CWePS: Chinese Web People Search

Jianhua Yin and Lili Jiang

Department of Computer Science and Technology,
Tsinghua University, Beijing, China
{jhyin12,juie.jiang}@gmail.com

Abstract. Name ambiguity is a big problem in personal information retrieval, especially given the explosive growth of Web data. In this demonstration, we present a prototype Chinese Web People Search system, called CWePS. Given a personal name as query, CWePS collects the top results from the existing search engines, and groups these returned pages into several clusters. Ideally, the Webpages in the same cluster are related to the same namesake. Specially, we propose a multi-stage strategy to deal with the Chinese personal name disambiguation on the Web and extract some prominent key phrases to describe each namesake.

1 Introduction

Much work has been done on English Web people search [2,4] and this paper presents a Chinese Web people search system. Compared with people search in English, Chinese text processing confronts additional challenges: first, the Chinese text is written in continuous character strings without any word gap in a sentence, which leads more difficulties to the task of tokenization; second, the substrings of a personal name also denote personal names, which is difficult to distinguish; thirdly, there is a higher personal name ambiguity in Chinese than that in other languages. Some related work has been studied, such as the Chinese Personal Name Disambiguation (CPND) campaign[3], which is held to explore personal name disambiguation in Chinese News. However, to the best of our knowledge, few research studies have been done on Chinese Web people search.

To handle these challenges, this paper presents a Chinese Web People Search system called CWePS. Given a personal name as query, we obtain some Web pages from the existing web search engine, and group the Webpages into different clusters using a multi-stage algorithm. On the first stage, we cultivate the context of the namesakes in Knowledge bases (e.g., Baidu Encyclopedia[1]), and then combine the Webpages related to the same namesake into new documents after matching the Webpages to these namesakes. On the second stage, we group these new documents and the unmatched Webpages into clusters using the Hierarchical Agglomerative Clustering (HAC) algorithm based on Can't Link (CL) and Must Link (ML) features. An example of the CL features is birthday for which each namesake can only have one instance, and the ML features are unique to each person like personal email address.

[1] http://baike.baidu.com/

J. Wang et al. (Eds.): WAIM 2013, LNCS 7923, pp. 792–795, 2013.

Our contributions are as follows: (1) we address the Chinese Web people search problem and develop a system, CWePS, to handle its special challenges; (2) we cultivate the Baidu Encyclopedia as well as the entity features to improve the disambiguation performance; (3) CWePS focuses on the Chinese Web People Search problem and gives some intuition for further research.

2 System Architecture

2.1 Preprocessing

Given a personal name, we use a data collector to fetch a list of search results from Google. First, we remove some noise from the results (e.g., the advertisement, HTML and JavaScript codes) using HtmlParser[2] . Second, we employ ICTCLAS[3] for the tasks of Chinese word segmentation and part of speech (POS) Tagging. Third, we discard some noisy Webpages, such as the ones without any variants of the queried name and the Webpages in which the queried name is tagged as a non-personal name. In addition, we use Email as ML feature and birthday as CL feature. These features are extracted by regular expression from contexts around the queried name.

We use the vector space model (VSM) to represent the Webpages as $W = \{w_1, w_2, ..., w_K\}$, in which w_i is the vector of features of the ith Webpage and K is the number of Webpages. We extracted three kinds of features, including nouns, verbs and named entities tagged by ICTCLAS. The weights of the features are measured by Term Frequency (TF) and Inverse Document Frequency (IDF). Compared with a word, the named entity makes more sense for entity (e.g., person) name disambiguation, thus, we assign the higher weights to the named entities.

2.2 Disambiguation

The disambiguation of Chinese personal name is divided into two stages. In the first stage, we extract the descriptions of the namesakes for each queried name from Baidu Encyclopedia as $B = \{b_1, b_2, ..., b_M\}$, in which M is the number of its namesakes in Baidu Encyclopedia. We calculate the similarity between each Webpage and the description of the namesakes by calculating the cosine similarity [1]. If the similarity is beyond a threshold, we conclude that the Webpage belongs to the corresponding namesake.

In the second stage, we collect the Webpages that are matched into the same namesake in the first stage, then we represent these new documents and the unmatched Webpages as $D = \{d_1, d_2, ..., d_N\}$ using the vector space model (VSM). We utilize the hierarchical agglomerative clustering (HAC) algorithm [1] based on the CL and ML features to group these new documents into new clusters. Next, we compute the $N \times N$ similarity matrix using Cosine similarity and we further build the CL and ML matrix between these documents. After that, we

[2] `http://htmlparser.sourceforge.net/`
[3] `http://www.ictclas.org/`

execute $N - 1$ steps of merging the currently most similar clusters as long as they don't have different CL features. The algorithm starts with merging the clusters with the same ML features and stops until the similarity between the most similar clusters is below a threshold.

3 System Demonstration

When a user searches "Wei Shen", the results of CWePS are shown in Figure 1. Each cluster is represented with its most frequent related person, place and organization. When the user clicks each cluster, the Webpages in that cluster will be shown on the right panel. On the left panel of Figure 2, the area of each sector is proportional to the quantity of Webpages of each cluster. When the user clicks any sector, the corresponding Webpages of that cluster will be shown on the right panel, meanwhile the related solid points will be highlighted to show the Webpages' rank distribution in Google search results.

From the results we observe: (1) the combination of the multiple features is useful for personal name disambiguation; (2) the CL and ML features can improve the disambiguation accuracy of personal name disambiguation; (3) some Webpages collected from the social networks (e.g., Weibo.com[4]) have little context but much structured information, which can be used in the future.

Fig. 1. The Result Interface of CWePS

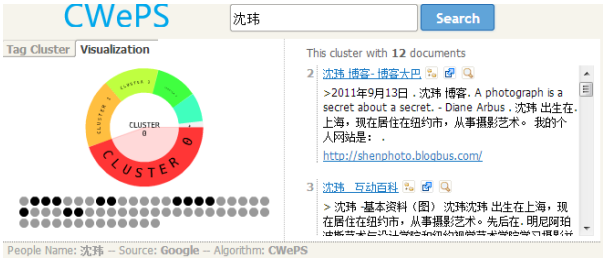

Fig. 2. The Visualization Interface of CWePS

[4] http://www.weibo.org/

Acknowledgement. This work was supported in part by National Natural Science Foundation of China under Grant No. 61272088.

References

1. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
2. Jiang, L., Wang, J., An, N.: Grape: A graph-based framework for disambiguating people appearances in web search. In: IEEE International Conference on Data Mining 2009, pp. 199–208 (2009)
3. Ying, C., Jin, P., Li, W.: The Chinese Persons Name Disambiguation Evaluation: Exploration of Personal Name Disambiguation in Chinese News. In: The First Conference on Chinese Language Processing (2010)
4. Javier, A.: Web People Search. PhD Thesis (2009)