EMSS: Entity Matching in a Semi-Supervised way

Jia R. Wu
University of Waterloo

Shaokai Wang
University of Waterloo

Abstract

1 Introduction

Entity matching (EM), also known as Entity Resolution (ER), in the world of data management refers to resolving duplicate entities to a single entity. Magellan is an end-to-end entity matching framework that utilizes machine learning to perform entity matching. Active learning in the context of machine learning describes the process of a user actively providing labels for model training. This paradigm of active human labeling is referred to as human-in-the-loop machine learning. In this following paper we are interested in applying active learning techniques to machine learning for Entity Matching.

We provide a comparison of Entity Matching across popular systems such as Magellan, Dedupe.io and JedAI. Comparisons are conducted with the standard DBLP-ACM dataset. Additionally, we provide a fully reproducible and containerized environment for execution requiring no external dependencies other than Docker (reference here). Finally, we extend Magellan with semi-supervised learning and demonstrate that comparable accuracy for EM can be achieved with less samples.

We make the following contributions:

- An evaluation of EM across common systems
- A framework for reproducible execution
- An active-learning extension for EM (semi-supervised)

2 Related Work

Perhaps don't need subsections here. A list may be sufficient.

Talk about dedupe primarily since it is what is being utilized

There has been much work done throughout the years to address entity matching. Efforts ranging from traditional machine learning solutions such as

Talk about structured and unstructured EM

2.1 Magellan

Magellan is described as an end-to-end EM framework as it provides the user with a toolset to conduct EM from the beginning of the workflow to the end.

tools to support the user through various EM scenarios, provides user guides

2.2 Dedupe

Discuss why they have no publication

2.3 JedAI

Talk about what this is.

3 Methods

Have a section on why the DBLP-ACM benchmark was used. Have a section on why we used Docker. Describe what was done to extend. (Active learning part)

4 Results

Describe the performance of using ActiveLearning in Magellan.

Add Key Magellan Result/Conclusion Here

5 Challenges

Describe the ease of use of each system here. Describe the potential performance of each system here.

Initially the challenge was determining why Magellan had not attempted to build

6 Threats To Validity

7 Future Work

8 Acknowledgments

9 Availability

All relevant scripts and data can be retireved from the following repository:

`https://github.com/JRWu/cs848w20`

Notes

[1]The predefined settings can be located here: https://github.com/filebench/filebench/wiki/Predefined-personalities

References