Two Birds with One Stone: A Graph-based Framework for Disambiguating and Tagging People Names in Web Search

Lili Jiang[1], Jianyong Wang[2], Ning An[3], Shengyuan Wang[2], Jian Zhan[2], Lian Li[1]
[1]School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Oracle Corporation, Nashua, NH, USA
jianglili06@lzu.cn
{jianyong,wwssyy,zhanjian}@tsinghua.edu.cn;ning.an@oracle.com;lil@lzu.edu.cn

ABSTRACT

The ever growing volume of Web data makes it increasingly challenging to accurately find relevant information about a specific person on the Web. To address the challenge caused by name ambiguity in Web people search, this paper explores a novel graph-based framework to both disambiguate and tag people entities in Web search results. Experimental results demonstrate the effectiveness of the proposed framework in tag discovery and name disambiguation.

Categories and Subject Descriptors: H.3.0 [Information Storing and Retrieval]: General

General Terms: Algorithms, Experiment, Measurement

Keywords: Name Disambiguation, Tagging, Clustering

1. INTRODUCTION

Submit a query "John Smith" to Google, and the top 100 returned Web pages may refer to at least 10 namesakes. One major challenge is how to locate all the information about a specific "John Smith" quickly. An ideal Web people search engine would return a list of clustered pages, and each of these clusters pertains to one specific "John Smith". Then the user might select the target "John Smith" and directly access all his relevant information. However, most popular search engines leave this difficult task to users.

Various approaches have been proposed to address Web people search and related challenges, including Vector Space Model for entity coreferencing, and clustering techniques in Web people name resolution [3]. There are also a few commercial people search engines including Wink (www.wink.com) and Spock (www.spock.com). Nevertheless, the increasing growth of Internet makes people search a big challenge.

We have observed that people tag information, such as locations and e-mail addresses, appears to be informative, and a combination of such tags can almost identify a unique target people. Based on this observation, we propose a novel weighted-graph based framework which makes full use of the extracted tag information and can be used to solve the problem of both disambiguating and tagging Web people names. The contributions of this work include: a) A novel weighted graph representation was proposed to model the relationships between tags; b) An effective graph clustering algorithm was devised to disambiguate and tag people names; c) An extensive performance study was conducted, which shows that the proposed framework achieves much better performance than previous approaches. In the rest of the work, we outline our framework in Section 2, evaluate it in Section 3, and conclude this paper in Section 4.

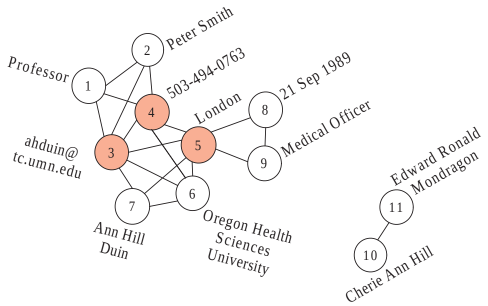

Figure 1: An Example of Tag Graph

2. THE FRAMEWORK FOR WEB PEOPLE NAME DISAMBIGUATION AND TAGGING

2.1 Graph Modeling

Given a people name as query, let document corpus $D = \{d_1, d_2, \ldots, d_k\}$ be the top k returned results from a search engine. After preprocessing, we extract seven types of tags T = {people name, organization, location, email address, phone number, date, occupation} from chunk windows around the query name in each cleaned document $d \in D$. $A = \bigcup_{i=1}^{7} T_i$, is a union of seven tag sets where T_i contains the non-repetitive tags with the type $t_i \in T$.

Regarding a people name, the proposed framework views the tag corpus A as an undirected labeled graph G=(V,E), where the node set $V = \{v_1, v_2, \ldots\}$ is a set of unique tags, and each edge $(v_i, v_j) \in E$ represents that the tags corresponding to v_i and v_j co-occur in a same document in D. We define the tags occurring in multiple documents as *bridge-tags*, and the maximal clique subgraph containing all tags in a single document as a *micro-cluster*. Figure 1 shows an example of tag graph, where nodes 3, 4 and 5 each corresponds to a *bridge-tag*, and there are in total four *micro-clusters*. Additionally, we also compute a weight for each node and each edge as follows.

Node Weighting. The seven types of tags have different characteristics when surrounding the people name. For

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

example, there is a low probability that an email or telephone number is shared by two namesakes, but two namesakes could be related to the same organization or location. This observation leads us to assign different type weights to nodes with different tag types. Intuitively, an ideal tag type should have the following property that each of its tags is uniquely associated with one namesake, while the more the number of distinct tags of one type for each namesake w.r.t. a query name, the less this tag type contributes in name disambiguation. In this work, we assign a type weight to each node according to the following heuristic: the higher the number of unique tags of a certain type, the smaller the weight of each node with this tag type. Thus, the tag weighting function is defined as Formula (1). Each tag type weight in the given corpus D is obtained by dividing the number of all unique tags by the number of unique tags with the corresponding type in T, and then taking the logarithm of that quotient. Finally, each weight is normalized using the sum of weights of all seven types.

$$typedweight(v) = \frac{\log_2 \frac{|A|}{|T_i|}}{\sum_{j=1}^{7} \log_2 \frac{|A|}{|T_j|}} \tag{1}$$

Where v denotes any tag with the type $t_i \in T$. $|A|$ is the number of non-repetitive tags in D, and $|T_i|$ is the number of unique tags of the type "t_i".

Edge Weighting. Assuming the conditional independence between any two unique tags, we have the joint probability for two connected nodes v and w, $p(vw) = p(v)p(w)$, then the edge between v and w, is weighted as Formula (2).

$$edgeweight(v, w) = \sum_{i=1}^{n} p_i(v)p_i(w) \tag{2}$$

where n is the number of documents in which tags v and w co-occur, $p_i(v)$ is the probability distribution obtained from the frequency of tag v divided by the total number of all tag occurrences in the ith document d_i.

2.2 Clustering Algorithm

We then develop a two-step clustering approach for the problem of name disambiguation. The first step is to simply identify the complete set of *micro-clusters* (i.e., the maximal clique subgraphs). The second step is to use a single link clustering method to further cluster the set of *micro-clusters* generated from the previous step into a final set of *macro-clusters*. Following we propose a new method to define the similarity between two intermediate clusters. In our method, each *micro-cluster* M is stored as a vector of tags $\overrightarrow{M}$. The similarity between any two *micro-clusters*, *ConnectivityStrength*, is estimated using cosine distance. If the *ConnectivityStrength* is above a predefined threshold, the tags in these clusters are considered to mention the same people entity and are merged into one cluster.

Differently from the TF/IDF formula, the weight of any tag v in $\overrightarrow{M}$ is defined by the following *Cohesion* formula.

$$Cohesion(v, \overrightarrow{M}) = typeweight(v) + \frac{\sum_{j=1}^{m} typeweight(w_j)}{m} + \sum_{j=1}^{m} edgeweight(v, w_j) \tag{3}$$

Formula (3) computes the importance of tag v to the *micro-cluster* M. It consists of three parts, the *typeweight* of tag v itself, the average *typeweight* of all the other tags in M, and the sum of *edgeweights* of all the edges starting from v and ending at other nodes in M. Compared with TF/IDF, which is a weight derived from term frequency and inverse document frequency, the *Cohesion* formula considers the connectivity of any *bridge-tag* with other tags in a *micro-cluster* based on both type weight and tag frequency, which is more effective in people disambiguation.

2.3 Tagging a Namesake

The output of the above clustering algorithm is a set of tag clusters, each of which represents a unique people entity. Tags with higher *Cohesion* weight in a cluster are selected to describe the corresponding namesake effectively. For demonstration purpose, Table 1 shows the tag clusters generated by our approach for three namesakes regarding the query name of "Marcy Jackson".

Table 1: An Example of Tag Clusters

Marcy Jackson^1	marcy.jackson@montgomerycollege.edu, Maryland Motor Vehicle Administration, Driver Education, Wheel Guidelines, 240-683-2589, ...
Marcy Jackson^2	Jackkeou@aol.com, Tallahassee, 850-894-0835
Marcy Jackson^3	editor, Wilton High School, Middlebury College, the World Languages Department, the University of Michigan, Veronica Foster, ...

3. EXPERIMENTAL EVALUATION

We evaluated our *Cohesion* weight based clustering algorithm in comparison with the top 5 best systems in [2]. We used the same dataset as the one in [2]. The measures used in the experiments are F_{EB} (the Extended B-cubed measure based on *Precision* and *Recall*) and F_{PI} (the harmonic mean of *Purity* and *Inverse Purity*) [1]. The threshold for *ConnectivityStrength* used in clustering was empirically estimated based on the training data.

The experimental results show that our proposed approach achieves much better overall performance (measured by F_{EB} and F_{PI}) in name disambiguation than the state-of-the-art systems. In addition, as shown in Section 2.3, our approach has another advantage, that is, it can output a set of selected tags to describe each cluster (corresponding to a namesake).

4. CONCLUSION

This paper proposes a novel weighted-graph based framework that utilizes extracted tags to both disambiguate and tag Web people names. Specially, we consider type weights for tags and propose a new way (i.e., *Cohesion*) to measure the importance of a tag in an unsupervised clustering algorithm. Experimental results show that the proposed approach outperforms all the solutions in a recent Web people search task [2]. In the future, we plan to apply the proposed framework in a real Web people search system.

5. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China under grant No. 60833003, a research award from Google, Inc., the Guangdong Provincial Government, and State Education Ministry of China under Grant No. 0712226-100097.

6. REFERENCES

[1] E. Amigo, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval Journal*, 2008.

[2] J. Artiles, J. Gonzalo, and S. Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for web people search task. In *Proc. Semeval 2007*.

[3] R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In *Proc. WWW 2005*.