Personal Name Disambiguation in Web Search Results Based on a Semi-supervised Clustering Approach

Kazunari Sugiyama and Manabu Okumura

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

Abstract. Most of the previous works that disambiguate personal names in Web search results often employ agglomerative clustering approaches. In contrast, we have adopted a semi-supervised clustering approach in order to guide the clustering more appropriately. Our proposed semi-supervised clustering approach is novel in that it controls the fluctuation of the centroid of a cluster, and achieved a purity of 0.72 and inverse purity of 0.81, and their harmonic mean F was 0.76.

Keywords: Information retrieval, Semi-supervised clustering, Personal name disambiguation.

1 Introduction

Personal names are often submitted to search engines as query keywords. However, in response to a personal name query, search engines return a long list of search results containing Web pages about several namesakes. For example, when a user submits a personal name such as "William Cohen" to the search engine Google[1], the returned results contain more than one person named "William Cohen." The results include a computer science professor, an U.S. politician, a surgeon, and others; these results are not classified into separate clusters but are mixed together.

Most of the previous works on disambiguating personal names in Web search results employ several types of unsupervised agglomerative clustering approaches [1], [2], [3], [4], [5]. However, it is hard for these approaches to guide the clustering process appropriately. Therefore, if some Web pages that describe the entity of a person are introduced in a semi-supervised manner, the clustering for personal name disambiguation would be much more accurate. Hereafter, we refer to such a 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 improve the clustering accuracy. Existing methods for semi-supervised clustering can be classified into the following two categories: (1) *constraint-based* [6], [7], [8] and (2) *distance-based* [9], [10]. These approaches aim at refining pure K-means algorithm [11] that needs to set the number of clusters K in advance. However, in our study, the number of namesakes in the Web search results is not known previously. Moreover, they do not consider controlling the fluctuation of the centroid of a cluster although these algorithms focus on introducing

[1] http://www.google.com/

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 250–256, 2007.

constraints and learning distances. We believe that in semi-supervised clustering, it is important to control the fluctuation of the centroid of a cluster that contains a seed page as well as to introduce constraints in order to obtain highly accurate clustering results. Focusing on this point, we propose a novel semi-supervised clustering approach that controls the fluctuation of the centroid of a cluster that contains a seed page.

2 Our Proposed Semi-supervised Clustering

In the following discussion, we denote the feature vector $\boldsymbol{w}^p$ of a 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. In our preliminary experiments for generating feature vectors for clustering in our task, we found that gain [12] is the most effective term weighting scheme. Using the gain scheme, we also define each element $w^p_{t_k}$ of $\boldsymbol{w}^p$ as follows:

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

where $df(t_k)$ is the number of search-result Web pages in which term t_k appears and N is the total number of search-result Web pages. In addition, 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 each term in the centroid vector of a cluster and t_k $(k = 1, 2, \cdots, m)$ denotes each term.

Our proposed approach controls the fluctuation of the centroid of a cluster that contains a seed page 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 the most similar cluster that contains a seed page, we weight each element of $\boldsymbol{w}^p$ by the distance $D(\boldsymbol{G}, \boldsymbol{w}^p)$ between $\boldsymbol{G}$ and $\boldsymbol{w}^p$. We employ the following as a measure of the distance: (i) Euclidean distance, (ii) Mahalanobis distance, and (iii) adaptive Mahalanobis distance. The adaptive Mahalanobis distance is a measure that overcomes the drawback of Mahalanobis distance in that the value of covariance tends to be large when the number of members of a cluster is small. Using Equations (1) and (2), we define the new centroid vector of cluster $\boldsymbol{G}^{new}$ after merging a certain cluster into its most similar cluster as follows:

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

where $\boldsymbol{w}^{p^{(G)}}$ and q are the feature vector $\boldsymbol{w}^p$ of a search-result Web page and the number of search-result Web pages $(q < n)$ in the cluster, respectively. When we merge clusters

Algorithm: Semi-supervised clustering
Input: Set of search-result Web page $p_i (i = 1, 2, \cdots, n)$, and seed pages $p_{s_j} (j = 1, 2, \cdots, u)$,
$Wp = \{p_1, p_2, \cdots, p_n, p_{s_1}, p_{s_2}, \cdots, p_{s_u}\}$.
Output: Clusters that contain the Web pages that refer to the same person.
Method:
1. Set each element in Wp as an initial cluster.
2. Repeat the following steps for all p_i $(i = 1, 2, \cdots, n)$ in Wp
  2.1 Compute the similarity between p_i and p_{s_j}.
    if the maximum similarity is obtained between p_i and p_{s_j},
      then merge p_i into p_{s_j} and recompute the centroid of the cluster using Equation (3),
    else p_i is stored as other clusters Oth, namely, $Oth = \{p_i\}$.
3. Repeat the following steps for all p_h $(h = 1, 2, \cdots, m, (m < n))$ in Oth
    until all of the similarities between two clusters are less than the predefined threshold.
  3.1 Compute the similarity between p_h and $p_r (r = h + 1, \cdots, m)$
    if the maximum similarity is obtained between p_h and p_r,
      then merge p_h and p_r and recompute the centroid of the cluster using Equation (4),
    else p_h is an independent cluster.
  3.2 Compute all of the similarities between two clusters.

Fig. 1. Our proposed semi-supervised clustering algorithm

that do not contain seed pages, we do not control the centroid of a cluster, and define the centroid vector of the cluster as follows:

$$\boldsymbol{G}^{new} = \frac{\left(\sum_{\boldsymbol{w}^{p^{(G)}} \in \boldsymbol{G}}^{q} \boldsymbol{w}^{p^{(G)}} + \boldsymbol{w}^{p}\right)}{q+1}, \tag{4}$$

Figure 1 shows the detailed algorithm of our proposed semi-supervised clustering approach.

3 Experiments

3.1 Experimental Data

In our experiments, we used the WePS corpus established for Web People Search Task [13]. The WePS corpus comprises 79 person sets, each of which corresponds to the top 100 search results of Yahoo![2] via its search API for a person name query. In other words, it contains approximately 7900 Web pages, and 49 and 30 personal names in the training and test sets, respectively.

3.2 Evaluation Measure

We evaluate clustering accuracy based on the *purity*, *inverse purity* and their harmonic mean F adopted in the Web People Search Task. Given a manual classification of the documents into a set of labels, the precision of each cluster P with respect to a label

[2] http://www.yahoo.com/

Table 1. Clustering accuracy obtained using agglomerative and our proposed semi-supervised clustering with one seed page

Clustering approach	Type of seed page	Purity	Inverse purity	F
Agglomerative clustering	no seed page	0.66	0.49	0.51
Semi-supervised clustering				
(i) Euclidean distance	(a) Wikipedia article	0.39	0.90	0.54
	(b) Top-ranked Web page	0.40	0.82	0.54
(ii) Mahalanobis distance	(a) Wikipedia article	0.44	0.96	0.55
	(b) Top-ranked Web page	0.47	0.81	0.60
(iii) Adaptive Mahalanobis distance	(a) Wikipedia article	0.48	0.88	**0.62**
	(b) Top-ranked Web page	0.50	0.78	*0.61*

partition L containing all documents assigned to the label, is the fraction of documents in P which belong to L. The purity is then defined as the weighted average of the maximum precision values of each cluster P, and the inverse purity is defined as the weighted average of the maximum precision values of each partition L over the clusters. Purity and inverse purity achieves maximum value of 1 when every cluster has one single member and when there is only one single cluster, respectively.

3.3 Experimental Results

3.3.1 Experimental Results Using Full Text in the Documents

We compare clustering accuracy obtained using agglomerative and our proposed semi-supervised clustering using full text in seed pages and search-result Web pages. In both approaches, we first determine the optimal similarity for merging similar clusters using the training set in the WePS corpus and then apply it to the test set in the corpus. This similarity is set to 0.0065. Moreover, in our semi-supervised clustering approach, we use the following two types of seed pages: (a) an article on each person in Wikipedia [14] and (b) the top-ranked Web page in the Web search results. We first conducted experiments using one seed page. However, every personal name in the test set of the WePS corpus does not have a 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 page in the Web search results as the seed page. We used Wikipedia article as a seed page for 16 persons and the top-ranked Web page for 14 persons in the test set of the WePS corpus. In a recent work that applies Wikipedia to personal name disambiguation, Bunescu and Paşca [15] identify and disambiguate named entities by using the structures of Wikipedia. Table 1 lists the clustering accuracies obtained using agglomerative and our semi-supervised clustering approach with one seed page.

Moreover, with regard to the adaptive Mahalanobis distance where the best F is obtained in the experiments using one seed page, we conduct further experiments by varying the number of seed pages. Figures 2 and 3 show the clustering accuracies obtained using multiple Wikipedia articles, Web pages ranked up to the top 5, respectively.

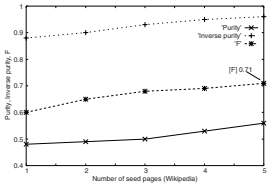

Fig. 2. Clustering accuracy obtained using multiple seed pages (5 Wikipedia articles)

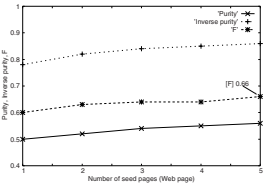

Fig. 3. Clustering accuracy obtained using multiple seed pages (Web pages ranked up to the top 5)

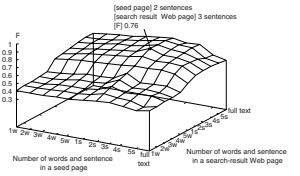

Fig. 4. Clustering accuracy obtained varying the number of words and sentences backward and forward from a personal name in a seed page and a search-result Web page in the case of the 5 seed pages (Wikipedia articles) shown in Fig. 2 ("w" and "s" denote "word" and "sentence," respectively)

3.3.2 Experimental Results Using Fragments in the Documents

We observed that the words that characterize the person often appear around a personal name. Therefore, we vary the numbers of words and sentences backward and forward from a personal name in the case where we used 5 Wikipedia articles as seed pages; in other words, the best value of F (0.71) is obtained in our experiment. In this experiment, using training set in the WePS corpus, we first search for the number of words or sentences around a personal name in a seed page and a search-result Web page that gives the best F. Figure 4 shows that the best F (0.76) is obtained when we use 2 and 3 sentences around a personal name in a seed page and a search-result Web page, respectively. After applying these number of sentences around a personal name to the test set of WePS corpus, we finally obtained the clustering accuracy, (purity:0.72, inverse purity:0.81, F:0.76).

3.4 Discussion

In the agglomerative clustering approach, in Table 1, the high purity (0.66) with low inverse purity (0.49) indicates that the agglomerative clustering tends to generate clusters that contain only one search-result Web pages.

In our proposed semi-supervised clustering approach, Table 1 shows that all the approaches outperform agglomerative clustering with regard to the values of inverse purity and F, although most of the purity values cannot outperform those obtained using agglomerative clustering. We consider that this is due to the effect of controlling the fluctuation of the centroid of a cluster that contains a seed page. In our proposed semi-supervised clustering approach, the best value of F (0.62) is obtained in the case where we employ the adaptive Mahalanobis distance with an Wikipedia article as a seed page. Moreover, in the semi-supervised clustering approach using multiple seed pages, Figures 2 and 3 indicate that the values of both purity and inverse purity improve as the

number of seed pages increases. This shows that introducing seed pages can guide the clustering process more appropriately.

In the experiments using fragments in the documents, we found that we can disambiguate a personal name more effectively by using several sentences than words around a personal name in a seed page and a search-result Web page in the training set of the WePS corpus. This is because we could acquire useful information from sentences that characterize an entity of a person. Moreover, the obtained clustering accuracy (purity:0.72, inverse purity:0.81, F:0.76) is comparable to the top result (purity:0.72, inverse purity:0.88, F:0.78) among the participant systems in Web People Search Task [13].

4 Conclusion

In this paper, we have proposed a semi-supervised clustering approach for disambiguating personal names in Web search results. Our approach is novel in that it realizes highly accurate semi-supervised clustering by controlling the fluctuation of the centroid of a cluster that contains a seed page. In our proposed semi-supervised clustering approach, we introduced some distance measures to control the centroid fluctuation. Experimental results show that our proposed approach achieved the best value of F (0.76) when we simultaneously used 2 sentences backward and forward from an ambiguous name in a seed page and 3 sentences backward and forward from an ambiguous name in a search-result Web page. In future work, we plan to use Web pages hyperlinked from a target page to disambiguate personal names in Web search results and extend our approach to disambiguate place names.

References

1. Mann, G.S., Yarowsky, D.: Unsupervised Personal Name Disambiguation. In: Proc. of the 7th Conference on Natural Language Learning (CoNLL-2003), pp. 33–40 (2003)
2. Pedersen, T., Purandare, A., Kulkarni, A.: Name Discrimination by Clustering Similar Contexts. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 226–237. Springer, Heidelberg (2005)
3. Wan, X., Gao, J., Li, M., Ding, B.: Person Resolution in Person Search Results: WebHawk. In: Proc. of the 14th International Conference on Information and Knowledge Management (CIKM 2005), pp. 163–170 (2005)
4. Bekkerman, R., McCallum, A.: Disambiguating Web Appearances of People in a Social Network. In: Proc. of the 14th International World Wide Web Conference (WWW2005), pp. 463–470 (2005)
5. Bollegala, D., Matsuo, Y., Ishizuka, M.: Extracting Key Phrases to Disambiguate Personal Names on the Web. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 223–234. Springer, Heidelberg (2006)
6. Wagstaff, K., Cardie, C.: Clustering with Instance-level Constraints. In: Proc. of the 17th International Conference on Machine Learning (ICML 2000), pp. 1103–1110 (2000)
7. Wagstaff, K., Rogers, S., Schroedl, S.: Constrained K-means Clustering with Background Knowledge. In: Proc. of the 17th International Conference on Machine Learning (ICML 2001), pp. 577–584 (2001)

8. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised Clustering by Seeding. In: Proc. of the 19th International Conference on Machine Learning (ICML 2002), pp. 27–34 (2002)
9. Klein, D., Kamvar, S.D., Manning, C.D.: From Instance-level Constraints to Space-level Constraints: Making the Most of Prior Knowledge in Data Clustering. In: Proc. of the 19th International Conference on Machine Learning (ICML 2002), pp. 307–314 (2002)
10. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.J.: Distance Metric Learning with Application to Clustering with Side-Information. Advances in Neural Information Processing Systems 15, 521–528 (2003)
11. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability, pp. 281–297 (1967)
12. Papineni, K.: Why Inverse Document Frequency? In: Proc. of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001), pp. 25–32 (2001)
13. Artiles, J., Gonzalo, J., Sekine, S.: The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task. In: Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 64–69 (2007)
14. Remy, M.: Wikipedia: The Free Encyclopedia. Online Information Review 26(6), 434 (2002)
15. Bunescu, R., Paşca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pp. 9–16 (2006)