CS 838 — Data Science: Principles, Algorithms, and Applications; Spring 2017

Stage 3: entity matching

Trang Ho, Thomas Ngo, Qinyuan Sun

1. Introduction

In this project stage, our team performed matching entities between two tables of education affiliations. The first table was extracted from the Academy of Management Conference (AOM) website, which contains personal information of the conference attendants in the year 2014. The personal information includes (1) individual name, (2) affiliation name, (3) country, (4) states/ province, (5) city, (6) contact numbers, (7) email address. Overall, this table consists of 9,532 entities at the individual level.

The second table was extracted from the World Higher Education Database (WHED), which contains information of unique education affiliations worldwide. This table provides information on (1) affiliation name, (2) country, (3) street, (4) city, (5) province/ states, (6) postal code, (7) telephone number, and (8) website address (if available). Overall, this table consists of 17,605 unique entities at the affiliation level.

In order to match individuals' affiliations on the first table to affiliations on the second table, we used their overlapped/relevant information: (1) affiliation name, (2) country, (3) province/states, (4) city, (5) website address, (6) individual email address. Our goal here is to get precision score of above 95% and recall score of as high as possible.

Subsequently, we carried out the following steps using Magellan:

- Pre-processing
- Down-sizing the AOM table and the WHED table
- Using a blocker to reduce the size of the potential-candidate set
- Sampling randomly 500 pairs of potential candidates for labelling
- Creating training and testing sets I and J
- Training and selecting the best classifier using cross-validation
- Evaluating performance on the testing set J

More details can be found below.

2. Matching procedure

Step 1. Pre-processing

In this step, we cleaned the two datasets by standardizing information on affiliation names, country, state/province, city, email server domain. For example, we standardized states by transforming "CA", "CA - California", "California" to "california" on both the AOM table and the WHED data.

Step 2. Down-sizing

Initially, we have 9,532 entities on the AOM table and 17,605 entities on the WHED data. After down-sizing, we have 4,000 AOM entities and 4,962 WHED entities

Step 3. Blocking

Our blocking consists of the following components:

- Blocking all tuple pairs that have different countries
- For American affiliations, blocking all tuple pairs that have different province/ states
- For all affiliations, blocking all tuple pairs that have neither (1) any overlap between AOM email domain and WHED affiliation website domain nor (2) sufficient overlap coefficient (i.e. greater than 0.5) between affiliation names

As a result, we reduced the size of our candidate set from 19,848,000 (=4,000 x 4,962) to 126,516.

This step took approximately 9 minutes.

In [ ]: