. By analysing challenges in data cleaning with Center of Human Rights Research Group and conducting user study on effectiveness of data manipulation tools, the proposed project aims to explore whether or not a visualization-based data cleaning tool would be more efficient for a boarder user group to perform intended tasks. A prototype driven by the new approach will be created as part of the project to better illustrate research problem and solution.

1 Research Question and Significance

Data cleaning is vital to ensuring data quality: a clean dataset enables efficient data processing and minimizes data misrepresentation. However, data cleaning can be extremely complicated: the process includes, but not limited to, categorizing values, identifying outliers, splitting variable, extracting and unifying values, and handling malformed input.

This can be a big problem, because data manipulation is not a task performed only by statisticians, but also researchers and business executives. These people do not necessarily have the technical background to manipulate a dataset. It is still possible for these users to accomplish the task by editing and inspecting entries manually, but deemed impossible as dimension of a dataset grows beyond factor of a thousand. This poses a challenge to the field of data science.

1.1 Current Research in Data Cleaning

Most data cleaning research focuses on making the earliest stage of data collection *error identification and correction*, more efficiently done. Researchers usually have two polarized approaches to solving the problem: either cleaning data automatically with help from software or through manual inspection.

1.1.1 Automatic Detection

Most representational research using formal approaches focus on data collection stage, specifically on values matching, verification, and schema translation (mapping existing data to common data model). These techniques help to ensure a higher data quality by reducing duplicated entries and correcting malformed input. In addition to data collection, researchers have also proposed qualitative data cleaning: by analysing transcript of the domain the dataset is describing to allow value mapping. Machine learning technique has also been used to discover patterns in large data set and further isolate data points that do not "fit".

1.1.2 Manual Correction

Automated data cleaning, however, has its own limitation; in particular the fact that operations that can be performed on a dataset cannot be enumerated and transformation needed in each dataset depends on its context and purpose. In order for data cleaning to remain flexible and accurate, I believe that data cleaning should remain a manual task.

The question then comes down to: *How can data cleaning be accessible to human users?* The challenges in manual data cleaning, as illustrated earlier, is that it requires expertise in programming, understanding of data, as well as a reasonably manageable data size. Data cleaning is therefore not available to a broad user group.

Jeffrey Heer, a HCI research in University of Washington, has foresaw this problem and therefore dedicated his research in attempting to develop a data wrangling tool that would be easy for human user to perform data cleaning through graphical user interface (GUI). He took a humanistic approach (predictive interaction) by allowing users to only use gestures such like drag-and-drop and highlighting to perform data cleaning tasks. The software will then use machine learning to infer what transformation users are intending to perform on the dataset using given gesture [1].

1.2 Research Question

The approach I wish to take in tackling the challenges in data cleaning is similar to Heer's: I believe that data cleaning should remain a cognitive task, and that more effort should go into making data cleaning more accessible to human users.

Data cleaning can be difficult because programming language is innately hard to understand; therefore human users often times have trouble formulating desirable operations and computer commands. Therefore, I believe that integrating data visualization into data cleaning could be a much effective way to manipulate dataset.

Data visualization is a natural way for people to understand dataset and identify anamolies. By viewing and selecting data that appear unconventional on graphs, users could potentially get a better idea of entry errors and therefore efficiently perform data cleaning. To give a short example, consider a malformed input of year **19997** instead of **1997**. By graphing years, user could easily spot such outlier. By selecting the data points on graph, user could then perform operation to remedy the error. Since statistical graphing and visualization is a universal tool, it could be easily transferable to detect other anamolies.

Therefore the proposed research question is: *Can visualization-based semi-automated tools help users more accurately and efficiently clean large datasets, compared to fully manual or automated approaches?*

1.3 Significance and Applications

The proposed approach is valuable due to its high flexibility and low requirement on users; technical background. Since each dataset has its own unique characteristics, graphical representations would provide liberty for users to select desirable feature to visualize and operate on.

For example, a potential dataset this proposed research will be the Syrian War casualties dataset from Center of Human Rights, which includes the location where these casualties casualties' are found. These location names are often in Arabic, therefore poses challenge to researchers when cleaning data. However, with help of data visualization, user can will be able to plot a choropleth of locations and group similar record by geographical proximity.

This approach can be generalized to various dimension and types of data and can also be useful in discovering subtle pattern or feature of dataset. Since visual representations impose minimal restriction on users' technical background, proposed approach can be valuable to a broader user group.

2 Project Design and Feasibility

The proposed senior honors thesis project will center on exploring effective data cleaning methodology via data visualization. The project will be consisted of four phases: research, design, development, and study.

2.1 Research Phase

During research phase, I will be working with Center of Human Right Research group, most liketly the Syria casualty dataset. This dataset, as illustrated in the previous section, contains casualties data on Syrian War. The dataset contains a lot of missing values and is messy due to difficulty in data collection. This dataset would be an appropriate source to analyse challenges in data cleaning and generate graphical solution. In

addition, past data cleaning approaches employed by the research group can be a valuable comparative analysis and starting point for research.

During this phase, more background research on existing data cleaning tool will be conducted and questionnaire about common challenges data scientists face in data cleaning, after which a concrete methodology of new data cleaning approach will be drafted and revised.

2.2 Design Phase

To evaluate the effectiveness of new data cleaning approach, the project will include a human study that aims to evaluate the functionality and effectiveness of new approaches.

The study will require human subjects to perform data cleaning tasks on a given dataset. In order to compare the effectiveness of data cleaning, human subjects would be exposed to different settings where they are given different tools to perform these tasks. The usability of these software will be evaluated using means like questionnaire, interaction, or variables such as session duration. The details of the study are to be worked out and revised with project advisor, which are then sent to Institutional Review Board (IRB) for approval.

In addition to study design, the project would also include a simple prototype of the proposed data cleaning tool. The purpose of the prototype is to be part of the usability research where user can use and rate effectiveness of different tools.

2.3 Development Phase

The development phase would consist of prototyping with existing software development tools, as described in the previous subsection.

2.4 Study Phase

The study phase would include conducting user study, data analysis, and evaluation. A detailed comparison report on effectiveness of existing and proposed data cleaning methodology will be written as final evaluation of the project.

3 Background

3.1 Technical Background

The most technically challenging part of the thesis would involve prototyping. By the time the thesis begin, I will have had completed a 12-week software engineering internship program, as well as in five semesters of coursework in software development. Prototype development in this proposed project, if well designed, can be completed without much difficulties.

As for the statistical component of the project, bulk of the data analysis work will involve hypothesis testing and data visualization, with which I have had completed at least one semester of coursework. I was also involved in researches that requires background in both, as outlined in the following section.

3.2 Past Researches

In the past, I have taken part in four research projects, both qualitative and quantitative, as well as both in team and as individual. I am currently a technical intern in a Human Computer Interaction (HCI) research lab Articulab under professor Justine Cassell in the CS department. I worked on both SARA and RAPT in system development, driving statistical models, conducting user study, and co-authoring on paper.

The third research project is a Statistics independent research under Professor Sam Ventura. The research centers on statistical software development, specifically R packages.

My last research project dated back to two year ago, was an independent qualitative research on social constructionism with Professor Bonnie Youngs. The research aims uncover the presence of social constructionism through anti-immigrant sentiment in France, using Interpretative phenomenological analysis (IPA) method. The findings were presented in Dietrich Undergraduate Colloqium.

3.3 Project Advisor

My primary project advisor will be Robin Mejia from Center of Human Rights and Statistics department, whom I was introduced to by my academic advisor Sam Ventura. Professor Rebecca Nugent, who is also my academic advisor in Statistics and professor in course 36-401, will also be co-advising on the project.

4 Feedback and Evaluation

During the project, I will be meeting with both Robin and Professor Nugent weekly to report progress. The project will have deliverables in forms of code, documentation, or written reports. The specific timeline of the project will be discussed prior to the beginning of the project.

5 Dissemination of Knowledge

The finding of this research will be presented to the Human Right Research Group and will also be included in articles for journal publications, such as the Journal of Visualization. The work will also be presented in Meeting of the Minds Undergraduate research Symposium as either a verbal or poster presentation.

References

[1] Jeffrey Heer, Joseph M. Hellerstein, and Sean Kandel. *Predictive Interaction for Data Transformation.* 7th Biennial Conference on Innovative Data Systems Research, January 2015.