Named Entity Deduplication by Hybrid Selection Methods

Shun-Hong Sie
Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan, R.O.C.
mayh@ntnu.edu.tw

Hao-Ren Ke
Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan, R.O.C.
clavenke@ntnu.edu.tw

In our previous works, we created the Taiwan Biographical Database, TBDB. It gathers data extracted from the revised Local Gazetteers of ChangHua County (新修彰化縣志). TBDB is an operational platform that helps historians conduct research. It provides the discovery of kinship and social relationships, and renders them in a map or diagram to assist historians in finding interesting topics or information hidden in personage histories. In TBDB, we have already developed an automatic named-entity-recognition algorithm and attain excellent results.

The named entities appearing in TBDB are important information for describing the life of the historical personage. Named entity recognition helps explore the data from full text of the biographies and construct relationship between historical persons. However, due to the divergence of the biographies, a long-time span of 300 years and the inconsistent writing style, even some meaningful words in the full text of the biographies may have low frequency, which means that it is difficult to extract those words by statistically based methods. Change of place names over time, different usage of nouns, and frequency variance of names in each biography make the situation even worse, these data do not have constituency.

During the time change, some nouns and place names obtain from ancient place names or aged person do not appear in current dictionaries or historical materials. The frequency of names in the biography varies, and could not obtain a clear threshold; besides, the distinguishment of persons with the same name is important to reduce the error rate of person relationship networks.

Some named entities in TBDB may denote different persons with the same name, and

these persons will cause wrong connection in generating social networks, and will lead to wrong impression for researchers. In view of this, this study attempts to solve this problem by proposing a hybrid disambiguation method, we will combine some methods which get good result and performance by linear combination.

By trying to use statistics threshold value in pilot study, we found a threshold of 2, which means that named entities must occur at least 3 times, can make deduplication more efficiently. We also try to adopt it to SNA computing by using betweenness and centrality, but got failure due to high computing. this paper attempts to find a better solution to solve the namesake problem.

As we mentioned earlier about high computing or threshold value problem, the statistic or SNA approach both have its limitations. Therefore, we have to find a new method to de-duplicate quickly with moderate cost, and undoubtedly get good result. This study will try to use a hybrid way to combine the result of multiple methods, each of which should achieve name deduplication with acceptable effectiveness and efficiency, by using weight parameters to adjust these methods we will find a good way to make it better than original method. The experiment will be divided into three parts. First of all, the researcher will choose some existent methods which have better performance compare to each selected method by preliminary test and get the baseline. Seconds linear combination, by using weight arguments, we try to combine multi methods in which have good performance together and do test. Finally, historians will evaluate the correctness of the proposed algorithm.

Keywords—Taiwan Biographical Database (TBDB), text retrieval, text mining, social network analysis (SNA), name entity recognition, de-duplicate

Introduction

In our previous works, we created the Taiwan Biographical Database, TBDB. It gathers data extracted from the revised Local Gazetteers of ChangHua County (新修彰化縣志). TBDB is an operational platform that helps historians conduct research. It provides the discovery of kinship and social relationships, and renders them in a map or diagram to assist historians in finding interesting topics or information hidden in personage histories. In TBDB, we have already developed an automatic named-entity-recognition algorithm and attain excellent results. We had extract lot of named entity from gazetteers, some of them may be duplicate which mean different person have same name and them will be set wrong connection to generate wrong social network.

The named entities appearing in TBDB are important information for describing the life of the historical personage. Named entity recognition helps explore the data from full text of the biographies and construct relationship between historical persons. However, due to the divergence of the biographies, change of place names over time, different usage of nouns, and frequency variance of names in each biography make the situation even worse, these data do not have constituency.

Pilot experiment

As fig 1 showing, named entity that repeat twice have 858 count take up 61% deducting those only showing once. Fig 2 show the wrong social network connection, it will make wrong image for researcher by wrong connection between the same named entity. Researchers try to solve this problem by use statistics threshold value approach, we found when set threshold 2 which mean these named entities must have least 3 times can make deduplication more efficiently. We also try to adopt it to SNA computing by using betweenness and centrality, but got failure due to high computing.

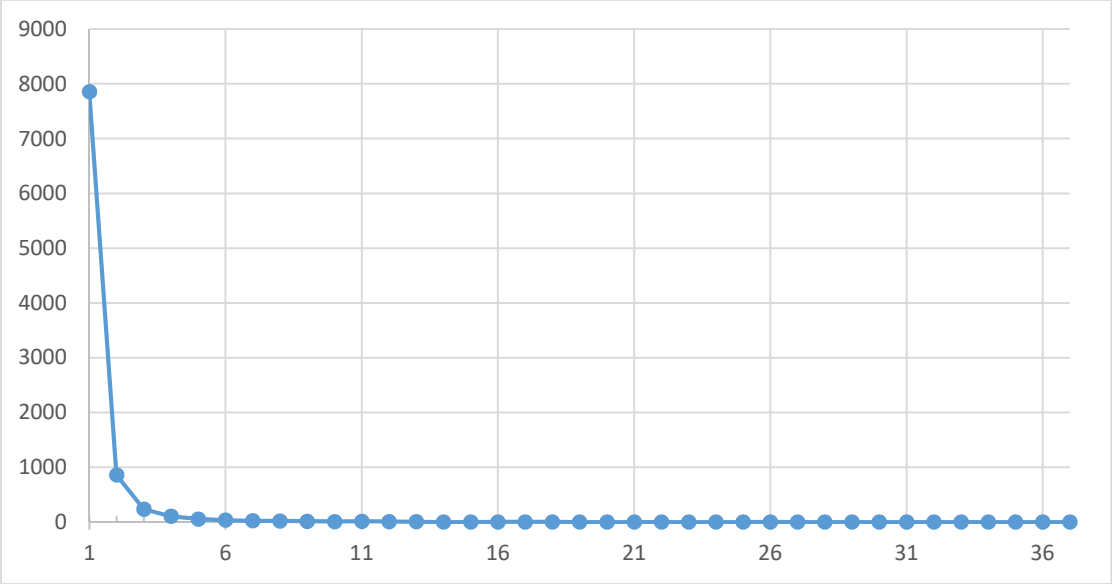

Fig 1. frequency distribution of named entity

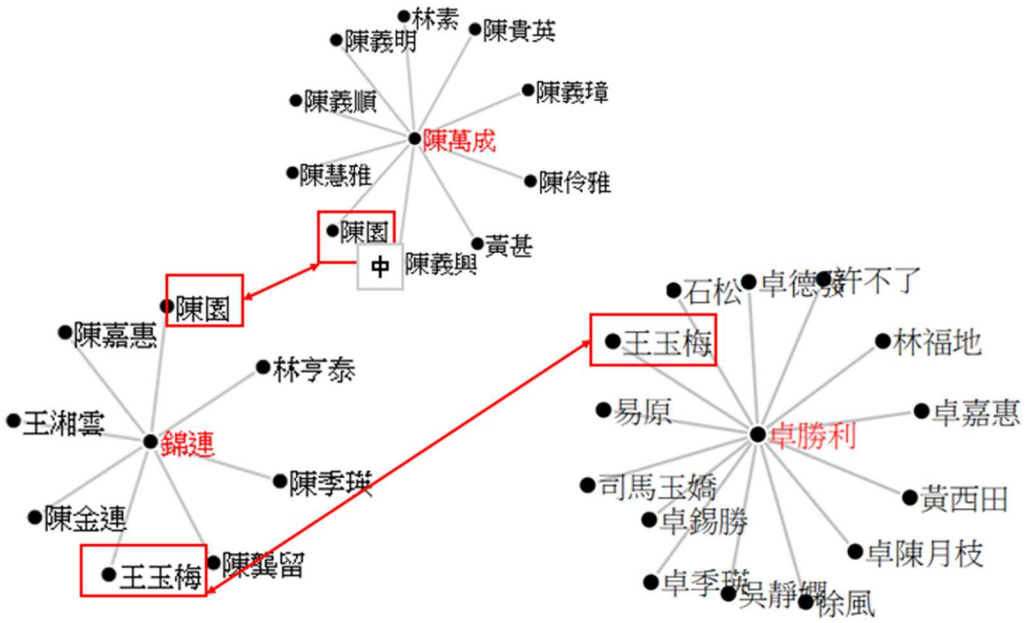

Fig 2: the wrong connection by duplicate named entity.

Fig 3 shows an example of social network different networks make connection by hung_shi which is a duplicated name entity. As we mentioned earlier about high computing or threshold value problem, the statistic or SNA approach both have its limitations. Therefore, we have to find a new method to de-duplicate quickly with moderate cost, and undoubtedly get good result.

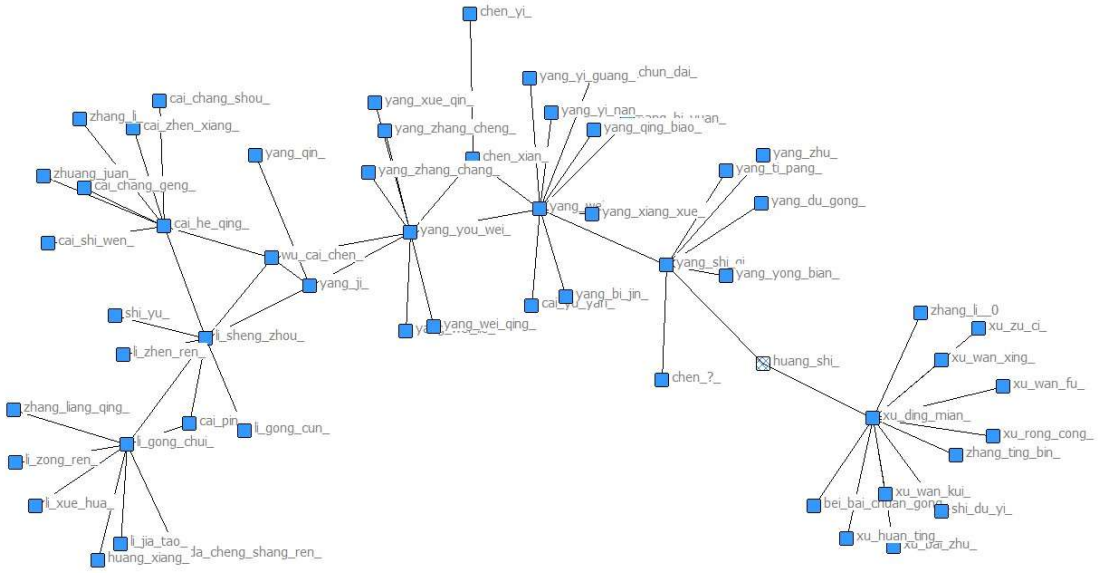

Fig3. The wrong connection by hung_shi drawn by UCINET.

Purpose methods

This study will try to use a hybrid way to combine the result of multiple methods, each of which should achieve name deduplication with acceptable effectiveness and efficiency, by using weight parameters to adjust these methods we will find a good way to make it better than original method.

The experiment will be divided into three parts. First of all, we choice some exist methods which have better performance by preliminary test and get the baseline. Seconds linear combination, by using weight arguments, we try to combine multi methods in which have good performance together and do test. Finally, evaluates the performance of the proposed algorithm with history experts. They will random selected from results to verify its correctness.