Abstract

Entity Resolution, the process of finding and matching data sets that belong to the same entity, is becoming a greater challenge as the volume and velocity of data explodes in so-called Big Data phenomenon. Entity resolution is used in many different fields such as Entity resolution is used in many different fields that use big data such as government, public health, web, shopping, and etc.

The goal of this project is to create an Interactive Visual Entity Resolution Interface to use in public health research area as part of privacy preserving interactive record linkage (PPIRL) research. The process of Entity Resolution reduces the complexity by grouping the data to particular entities to match and link entities. However, generally speaking, computers, which can be referred to programs or algorithms, by themselves, cannot perform the tasks of entity resolution effectively. Therefore, the Entity Resolution Interface in this project will aid a user to manually resolve complex entities that cannot be resolved automatically by computer.

The dataset for this project contains personal data: first name, last name, date of birth, registration number, gender, and ethnicity. Privacy protection is also a big challenge of this project because the data handled in this project are related with identification. The interface need to conceal most of the information. If two records have exactly same information, the interface hides all of the credentials to protect them. However, if two records contain similar information but slightly different, then the interface shows only parts of the credentials to determine the differences of two or more records that can belong to same entity. Furthermore, if the user cannot determine the identity of the two records by the given parts of the credentials, the interface provides few more information to the user to identify the two records.

The privacy protection problem leads to the biggest challenge of this project. Not only separating the data into proper and separate columns, the interface interacts with the user and show the credentials stage by stage to protect the privacy. To make this feasible, the interface first needs to separate the data into different columns. Then, the separated data need to be clickable and the data needs to be changed if the clickable data are clicked. This is a challenge because various languages and techniques as well as the linking process between them were required to perform the tasks needed to build the interface.

The interface will be displayed on the web and the data will be imported from csv files. Therefore, this project requires some basic knowledge of database and network programming (WebSocket API). In addition, knowledge of web programming (HTML/CSS, Javascript, DOM), Javascript Object Notation (JSON), Javascript library (JQuery) and Javascript data visualization library (d3.js) will be needed in this project.

Outline

Introduction

* Three steps(tasks) of entity resolution in this project
  + Deduplication
  + Record Linkage
    - Privacy Protection
  + Reference Mathing (not started)

Body1 Deduplication

* Deduplication
  + Deduplication is a process of grouping (clustering) the data from uncertain dataset.
  + How cluster is done in this project
* Getting the data from database
* load\_data.js file
  + import database from tow database (csv) files
    - cluster
    - uncertain
  + init() function
    - leads to other functions
    - string encode
    - string decode
    - draw table

Body 2 Record Linkage

* Record Linkage
  + Also called Record Matching.
  + Match similar data.
  + One step further from deduplication
  + How records are linked together for comparison
* Draw Table
  + Basic design of a table (HTML/CSS)
  + Receiving JSON data
  + Displaying table
  + Clickable cells
  + Changing the data when clicked

Body 3 Reference Matching

* Reference Matching
  + If the records belongs to the same entity, the interface keeps them into same record.
  + If they are not, then it will separate them.
  + The final process of this part is not yet finished in the project.

Conclusion

* Wrap up