The Good and Bad Data

Poonam Kumari
University at Buffalo
poonamku@buffalo.edu

Oliver Kennedy
University at Buffalo
okennedy@buffalo.edu

ABSTRACT

Data uncertainty arises from sensor errors, ambiguous human generated inputs, and more. Data cleaning tools help clean dirty data and provide results which are termed as clean data. But how do we define dirty data? For example deleting missing values from data set would help the analyst in one context, while in a different application, deleting missing value might produce bias results. In this paper we go through iterative data cleaning process and the challenges faced by analysts. We conclude by presenting a study design for future work which addresses few of the challenges emerging as part of data cleaning.

1. INTRODUCTION

Data cleaning is an iterative process which is tailored to the requirements of a specific analysis task. Analysts face many challenges while cleaning data which is termed dirty due to incorrect, missing, and duplicate data that are identified and repaired. Although the design and implementation of data cleaning algorithms has improved significantly, these softwares still require some level of expertise from defining data quality rules to manually identifying and fixing errors. This paper presents the challenges faced by analyst, and what changes would help make data cleaning reliable.

2. BACKGROUND

Studies have been conducted to address the challenges faced by analysts during data cleaning process. [1] highlights the issues in current data cleaning work flows and identifies opportunities for facilitating and automating rapid human-in-the-loop interactivity. Interviews were conducted in [2] to analyze the challenges faced by analysts in phases of data cleaning. A visual analytic tool for bridging the gap in programming proficiency of analysts is proposed. Several other cleaning techniques [4,13,14,15] and tools [5,6,7,8,9,10,11,12] have been designed to make the cleaning task easier for analyst. In order to work with these cleaning techniques and tools, analysts need some level of programming proficiency,[2] categorizes analysts into hackers, scripters and application users.

Apart from programming proficiency analysts need to know the definition of dirty data in a particular context. Missing data is often considered as an example of dirty data, and the most easy fix is to delete missing data. But missing data might have some co-relation with other values in the given dataset. So how can an analyst define good data, for example deleting missing values from data set would help the analyst in one context, while in a different application, deleting missing value might produce bias results.

We conducted informal interviews with users who deal with dirty data on a day-to-day basis in order to understand the cleaning work flow and identify pain points.

3. STUDY

We interviewed two users, one of the user worked for a news agency and the other is a professor at an university. Although both users deal with different application data, we identified a common theme in the data cleaning process used. In both cases, data is presented in the form of a csv file and the initial step is to guess the column names based on data distribution. The next step in the work flow is to identify missing values and duplicate data and take corrective action. One of the recurring pain point was the decision to ignore or consider the missing value in analysis. Initially missing data was discarded by both the users. Duplicate data was resolved in the next step using scripts written in R or python. Application specific tasks were performed to clean the data further and then the results were analyzed using visualization tools, which were presented to the domain expert who helped identify outliers.

3.1 Challenges

Few of the challenges that emerged from this study are the scale of the data, definition of correct data, programming proficiency of the user, domain knowledge, and lack of additional information to resolve duplicates, establishing trust in the results produced by cleaning tools. One common pain point faced by both the users is defining good data. [1] proposes an iterative feedback loop for data cleaning which involves both analyst and domain expert. Domain knowledge is needed to define good data and inclusion of domain expert in the cleaning approach helps inspecting or modifying the cleaning pipeline at any step. We do not have to wait till the end of the workflow in order to identify outliers.

Proceedings of the VLDB Endowment, Vol. 11, No. 4
Copyright 2017 VLDB Endowment 2150-8097/17/12.
DOI:

4. CONCLUSION

Domain expertise is required to distinguish between good and dirt data. In order to clean and trust data effectively we need to define what good data is, which differs based on the application domain.

5. REFERENCES

1. Krishnan, Sanjay, et al. "Towards reliable interactive data cleaning: a user survey and recommendations." HILDA@ SIGMOD. 2016.
2. Kandel, Sean, et al. "Enterprise data analysis and visualization: An interview study." IEEE Transactions on Visualization and Computer Graphics 18.12 (2012): 2917-2926.
3. X. Chu, I. Ilyas, S. Krishnan, and J. Wang. Data cleaning: Overview and emerging challenges. In SIGMOD Tutorial, 2016.
4. A. Chalamalla, I. F. Ilyas, M. Ouzzani, and P. Papotti. Descriptive and prescriptive data cleaning. In SIGMOD, pages 445456, 2014.
5. Z. Chen and M. Cafarella. Integrating spreadsheet data via accurate and low-effort extraction. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 11261135. ACM, 2014.
6. X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In SIGMOD, pages 12471261, 2015.
7. M. Dallachiesa, A. Ebaid, A. Eldawy, A. K. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang. Nadeef: a commodity data cleaning system. In SIGMOD Conference, pages 541552, 2013
8. H. Galhardas, D. Florescu, D. Shasha, and E. Simon. Ajax: an extensible data cleaning tool. In ACM Sigmod Record, volume 29, page 590, 2000.
9. D. Haas, J. Ansel, L. Gu, and A. Marcus. Argonaut: Macrotask crowdsourcing for complex data processing. PVLDB, 8(12):16421653, 2015.
10. D. Haas, J. Wang, E. Wu, and M. J. Franklin. Clamshell: Speeding up crowds for low-latency data labeling. PVLDB, 9(4):372383, 2015.
11. S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In ACM Human Factors in Computing Systems (CHI), 2011.
12. Z. Khayyat, I. F. Ilyas, A. Jindal, S. Madden, M. Ouzzani, P. Papotti, J.-A. Quian-Ruiz, N. Tang, and S. Yin. Bigdansing: A system for big data cleansing. In SIGMOD, pages 12151230, 2015.
13. S. Krishnan, J. Wang, M. J. Franklin, K. Goldberg, T. Kraska, T. Milo, and E. Wu. Sampleclean: Fast and reliable analytics on dirty data. IEEE Data Eng. Bull., 2015.
14. S. Krishnan, J. Wang, K. Goldberg, M. Franklin, and T. Kraska. Privateclean: Data cleaning and differential privacy. In SIGMOD Conference, 2016.
15. S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg. Activeclean: Interactive data cleaning while learning convex loss models. In Arxiv: http:// arxiv.org/ pdf/1601.03797.pdf , 2015.