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. Matching may be done in a probabilistic or a deterministic (rule-based) manner. Magellan is an end-to-end entity matching framework that utilizes machine learning to perform both manners of 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 and Dedupe.io [1]. Comparisons are conducted with two datasets BUY and RESTAURANT (add names for these later). Additionally, we provide a fully reproducible and containerized environment for execution requiring no external dependencies other than Docker [2]. 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

In this work, we focus on structured entity matching when resolving two csv tables.

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 through the end. These tools include user guides, strategies for blocking, sampling and matching. The following is a description of the Magellan workflow.

Initially, Magellan computes the cartesian product of both tables. From this resultant set of ordered pairs, an overlap blocker is applied for a resulting set C. The blocking technique used within Magellan is an overlap blocking one. Attributes are tokenized into q-grams and blocked together if they share 1 minimum overlap bewteen q-grams. Subsequently, a sample denoted by S of 450 tuples is taken from this ordered set C. Users are expected to add labels to these, denoting if they are identical entities or not.

I and J are computed from this labeled set S.

I is training set. J is testing set.

Feature vectors are then extracted from tables I and J. An example of a feature in this instance would be a function that maps a tuple pair into a value. If the attributes are textual, a feature could be a 3-gram jaccard score. Sample set S is thus split into training and testings sets, and six various machine learning strategies are applied to match values. The six included are Decision Tree, Support Vector Machine, Random Forest, Logistic Regression, Linear Regression and Naive Bayes. Each matcher is evaluated on the test set J and the test set I for precision and recall.

(verify if true) In theory, the matcher can then be applied to the entire dataset.

2.2 Dedupe.io

Discuss why they have no publication Dedupe.io is available as both an open source and as a paid service for EM. We will focus on the open source implementation implementation of dedupe in this paper.

The workflow of Dedupe.io is

The bread and butter of Dedupe.io consists of 3 parts, a learnable matching function, an active labeling component and a clustering component.

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 retrieved 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

[1] Bilenko, M. Learnable similarity functions and their applications to clustering and record linkage.

[2] Merkel, D. Docker: Lightweight linux containers for consistent development and deployment. *Linux J. 2014*, 239 (Mar. 2014).