2020-02-04

CAS 771 - Big Data Systems and Applications

Final Project Proposal

Baran Kaya, kayab@mcmaster.ca, 400284996

Contents

1. Introduction

The following is a proposal for entity matching systems/algorithms comparison project. In this project, commercial systems and academic algorithms are going to be compared with their performance and features.

Problem: What are the main difference between commercial entity matching systems and research based entity matching algorithm tools.

2. Background

Entity matching (record linkage, deduplication) is one of the biggest challenges in data management area. One of the common entity matching problem is storing same real-world entity's data in different tables. For example, if the company has 2 different tables for customers and billing and they both store the customer's name with different formats (J. Smith vs John Smith), this would create correctness problem in the data. There are different algorithms and software tools to solve this problem. This project's aim is to compare research based tools to real-world ready commercial tools.

3. Summary

Goal of the project is comparing academic and real-world ready systems. Since it is a course project, comparing 6-8 systems and writing a report could take 2-3 months. I found 10 different systems that consist of both research based open source systems and commercial systems. Since most of the commercial systems require payment, I can only use them in free trial period. That is way I couldn't add systems like IBM InfoSphere because it doesn't have a free version. In the first stage, I am going to check these 10 tools and try to find the best ones for the project.

Commercial entity matching systems:

- DataLadder (30 days free trial)
- SAS DataFlux (14 days free trial)
- WinPure
- Management Ware – Deduplication Tool

Research based entity matching systems:

- Magellan [1]
- Dedupe
- SERF (Stanford Entity Matching Framework) [2]
- Python Record Linkage Tool
- Febrl (Freely Extensible Biomedical Record Linkage)
- FRIL

4. Outcome

While comparing research based and commercial systems, all of them are going to be used in the same scenario with the same data. After that, their results are going to be compared with respect to compute time (performance), result accuracy and their features that they used while calculating the results. Project's and these experiments' aim is to find the main focus of each system. My first predictions about these systems is like that: Commercial systems would focus on the performance while the researched based ones would focus on the algorithm accuracy.

5. Value

The value of this project is finding out what are the main focuses for both commercial entity matching systems and research based algorithm tools.

6. Methods

All of the systems are going to be tested with same data in the same situation. First of all, I need to get entity matching ready data and have to work on it for specific scenarios. After that, research based tools experiment starts. These step consists of 2 parts: experiment

and algorithm's paper review. Then, commercial tools experiment starts. Since they are limited time offer, I am going to use them after researched based ones. So that, I can get some data management experience form the previous experiments and work on them easily and fast. Lastly, I am going to prepare a final report about the experiment results.

7. Schedule

Getting data & preparing it	February 10-16, 2020
Getting all the systems	February 17-24, 2020
Research based system experiments	February 24 - March 15, 2020
Commercial system experiments	March 16 – March 29, 2020
Preparing the final report	March 30 – April, 2020

8. Conclusion

The results of the project will show how commercial products and academic research algorithms differ from each other. Also, with this project, I am going to gain big data management experience.

9. References

[1] Konda, Pradap, et al. "Magellan: Toward building entity matching management systems." *Proceedings of the VLDB Endowment* 9.12 (2016): 1197-1208.

[2] Benjelloun, Omar, et al. "Swoosh: a generic approach to entity resolution." *The VLDB Journal* 18.1 (2009): 255-276.

[3] Bohannon, Philip, et al. "A cost-based model and effective heuristic for repairing constraints by value modification." *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. 2005.

Entity Matching Systems:

- DataLadder: https://dataladder.com/data-deduplication-software/
- SAS DataFlux: http://support.sas.com/software/products/dfdmstudioserver/index.html#s1=1
- WinPure: https://winpure.com/deduplication-software.html#freedemo
- Management Ware – Deduplication Tool: https://datacleansingmatching.com/data-cleansing-data-matching-software-download/

- Magellan: https://sites.google.com/site/anhaidgroup/projects/magellan
- Dedupe: https://github.com/dedupeio/dedupe
- SERF (Stanford Entity Matching Framework): http://infolab.stanford.edu/serf/
- Python Record Linkage Tool: https://recordlinkage.readthedocs.io/en/latest/about.html
- Febrl: https://sourceforge.net/projects/febrl/
- FRIL: http://fril.sourceforge.net/