Web Appearance Disambiguation of Personal Names Based on Network Motif

Kai-Hsiang Yang[*], Kun-Yan Chiou[#], Hahn-Ming Lee[*,#], Jan-Ming Ho[*]

* Institute of Information Science, Academia
Sinica, Taipei,Taiwan
{khyang, hmlee, hoho}@iis.sinica.edu.tw

# Department of Computer Science and
Information Engineering National Taiwan
University of Science and Technology
{M9315053, hmlee} @mail.ntust.edu.tw

Abstract

Searching for information about a particular person is a common activity on search engines. However, current search engines do not provide any special function for search a person. Previous research has solved the problem by using additional background knowledge, such as a friend list, to cluster the searched web pages. However, it is still difficult to retrieve and choose suitable background knowledge. In this paper, we propose a Web Appearance Disambiguation (WAD) system to solve the problem by only using the hyperlink structures between web pages. The key idea of the WAD system is to find out smaller node motifs as evidences of close relationship between pages for clustering searched web pages. Our experimental results show that, under no background knowledge, the performance of the WAD system achieves 70% for the F-measure.

1. Introduction

Searching for personal information is one of the most popular search types for search engines. According to the statistics in [11], 5-10% of the queries from the AllTheWeb site include person names. However, current search engines do not provide any special function for searching a person. When a person name is sent to a search engine as a query, the search engine returns many web pages, but we still do not know which pages relate to the person we are interested in. This is known as the "person name ambiguity problem" for web appearance.

To solve the problem Bekkerman and McCallum [2] proposed two unsupervised methods using additional background knowledge, such as a friend list, to cluster the searched web pages; however, it is not the case for search people in reality, because relationship between people is very hard to correctly gather.

In this paper we propose a Web Appearance Disambiguation (WAD) system to cluster the searched web pages from a search engine, so that each group only refers to a specific person. The clustering only depends on the hyperlink structures between web pages. The basic idea in WAD is that when two searched web pages contain some common hyperlinks in the content, the probability that these two pages mention the same people is high. Hence, the WAD system tries to find the connection patterns between pages, and uses these evidences to cluster the searched pages. Our experimental results show that, on average, the clustering performance of the WAD system achieves 70% for the F-measure.

The remainder of this paper is organized as follows. In Section 2, we define the problem and briefly review related works and discuss their limitations. Section 3 introduces the WAD system and its components. Our experimental results are presented in Section 4. Finally, we give the conclusions in Section 5.

2. RELATED WORKS

The person name disambiguation problem has already been studied in many fields, such as in citation matching [6, 9, 12], duplicate record detection and elimination [1, 5], etc. For the problem on Web appearances, Mann and Yarowsky [7] proposed a bottom-up agglomerative clustering approach to group web pages based on a rich biographical feature about a person,. However, these features often relate to private data, so it makes the approach difficult to implement in real world.

Besides, Lloyd, etc. [4] proposed an approach, which uses context data as features to generate vectors, and then applies a k-mean clustering technique to group related pages. But the F-measure metric of the approach is not high enough. Recently, Bekkerman and McCallum [2] proposed two unsupervised methods to solve the problem. As mentioned in the Introduction, relationship between people is very hard to correctly gather.

The person name disambiguation problem is defined as: given a person name N, let $P = \{P_1, ..., P_i\}$ be the set of Web pages replied from a search engine in response to a query about N. The solution of the problem is providing a function f that clusters the Web pages of P into several clusters, $C = \{C_1, ...,C_j\}$, so that each cluster C_j refers to a particular person. The optimal solution of the function f is when $j = i$ and pages in each C_j are related to the same person.

The evaluation metric we use for this clustering problem is the F measure. Suppose that the answer to a problem contains i classes of pages, where each class refers to a particular person, and a clustering algorithm outputs j clusters of Web pages. The F measure of the clustering algorithm is then defined as follows.

For cluster j and class i, Recall(i, j) = n_{ij}/n_i and Precision(i, j) = n_{ij}/n_j, where n_{ij} is the number of members of class i in cluster j; n_j is the number of members of cluster j; and n_i is the number of members of class i. The F measure of cluster j and class i is then given by:

$$F(i, j) = (2 * \text{Recall}(i, j) * \text{Precision}(i, j))/ \\ ((\text{Precision}(i, j) + \text{Recall}(i, j)).$$

3. SYSTEM ARCHITECTURE

The architecture of the WAD system comprises five components: (1) a web search component to collect the search results from search engines; (2) a dataset augmentation component to augment the web page dataset by following the hyperlink information given in the web pages. (3) A popular node removal component to remove several popular web sites which may cause too much hyperlink connections between pages, and (4) a network motif detection component to detect few node network motifs. (5) A web page grouping component is used to cluster the web pages into several groups, and returns the grouped results to the user.

3.1 Web Search and Dataset Augmentation

For the searched web pages, denoted by *SP*, the augmentation component tries to enlarge the dataset by following the hyperlink information in the pages. We propose two kinds of augmentation process: (1) an l-level augmentation process, and (2) a host-based augmentation process.

3.1.1 *l*-level augmentation process

The *l*-level augmentation process is based on two types of hyperlink information: **Outgoing link (O-link)** and **Incoming link (I-link)**. In practice, the I-links can be obtained by using the "link:" function provided by Google.

The *l*-level augmentation process is an iterative procedure that enlarges the *SP* in *l* steps. Figure 1 depicts an example of 2-level augmentation process. We denote the augmented dataset by *AP*.

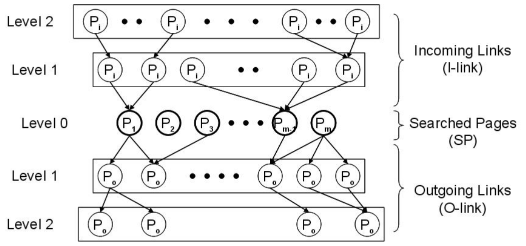

Figure 1: Two-Level Dataset Augmentation

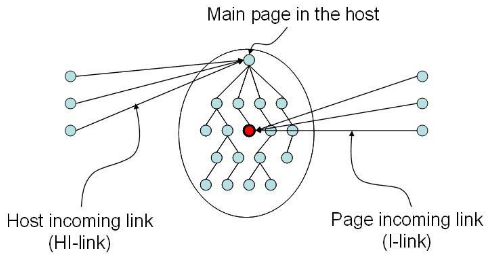

Figure 2: The host incoming link

3.1.2 *Host-Based augmentation process*

According to [3], the number of I-links is far smaller than the number of O-links. To balance the augmented dataset, we then define a host-based augmentation process to collect more I-links. As shown in Figure 2, if there are pages with hyperlinks to the domain name of a page p, those pages will be collected into the dataset by the host-based augmentation process. We denote the augmented dataset by *HAP*.

3.2 Popular Node Removal

A popular page usually links to or is linked by many pages. In the WAD system, this kind of pages easily makes pages to connect to one another, where the probability that the connected pages mention the same people is very low. Hence, a popular node removal (*PNR*) component is defined to remove such popular pages before detecting the hyperlink connections. Here, we use a heuristic to complete the job which works as follows. For a page p, suppose that S_{I-link} is the number of I-links and $S_{HI-link}$ is the number of HI-links. If the product of S_{I-link} and $S_{HI-link}$ is larger than a pre-defined threshold T, it will be removed from the dataset. After the popular node removal process, the WAD system constructs a directed graph to represent the hyperlink structure. Let graph *GLS* = {V,E} be the link structure graph for the augmented dataset, where the nodes of the graph are the Web pages and an edge between any pair of nodes di and dj exists iff di and dj are connected by the I-link, O-link or HI-link.

3.3 Network Motif Detection

The goal of the network motif detection is to find out all connection patterns in the graph which are important evidence to show the connected pages are related to the same person. It is nature to believe that if two pages are connected by a short hyperlink path, the probability that they are related to the same person is high. Hence, we want to find out all small node network motifs in the graph G_{LS}. The network motifs [1, 8] are basically the structural elements (sub-graphs) that form the basic elements of complex networks. To consider all possible connections between two pages, the motifs with path length <= 5 include the 2-node (1 type), 3-node (3 types), 4-node (4 types) and 5-node (10 types) motifs, as shown in Figures 3.

To detect whether two pages are connected, we define two kinds of detections: (1) page-based detection (PD), and (2) host-based detection (HD). In the first PD mode, two hyperlinks meet *iff* all the URLs are the same. In the second HD mode, two hyperlinks are defined as meeting *iff* the domain names of their URLs are the same. Detection using this mode may provide high recall.

3.4 Web Page Grouping

The purpose of the last component in the WAD system is to cluster the web pages, so that each group only refers to a particular person. To achieve the goal, we use the Warshall's algorithm [10] to compute the transitive closure of the graph G_{LS}.

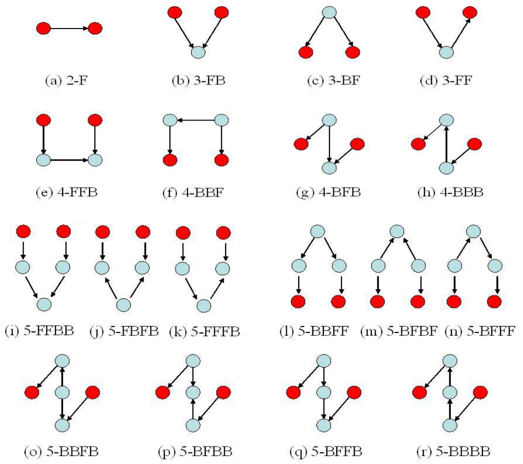

Figure 3: 2-5 node network motifs.

4. EXPERIMENTS

The dataset we use in our experiments is the same as the Bekkerman and Mc-Callum's work [2], and it is the only one available dataset we can obtain. It contains 12,000 Web pages retrieved from Google by querying 12 person names. The non-textual formats, HTTPD error pages, and empty pages are removed from the dataset, and each remaining page is assigned to an answer category. The "other" category in the answer set contains web pages that do not contain a person name; thus, we do not measure the F-measure of this category.

4.1 *Effects of different network motifs*

We first evaluate the effect of each network motif on the precision, recall, and F-measure for the 12 different personal names. As shown in Figure 4, the recall rate clearly increases when large node network motifs are used, and the precision is still above 90% even when 4-node network motifs are used. However, using 5-node network motifs would introduce too much noise. Note that the F-measures for 4-node and 5-node network motifs are almost the same, but the precision drops dramatically for 5-node network motifs. Hence, it is clear that 2-node, 3-node, and 4-node network motifs yield high precision, and 4-node network motifs produce a better F-measure.

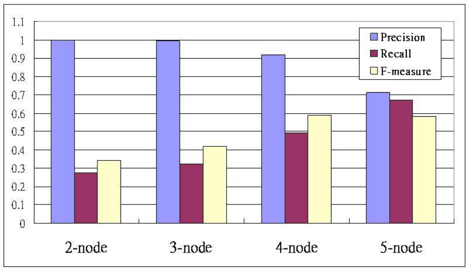

Figure 4: Performance of different network motifs.

4.2 *Effects of network motif detection*

In this experiment, we evaluate the effect of two network motif detections: page-based detection (PD) and host-based detection (HD). Intuitively, we can predict that the PD approach provides high precision, but low recall; whereas the HD approach produces low precision, but high recall. Figure 5 depicts the average F-measure of different approaches. The F-measure of 2-node and 3-node network motifs using the HD approach is 20% higher than that achieved by the PD approach. The reason is that the HD approach improves the recall rate substantially. For 4-node network motifs, the F-measure derived by the HD approach only has 5% improvement over that of the PD approach; for 5-node network motifs, the F-measure is almost the same for both approaches. The results show that the HD approach is suitable for the 2-node, 3-node, and 4-node, because it can match two pages that have the same host name in their URLs, which increases the recall rate. We manually analyzed the dataset and found that many detected 5-node network motifs are hyperlinks to web sites that provide web space for everyone to save their homepages. This kind of hyperlink relation introduces too much noise into the dataset, which is difficult to filter out.

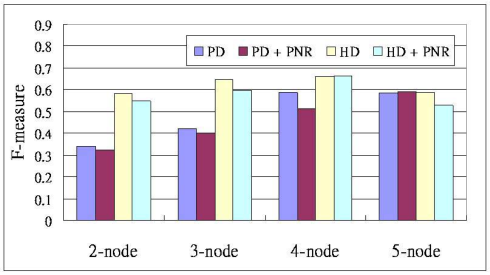

Figure 5: Performance of network motif detections.

4.2.3 *Effects of popular node removal*

We now turn to see the effect of popular node removal and compare the PD approach with PD+PNR, and the HD approach with HD+PNR, as shown in Figure 5. The results show that applying the PNR reduces the F-measure for all network motifs. The reason is that the precision for 2-node, 3-node and 4-node network motifs are high (over 90%), but applying the PNR function removes some popular pages, which reduces the recall rate. The decrease in recall dominates the F-measure, so the F-measure becomes lower.

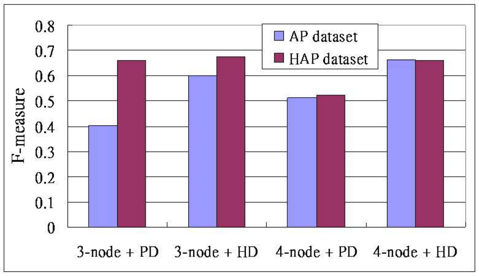

Figure 6: Performance of I-link and HI-link.

4.2.4 *Effects of augmentation processes*

We now evaluate the effects of two augmentation processes: the *l*-level augmentation process, and the host-based augmentation process. Figure 6 depicts the F-measures for the two datasets. It is obvious the F-measure for the HAP dataset is better than that for the AP dataset under 3-node network motifs. However for 4-node network motifs, the F-measures of the HAP and AP datasets are almost the same. Hence, the results show that the HAP dataset is suitable for 3-node network motifs.

4.2.5 *Combination of network motifs*

Figure 7 shows the results of different combination of network motifs. We use the combination of 2-node, 2+3-node, 2+3+4-node, and 2+3+4+5-node motifs under different datasets (AP or HAP) and different network motif detection approaches (PD or HD). Clearly, the F-measure increases when more network motifs are used under the AP+PD; and the best result in this combination is 67%. The reason is that the precision decreases quickly because an error hyperlink connection in the transitive closure will dramatically affect the performance. In our experiments, the best result is derived when using 2+3+4 node network motifs under the HAP dataset and the PD approach (HAP+PD), which reaches 70% of the F-measure. More importantly, the best performance of this setting is very close to using the 4-node network motifs and the HD approach only. Main difference is that the recall rate of the former is 70%, while that of the latter is only 60%.

6. CONCLUSIONS

In this paper, we have proposed a Web Appearance Disambiguation (WAD) system that disambiguates web appearances of personal names by only using the hyperlink structures between web pages. The WAD system tries to find out network motifs as evidence to group searched pages. The WAD system is unsupervised, and does not require any background knowledge about the people. Our experimental results show that, under no background knowledge, the performance of the WAD system achieves 70% for the F-measure.

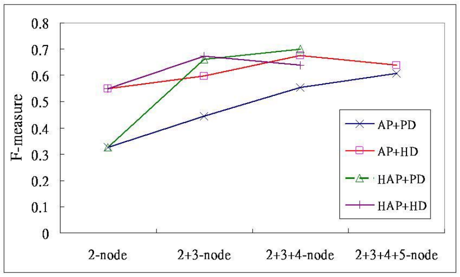

Figure 7: Performance for combinations of motifs.

ACKNOWLEDGMENT

This work was supported by the project of NSC 95-2221-E-001-021-MY3.

REFERENCE

[1] I. Bhattacharya and L. Getoor, “Reduplication and group detection using links”, *Proceedings of LinkKDD conference on Link Analysis and Group Detection*, 2004.

[2] R. Bekkerman and A. McCallum, “Disambiguating web appearances of people in a social network”, *Proceedings of the 14th international conference on WWW*, May 2005, pp. 463-470.

[3] G.R. Xue, Q. Yang, H.-J. Zeng, Y. Yu, and Z. Chen, “Exploiting the hierarchical structure for link analysis”, *Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR’2005)*, Salvador, Brazil, 2005, pp. 186-193.

[4] L. Lloyd, V. Bhagwan, D.F. Gruhl, and A. Tomkins, “Disambiguation of references to individuals”, *IBM Research Report*, RJ10364, 2005.

[5] X. Dong, A. Halevy, and J. Madhavan, “Reference reconciliation in complex information spaces”, *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ACM Press, New York, NY, USA, 2005, pp. 85-96.

[6] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, “Two supervised learning approaches for name disambiguation in author citations”, *Proceedings of the joint conference on Digital libraries*, ACM Press, New York, NY, USA, 2004, pp. 296-305.

[7] G. Mann and D. Yarowsky, “Unsupervised personal name disambiguation”, *Proceedings of 7th Com-putational Natural Language Learning Conference*, Alberta , 2003, pp. 33-40.

[8] R. Milo, S. Shen-Or, “Network motifs: simple building blocks of complex networks”, *Science*, volume 298, 2002, pp. 824-827.

[9] B.W. On, D. Lee, J. Kang, and P. Mitra, “Comparative study of name disambiguation problem using a scalable blocking-based framework”, *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, NY, USA, 2005, pp. 344-353.

[10] S. Warshall, “A theorem on boolean matrices”, *Jour-nal of the ACM*, 9(1):11–12, June 1962.

[11] R. Guha and A. Garg, “Disambiguating people in search”, Stanford University, 2004.

[12] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser, “Identity uncertainty and citation matching”, *Advances in Neural Information Processing Systems*, 2002.