Selecting Hierarchical Clustering Cut Points for Web Person-Name Disambiguation

Jun Gong
Department of Information System
Beihang University
XueYuan Road HaiDian District, Beijing, China
jungong@ymail.com

Douglas W. Oard
College of Information Studies/UMIACS
University of Maryland
College Park, Maryland USA
oard@umd.com

ABSTRACT

Hierarchical clustering is often used to cluster person-names referring to the same entities. Since the correct number of clusters for a given person-name is not known *a priori*, some way of deciding where to cut the resulting dendrogram to balance risks of over- or under-clustering is needed. This paper reports on experiments in which outcome-specific and result-set measures are used to learn a global similarity threshold. Results on the Web People Search (WePS)-2 task indicate that approximately 85% of the optimal F_1 measure can be achieved on held-out data.

Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Search and Retrieval – *search process, clustering.*

General Terms

Algorithms, Performance, Experimentation.

Keywords

Person-Name Disambiguation, Clustering.

1. INTRODUCTION

In the real world, we often face the problem of name disambiguation; this problem becomes more serious in Web IR due to the huge amount of information available on the Web. Web person-name disambiguation has attracted considerable attention, and many approaches have been proposed [1]. Among them, *Hierarchical Agglomerative Clustering* (HAC) [2] is one of the most widely used. HAC views each Web page containing person-names as a seed cluster, iteratively combining the most similar pair of clusters to form a larger cluster that replaces the pair. Some stopping criterion is used to terminate the agglomeration process in a way that balances the risk of merging references to different people with the risk of failing to merge references to the same person. We can in principle base the stopping point on both global criteria (e.g., the resulting total number of clusters) and on local criteria (e.g., the similarity of items to be merged). We operationalize that approach by using the local criteria to assign a merging score and then thresholding that score based on the global criteria. As Figure 1 illustrates, we can view this process as first creating a dendrogram that represents all possible merging decisions (all the way to one single large cluster), and then selecting an appropriate cut point that partitions the set pages into disjoint clusters. As others have done, we adopt the simplifying "one person per document" heuristic [1]: we treat all mentions of an ambiguous person-name that are found in the same Web page as referring to the same person. This assumption essentially reduces our original problem of clustering individual person-name mentions to the well-studied problem of clustering Web pages.

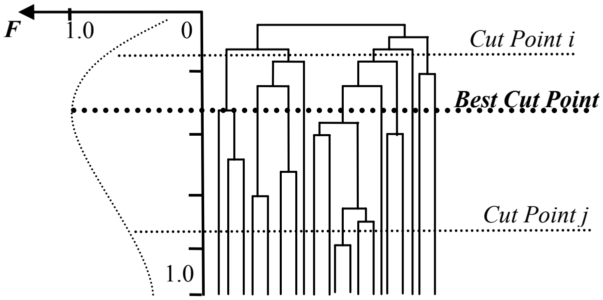

Figure 1. The dendrogram cut-point perspective.

2. METHODS

Our principal interest in this work is in cut point selection based on global features, so we use well-known similarity-based techniques to generate merging scores. In particular, we define a similarity measure between any pair of pages based on five feature types: all automatically detected named entities, the specific matches to the query name, outlink structure, focal text around mentions, and full text. This page similarity measure is described in detail in [3]. As a cluster merging score we use either the minimum inter-cluster page similarity (for "complete link") or the maximum inter-cluster page similarity (for "single link").

As a basis for selection of a global cut-point on merging scores, we define two cut-point dependent features: (1) the merging score threshold value for a cut-point, and (2) the number of clusters that would result from a cut-point. We also define two additional features that depend on the query but not on the specific cut point that is chosen: (3) the mean across all inter-page similarity values (4) the standard deviation across all inter-page similarity values. Some query result sets exhibit a greater central tendency than others; we model that effect with one additional feature. We design a similarity density d_i for page i as:

$$d_i = \sum_j s_{ij} \Big/ (N-1)$$

Where N is the number of pages in the result set (which we set to 100 for training and 150 for testing). We then define a fifth query-dependent (but cut-point-independent) feature: (5) the standard deviation across all pages of the mean inter-page similarity between a page and all other pages (i.e., the standard

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

deviation of *d*). To see why these features might be useful, we can view inter-page similarity (or each page's inter-page similarity density) as a random variable *S* (or *D*) and then examine the Cumulative Distribution Function (CDF) for different person-name queries. Figure 2 illustrates those distributions for "Amanda Lentz" and "Cheng Niu." The distributions are broadly similar, but with notable inter-query differences.

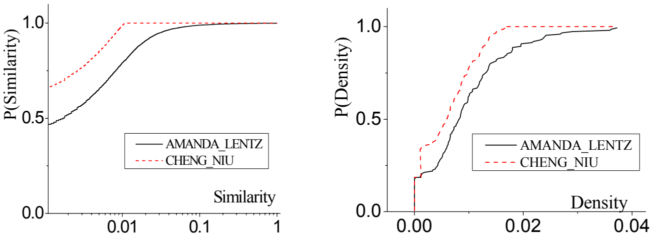

Figure 2. CDF for results from different person-names.

3. EXPERIMENTS

We model our problem of cut-point selection as binary classification using a Support Vector Machine (SVM) with a radial basis function kernel. For each candidate cut-point threshold (C_i) in a dendrogram, we must decide whether to cut at that point based on our five features and on what we have leaened from the 30 person-name training query result sets that are available in the Web People Search (WePS)-1 test collection. We then test the system on the 30 person-name query result sets in the WePS-2 test collection.[1] Our goal is to optimize macro-averaged *F* (the balanced harmonic mean of recall and precision, averaged over ground-truth clusters). We therefore define F_{best} as the optimal (*post hoc*) value of *F* for any cut-point threshold, and we seek to optimize F/F_{best} by varying our process for SVM training. During training, we calculate F_{best} once for each person-name query result set and we calculate *F* for each candidate cut-point threshold on that same result set. We define an acceptance region *r*, a value between 0 and 1 that determines how close to F_{best} a result must be to be used as a positive exemplar during the training process. If $F > rF_{best}$ the cut-point is used. During testing, our SVM may identify more than one suitable cut-point threshold for a given person-name query result set. In such cases, we choose the one that yields a number of clusters closest to the average number of clusters yielded by all such thresholds.

Tables 1 and 2 show some results for an experiment in which we fixed *r=0.4* and trained on thirty WePS-1 person-name queries. The mean F/F_{best} over 30 WePS-2 queries was 0.84% and 82% for single-link and complete-link clustering, respectively.

Table 1. WePS-2 single-link clustering results (*r=0.4*).

	F_{best}	F	F/F_{best}
Tom Linton	0.7768	0.6389	82%
Janelle Lee	0.9333	0.7832	84%
David Tua	1.0000	0.9480	95%
Gideon Mann	0.9698	0.8333	86%
Mike Robertson	0.7425	0.6722	91%

Figure 3 illustrates the sensitivity of F/F_{best} to *r*, again training on 30 WePS-1 queries and testing on 30 WePS-2 queries. Values of *r* between 0.1 and 0.7 are reasonable choices, and the slight advantage of single-link over complete-link clustering that was evident at *r=0.4* persists overt that range.

Table 2. WePS-2 Complete-link clustering results (*r=0.4*).

	F_{best}	F	F/F_{best}
Tom Linton	0.7653	0.7106	93%
Janelle Lee	0.9167	0. 7809	85%
David Tua	1.0000	0.7999	80%
Gideon Mann	0.9443	0.7074	75%
Mike Robertson	0.8385	0.8319	99%

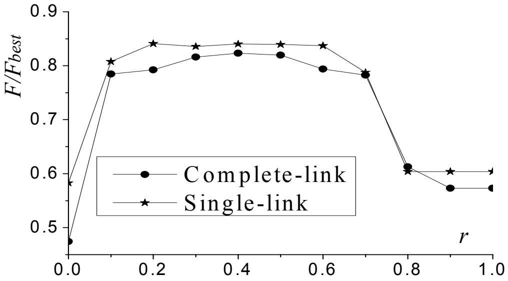

Figure 3. Sensitivity of F/F_{best} to *r* on WePS-2.

4. CONCLUSION AND FUTURE WORK

Our results demonstrate that the method we have described chooses an appropriate cut-point threshold reasonably well, which is an important capability when using hierarchical clustering for Web person-name disambiguation. Similar challenges also arise in other applications. For example, search engines seek to cluster news pages from different sources so that pages talking about the same events can be presented together. As it is impossible to know *a priori* how many events will be described in today's news, the risks of over- or under-clustering must be balanced. Our decomposition of that problem into assigning local merge scores and then choosing a global cut-point threshold is useful, but it prevents us from jointly learning how best to combine local and global evidence. We are, therefore, also interested in exploring applications such as news clustering for which more data might be available (e.g., in the Topic Detection and Tracking collections) to train joint models. What we learn from those applications might then be brought back to inform our future work on the WePS task.

5. ACKNOQLEDGEMENTS

The research is supported by the National Natural Science Foundation of China under Grant No.70671007.

REFERENCES

[1] J. Artiles, J. Gonzalo, and S. Sekine. The SemEval-2007 WePS Evaluation. Proceedings of SemEval, 2007.

[2] C. S. Manning and H. Schutze. Foundations of Statistical Natural Language Processing, The MIT Press, 500-512.

[3] J. Gong and D. Oard. Determine the Entity Number in Hierarchical Clustering for Web Personal Name Disambiguation. Workshop (WePS 2009), 18th WWW Conference, April 2009.

[1] http://nlp.uned.es/weps/