TITPI: Web People Search Task Using Semi-Supervised Clustering Approach

Kazunari Sugiyama
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuta, Midori, Yokohama,
Kanagawa 226-8503, Japan
sugiyama@lr.pi.titech.ac.jp

Manabu Okumura
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuta, Midori, Yokohama,
Kanagawa 226-8503, Japan
oku@pi.titech.ac.jp

Abstract

Most of the previous works that disambiguate personal names in Web search results employ agglomerative clustering approaches. However, these approaches tend to generate clusters that contain a single element depending on a certain criterion of merging similar clusters. In contrast to such previous works, we have adopted a semi-supervised clustering approach to integrate similar documents into a labeled document. Moreover, our proposed approach is characterized by controlling the fluctuation of the centroid of a cluster in order to generate more accurate clusters.

1 Introduction

Personal names are often submitted to search engines as query keywords, as described in a report[1] indicating that about 10% of the English queries from the search engine *ALLTheWeb*[2] contain personal names. However, in response to a personal name query, search engines return a long list of search results containing that contains Web pages about several namesakes. For example, when a user submits a personal name like "William Cohen" as a query to the search engine Google[3], the returned results represent more than one person named "William Cohen." In the results, a computer science professor, an American politician, a surgeon, and others are not classified into separate clusters but mixed together.

Most of the previous works on disambiguating personal names in Web search results employ several kinds of agglomerative clustering approach as described in Section 2. However, in these approaches, a lot of clusters that contain only one element tend to be generated, depending on a certain criterion for merging similar clusters. In addition, in person search results from the World Wide Web (WWW), we can often observe that a small number of entities have a lot of search-result Web pages, while others have only one or two. In light of these facts, if a labeled Web page that describes a person is introduced, clustering for personal name disambiguation would be much more accurate. In the following, we refer to such a labeled Web page as the "*seed page*." Then, in order to disambiguate personal names in Web search results, we introduce semi-supervised clustering that uses the seed page to aid the clustering of unlabeled search-result Web pages. Our semi-supervised clustering approach is characterized by controlling the fluctuation of the centroid of a cluster.

2 Related Work

(Mann and Yarowsky, 2003) first extract biographical information, such as birthdates, birthplaces, occupations, and so on. Then, for each document, they generate a feature vector composed of the extracted biographical information, proper nouns, and the TF-IDF score computed from the documents in the search results. Finally, using this feature vector, they disambiguate personal names by generating clusters based on a bottom-up centroid agglomera-

[1] http://tap.stanford.edu/PeopleSearch.pdf
[2] http://www.alltheweb.com/
[3] http://www.google.com/

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 318–321,
Prague, June 2007.

tive clustering algorithm. (Wan et al., 2005) employ an approach similar to that of (Mann and Yarowsky, 2003), and have developed a system called *WebHawk*.

(Pedersen et al., 2005) recently proposed a method for discriminating names by clustering the instances of a given name into groups. They extract the context of each instance of an ambiguous name and generate second-order context vectors using significant bigrams. The vectors are then clustered such that instances that are similar to each other are grouped into the same cluster.

(Bekkerman and McCallum, 2005) propose the following three unsupervised approaches: (1) an approach based on the hyperlink structures of Web pages; (2) an approach based on agglomerative/conglomerative double clustering (Bekkerman et al., 2005); and (3) a hybrid approach combining the first two.

(Bollegala et al., 2006) first agglomeratively cluster a set of documents and then select key phrases from the resulting clusters to distinguish different namesakes. They extract key phrases from the documents and merge the clusters according to the similarity between the extracted phrases.

3 Our Proposed Approach

In this section, we first review the pure agglomerative clustering approach that most of the previous related works employ and then describe our proposed semi-supervised clustering approach.

In the following discussion, we denote the feature vector $\boldsymbol{w}^p$ of a search-result Web page p in a set of search results as follows:

$$\boldsymbol{w}^p = (w^p_{t_1}, w^p_{t_2}, \cdots, w^p_{t_m}), \tag{1}$$

where m is the number of distinct terms in the Web page p, and t_k $(k = 1, 2, \cdots, m)$ denotes each term. Stop words were eliminated from all Web pages in the search results based on the stopword list[4], and stemming was performed using Porter stemmer[5]. In our preliminary experiments, we found that gain (Papineni, 2001) is the most effective term weighting scheme for generating feature vectors for clustering in this kind of task. Using the gain scheme, we also define each element $w^p_{t_k}$ of $\boldsymbol{w}^p$ as follows:

[4]ftp://ftp.cs.cornell.edu/pub/smart/english.stop

[5]http://www.tartarus.org/~martin/PorterStemmer/

Algorithm: Agglomerative clustering
Input: Set of search-result Web page $p_i (i = 1, 2, \cdots n)$,
$P = \{p_1, p_2, \cdots p_n\}$.
Output: Clusters that contain the Web pages that refer to the same person.
Method:
1. Set the each element in P as initial clusters.
2. Repeat the following steps for all p_i $(i = 1, 2, \cdots, n)$ in P
until all of the similarities between two clusters are less than the predefi ned threshold.
2.1 Compute the similarity between p_i and p_{i+1}
if the similarity is greater than the predefi ned threshold, then merge p_i and p_{i+1}, and recompute the centroid of the cluster using Equation (3),
else p_i is an independent cluster.
2.2 Compute all of the similarities between two clusters.

Figure 1: Agglomerative clustering algorithm.

$$w^p_{t_k} = \frac{df(t_k)}{N}\left(\frac{df(t_k)}{N} - 1 - \log\frac{df(t_k)}{N}\right),$$

where $df(t_k)$ is the document frequency of term t_k, and N is the total number of search-result Web pages.

We also define the centroid vector of a cluster $\boldsymbol{G}$ as follows:

$$\boldsymbol{G} = (g_{t_1}, g_{t_2}, \cdots, g_{t_m}), \tag{2}$$

where g_{t_k} is the weight of the centroid vector of a cluster, and t_k $(k = 1, 2, \cdots, m)$ denotes each term.

3.1 Agglomerative Clustering

In pure agglomerative clustering, initially, each Web page is an individual cluster, and then two clusters with the largest similarity are iteratively merged to generate a new cluster until this similarity is less than a predefined threshold. The detailed algorithm is shown in Figure 1. In this algorithm, the new centroid vector of cluster $\boldsymbol{G}^{new}$ after merging a cluster into its most similar cluster is defined as follows:

$$\boldsymbol{G}^{new} = \frac{\left(\sum^n_{\boldsymbol{w}^{p(G)} \in \boldsymbol{G}} \boldsymbol{w}^{p(G)} + \boldsymbol{w}^p\right)}{n+1}, \tag{3}$$

where $\boldsymbol{w}^{p(G)}$ and n represent the feature vector $\boldsymbol{w}^p$ of a search-result Web page and the number of search-result Web pages in the centroid cluster, respectively.

3.2 Our Proposed Semi-supervised Clustering

As described in Section 1, if a seed page that describes a person is introduced, the clustering for personal name disambiguation would be much more accurate. Therefore, we apply semi-supervised clustering to disambiguate personal names in Web

Algorithm: Semi-supervised clustering
Input: Set of search-result Web page $p_i (i = 1, 2, \cdots n)$,
and a seed page p_{seed}, $P = \{p_1, p_2, \cdots p_n, p_{seed}\}$.
Output: Clusters that contain the Web pages that refer to the same person.
Method:
1. Set the each element in P as initial clusters.
2. Repeat the following steps for all p_i $(i = 1, 2, \cdots, n)$ in P.
   2.1 Compute the similarity between p_i and p_{seed}.
   if the similarity is greater than the predefined threshold,
   then merge p_i into p_{seed} and recompute the centroid of
   the cluster using Equation (4),
   else p_i is stored as other clusters Oth, namely, $Oth = \{p_i\}$.
3. Repeat the following steps for all p_j $(j = 1, 2, \cdots, m, (m < n))$
   in Oth until all of the similarities between two clusters are less than
   the predefined threshold.
   3.1 Compute the similarity between p_j and p_{j+1}
   if the similarity is greater than the predefined threshold,
   then merge p_j and p_{j+1}, and recompute the centroid of the cluster
   using Equation (3),
   else p_j is an independent cluster.
   3.2 Compute all of the similarities between two clusters.

Figure 2: Semi-supervised clustering algorithm.

search results. Our proposed approach is novel in that it controls the fluctuation of the centroid of a cluster when a new cluster is merged into it. In this process, when we merge the feature vector $\boldsymbol{w}^p$ of a search-result Web page into a particular centroid $\boldsymbol{G}$, we weight each element of $\boldsymbol{w}^p$ by the distance between $\boldsymbol{G}$ and $\boldsymbol{w}^p$. As a measure of the distance, we employ the Mahalanobis distance (Hand et al., 2001) that takes into account the correlations of the data set in the clusters. Using Equations (1) and (2), we define the new centroid vector of cluster $\boldsymbol{G}^{new}$ after merging a cluster into its most similar cluster as follows:

$$\boldsymbol{G}^{new} = \frac{\left(\sum_{\boldsymbol{w}^{p(G)} \in \boldsymbol{G}}^{n} \boldsymbol{w}^{p(G)} + \frac{\boldsymbol{w}^p}{D_{mhl}(\boldsymbol{G}, \boldsymbol{w}^p)}\right)}{n+1}, \quad (4)$$

where $\boldsymbol{w}^{p(G)}$ and n are the feature vector $\boldsymbol{w}^p$ of a search-result Web page and the number of search-result Web pages in the centroid cluster, respectively. In Equation (4), the Mahalanobis distance $D_{mhl}(\boldsymbol{G}, \boldsymbol{w}^p)$ between the centroid vector of cluster $\boldsymbol{G}$ and the feature vector $\boldsymbol{w}^p$ of search-result Web page p is defined as follows:

$$D_{mhl}(\boldsymbol{G}, \boldsymbol{w}^p) = \sqrt{(\boldsymbol{w}^p - \boldsymbol{G})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}^p - \boldsymbol{G})},$$

where $\boldsymbol{\Sigma}$ is the covariance matrix defined by the members in the centroid of a cluster. Figure 2 shows the detailed algorithm of our proposed semi-supervised clustering.

In our semi-supervised clustering approach, we use the following two kinds of seed page: (a) the article on each person in Wikipedia, and (b) the top ranked Web page in the Web search results. However, not every personal name in the test data of Web People Search Task has an corresponding article in Wikipedia. Therefore, if a personal name has an article in Wikipedia, we used it as the seed page. Otherwise, we used the top ranked Web pages in the Web search results as the seed page. Table 1 shows personal names classified based on each seed page used in our experiment.

Table 1: Personal names and two kinds of seed page.

Seed page	Personal name
(a) Wikipedia article	Arthur Morgan, George Foster, Harry Hughes, James Davidson, James Hamilton, James Morehead, Jerry Hobbs, John Nelson, Mark Johnson, Neil Clark, Patrick Killen, Robert Moore, Stephen Clark, Thomas Fraser, Thomas Kirk, William Dickson (16 names)
(b) The top ranked Web page	Alvin Cooper, Chris Brockett, Dekang Lin, Frank Keller, James Curran, Jonathan Brooks, Jude Brown, Karen Peterson, Leon Barrett, Marcy Jackson, Martha Edwards, Sharon Goldwater, Stephan Johnson, Violet Howard (14 names)

4 Evaluation Results & Discussion

Tables 2 and 3 show evaluation results in each document set obtained using pure agglomerative clustering and our proposed semi-supervised clustering, respectively. "Set 1," "Set 2," and "Set 3" contain the names from participants in the ACL conference, from biographical articles in the English Wikipedia, and from the US Census, respectively. According to these tables, we found that, although agglomerative clustering outperforms our proposed semi-supervised clustering by 0.21 in the value of purity, our proposed semi-supervised clustering outperforms agglomerative clustering by 0.4 and 0.06 in the values of inverse purity and F-measure, respectively. This indicates that our proposed method tends to integrate search-result Web pages into a seed page and a small number of clusters are generated compared with agglomerative clustering. In terms of these facts, it is easier for a user to browse Web pages clustered based on each personal entity. On the other hand, the small values of purity indicate that irrelevant search-result Web pages are often contained in the generated clusters. Therefore, we can guess that irrelevant search-result Web pages are integrated into a seed page. In fact, we observed that more than 50 search-result Web pages could be grouped together with a seed page.

Table 2: Evaluation results in each document set obtained using agglomerative clustering.

Document set	Purity	Inverse purity	F-measure (alpha=0.5)
Set 1	0.58	0.51	0.45
Set 2	0.67	0.47	0.53
Set 3	0.72	0.47	0.55
Global average	0.66	0.49	0.51

Table 3: Evaluation results in each document set obtained using our proposed semi-supervised clustering.

Document set	Purity	Inverse purity	F-measure (alpha=0.5)
Set 1	0.53	0.86	0.62
Set 2	0.42	0.89	0.55
Set 3	0.41	0.92	0.55
Global average	0.45	0.89	0.57

Table 4 shows the evaluation results obtained using each seed page. The value of F-measure obtained using seed page (a) (0.55) is comparable to that obtained using seed page (b) (0.60). In addition, we could observe that some Wikipedia articles are under updating. Therefore, if the Wikipedia articles are continuously updated, the reliability of Wikipedia as a source of seed pages will be promising in the future. Moreover, observing the results of each person in detail, we found that the purity values are improved when we use a seed page that describes the person using more than about 200 words. On the other hand, in the case where a seed page describes a person with less than 150 words, or describes not only the target person but also some other persons, we could not obtain high purity values.

5 Conclusion

In this paper, we described our participating system in the SemEval-2007 Web People Search Task (Artiles et al., 2007). Our system used a semi-supervised clustering which controls the fluctuation of the centroid of a cluster. The evaluation results showed that our proposed method achieves high scores in inverse purity, with the lower scores in purity. This fact indicates that our proposed method tends to integrate search-result Web pages into a seed page. This clustering result makes it easier for a user to browse the results of a person Web search. However, in the generated cluster with a seed page, irrelevant search-result Web pages are also contained. This problem can be solved by introducing multiple seed pages. In our experiment, we used the full contents of search-result Web pages and a seed page. We consider that this can cause lower scores in purity. Therefore, in future work, in order to improve the accuracy of clustering, we plan to conduct further experiments by introducing multiple seed pages and using parts of search-result Web pages and seed pages such as words around an ambiguous name.

Table 4: Evaluation results based on each seed page obtained using our proposed semi-supervised clustering.

Seed page	Purity	Inverse purity	F-measure (alpha=0.5)
(a) Wikipedia article	0.44	0.96	0.55
(b) The top ranked Web page	0.47	0.81	0.60

References

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. 2007. The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In *Proceedings of Semeval 2007, Association for Computational Linguistics*.

Ron Bekkerman, Ran El-Yaniv, and Andrew McCallum. 2005. Multi-way Distributional Clustering via Pairwise Interactions. In *Proceedings of the 22nd International Conference on Machine Learning (ICML2005)*, pages 41-48.

Ron Bekkerman and Andrew McCallum. 2005. Disambiguating Web Appearances of People in a Social Network. In *Proceedings of the 14th International World Wide Web Conference (WWW2005)*, pages 463-470.

Danushka Bollegala, Yutaka Matsuo and Mitsuru Ishizuka. 2006. Extracting Key Phrases to Disambiguate Personal Names on the Web. In *Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing2006)*, pages 223-234.

David J. Hand, Heikki Mannila and Padhraic Smyth. 2001. Principles of Data Mining. MIT Press, 2001.

Gideon. S. Mann and David Yarowsky. 2003. Unsupervised Personal Name Disambiguation. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pages 33-40.

Kishore Papineni. 2001. Why Inverse Document Frequency? In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 25-32.

Ted Pedersen, Amruta Purandare, and Anagha Kulkarni. 2005. Name Discrimination by Clustering Similar Contexts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing2005)*, pages 226-237.

Xiaojun Wan, Jianfeng Gao, Mu Li, and Binggong Ding. 2005. Person Resolution in Person Search Results: WebHawk. In *Proceedings of the 14th International Conference on Information and Knowledge Management (CIKM 2005)*, pages 163-170.