A Bipartite Graph Based Social Network Splicing Method for Person Name Disambiguation

Jintao Tang\textsuperscript{1,2}, Qin Lu\textsuperscript{1}, Ting Wang\textsuperscript{2}, Ji Wang\textsuperscript{3}, Wenjie Li\textsuperscript{1}
\textsuperscript{1}Department of Computing, Hong Kong Polytechnic University, Hong Kong
\textsuperscript{2}College of Computer, National University of Defense Technology, Changsha, P.R. China
\textsuperscript{3}National Laboratory for Parallel and Distributed Processing, Changsha, P.R. China
\{tangjintao, tingwang, wj\}@nudt.edu.cn, \{csluqin, cswjli\}@comp.polyu.edu.hk

ABSTRACT
The key issue of person name disambiguation is to discover different namesakes in massive web documents rather than simply cluster documents by using textual features. In this paper, we describe a novel person name disambiguation method based on social networks to effectively identify namesakes. The social network snippets in each document are extracted. Then, the namesakes are identified via splicing the social networks of each namesake by using the snippets as a bipartite graph. Experimental results show that our method achieves better result than the top performance of WePS-2 in identifying different namesakes.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Clustering; I.2.7 [Artificial Intelligence]: Natural Language Processing—Text analysis

General Terms
Algorithms, Experimentation

Keywords
Person name disambiguation, Social network, Bipartite graph.

1. INTRODUCTION
Name is not a unique identifier for a person. Normally, when you search for a person, say, "Andrew McCallum", a search engine will use it as a keyword and return pages relevant to any person with this name without distinguishing different "Andrew McCallum"s. Recently, person name disambiguation has become as an open challenge in many web-mining areas such as social network extraction, web people search and so on.

Recent works on person name disambiguation were reported at the WePS workshops [1]. The typical approach uses the set of webpages that contain the person name to cluster the pages based on the features contained in the pages [2]. However, we thought that the most important issue is to pinpoint each specific namesake based on the features that can uniquely identify them.

It is reasonable to assume that different namesakes have different social circles. Social network information can be a discriminative feature to identify the different namesakes. Existing works use social networks as common word co-occurrence features [3, 4]. However, there usually are millions of webpages mentioned the searched name \( n \), so the names that co-occur with \( n \) in \( d_i \) referred as its neighbors and denoted by \( s_i \), forms the social network snippet for \( n \) as \( d_i = \{s_i, s_2, \ldots, s_n\} \). Variants of \( s_i \) are also resolved and included to address the co-reference problem. The variants of \( n \) include abbreviated forms, last name and first name rotated forms, middle name initialized forms, middle name dropped forms, and combinations of them. From the relationship between \( n \) and its neighbors, we can construct a candidate-neighbor bipartite graph with the set of candidate nodes \( C_n \) and the set of neighbor nodes \( S_n \) as follows. A candidate node \( c_i \) is created for each possible specific namesake of \( n \). First, each document \( d_i \) in \( D_n \) is considered as a candidate node. A neighbor node \( s_j \) is created for each name entity in the documents. A social edge \( e_{ij} \) is created between the candidate node \( c_i \) and a neighbor node \( s_j \) if \( s_j \) co-occur with \( n \) in the same paragraph in \( c_i \). The weight \( w_{ij} \) of edge \( e_{ij} \) is the normalized frequency when both \( s_j \) and \( n \) occur in the same paragraph of \( d_i \). Figure 1 shows an example of a candidate-neighbor graph.

Figure 1. An Example of Candidate-Neighbor Bipartite.

Then we can cluster the bipartite graph to gradually merge similar candidate nodes. It is reasonable to assume that if two candidates share many neighbors, it is likely that they refer to the same person. So we propose a similarity measure based on the bipartite graph, and use a bottom-up method to cluster the candidate nodes for person name disambiguation. The distance between two candidate nodes \( c_i \) and \( c_j \) is,

\[
dis(c_i, c_j) = \sqrt{\sum_{v \in V} (w_{ij} - w_{ij})^2}
\]  

According to this similarity measure, the normalized centroid
method with dimension measures proposed in [5] is applied to clustering of the candidate nodes. The co-occurrence names in the million-scale webpages are high dimensional and sparse, which make the “curse of dimensionality” emerge as a great challenge. We use the dimension array data structure [5], to reduce clustering complexity effectively. Furthermore, we use the clusters as new candidate nodes to generate a new candidate-neighbor bipartite graph to gradually cluster the candidate nodes until the result is stable.

3. EXPERIMENTS
The WePS-2 task dataset [1] is used in our experiment. The WePS-2 test set consists of 30 names, each of which has 150 pages. A similarity threshold for the clustering algorithm to control the granularity of clusters is determined experimentally. Since our objective is to correctly identify different namesakes, we use the WePS-1 dataset to train the similarity threshold by appointing a strict correctness condition. Then, we test the webpages clustering performance of the proposed method by using the B-Cubed measure to compare with the Top 3 systems reported in WePS-2. We also design a two-step clustering method based on [6] to incorporate the bipartite graph based method with textual features so the cluster result from our method is considered as a new document and then the textual features such as URL, token, title and bigram are used to compute the textual similarity among these documents.

Table 1. Clustering Results on WePS-2 Dataset

<table>
<thead>
<tr>
<th></th>
<th>BEP</th>
<th>BER</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WePS2 Top 1</td>
<td>0.87</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>WePS2 Top 2</td>
<td>0.85</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>WePS2 Top 3</td>
<td>0.93</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>Bipartite Graph</td>
<td><strong>0.95</strong></td>
<td>0.53</td>
<td>0.68</td>
</tr>
<tr>
<td>Two-Step</td>
<td>0.90</td>
<td>0.78</td>
<td><strong>0.83</strong></td>
</tr>
</tbody>
</table>

Table 1 shows that the bipartite graph based clustering method is quite discriminative. The B-Cubed precision (BEP) of this method is the highest among all the compared systems. The reason is that the social circles differ a lot in different namesakes and the bipartite graph based method can catch the differences of the social networks. The two-step clustering method also shows high precision as it uses the bipartite graph based method as its first step. It gives the best F1 value because the use of textual features obviously improves the recall for document clustering.

Figure 2. The F1 Performance of Namesakes identification.

However, the B-Cubed measures are not designed for name entity disambiguation. When applied to namesakes being clustered, BEP does not penalize the same namesake being in separate clusters and BER does not penalize different namesakes being in the same cluster. So the criterion for namesakes clustering should be that a namesake is uniquely identified by a cluster. In other words, a cluster can identify a specific person if and only if the webpages in this cluster belong to this person. So, the precision, recall and F1 on the clustering results to namesakes are examined. Figure 2 shows the evaluation on our two proposed methods through the evaluation on namesakes’ identification. Contrary to the result in Table 1, the namesake F1 shows that the bipartite graph based method’s average F1 is 0.563 with standard deviation of 0.094, whereas for the two-step clustering method, its average F1 is only 0.231 with standard deviation of 0.154. This means that the bipartite graph based method is better in its identification of unique namesakes as well as having less volatility. Since the testing data is a closed set, further investigation shows that some persons do not have social network information in these webpages. In these cases, the two-step clustering has better ability to use the additional textual feature to identify similarity of documents as indicated in Table 1 to get better clustering performance. However, the additional textual features also introduce noises, which indeed affects its ability to uniquely identify namesakes.

4. CONCLUSION
A new algorithm for person name disambiguation is proposed to use social network information to identify different namesakes. A bipartite graph based method is used to effectively splice the social networks in the high dimensional and sparse data. As a result, different namesakes are identified and the social network of each specific person is generated. Experimental results show that our proposed method is discriminative to identify the namesakes and has better performance than the top performers in WePS-2.

5. ACKNOWLEDGEMENT
The research is supported by NSF of China (60873097, 60933005), the National Grand Fundamental Research Program of China (2011CB302603), and the Human Language Tech Lab, Dept. of Computing, Hong Kong Polytechnic University.

6. REFERENCES