

# A Comparison of Retrieval-Based Hierarchical Clustering Approaches to Person Name Disambiguation

Christof Monz  
ISLA, University of Amsterdam  
Science Park 107  
1098 XG Amsterdam, The Netherlands  
c.monz@uva.nl

Wouter Weerkamp  
ISLA, University of Amsterdam  
Science Park 107  
1098 XG Amsterdam, The Netherlands  
w.weerkamp@uva.nl

## ABSTRACT

This paper describes a simple clustering approach to person name disambiguation of retrieved documents. The methods are based on standard IR concepts and do not require any task-specific features. We compare different term-weighting and indexing methods and evaluate their performance against the Web People Search task (WePS). Despite their simplicity these approaches achieve very competitive performance.

## Categories and Subject Descriptors:

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [Information Systems Applications]: H.4.2 Types of Systems

**General Terms:** Algorithms, Measurement, Performance, Experimentation

**Keywords:** Clustering, Person name disambiguation

## 1. INTRODUCTION

Searching for people is one of the most common web search tasks. An increasing number of people now have a web presence; be it directly through their home page or indirectly, .e.g., through their employer’s or sports club’s web page. This also means that there is a growing number of web pages that are associated with different persons sharing the same name. From a user’s perspective it can be a tedious task to discriminate among the returned search results between the different people. A much preferred solution to this is a system that automatically groups the results in a way that disambiguates between the different people sharing a name.

The Web People Search Evaluation workshops (WePS) [2] focus on the problem of people disambiguation in search and provide data sets to evaluate different approaches.

Our approach described here has been developed and evaluated in the context of the WePS challenge. While most previous approaches tried to integrate person-specific features into their disambiguation approach, our approach is fairly general, requiring hardly any specific adaptations to the person disambiguation task. Yet, at the same time our approach is not only robust but also achieved the second highest performance out of 78 submitted runs in the latest WePS evaluation [3].

Copyright is held by the author/owner(s).  
SIGIR’09, July 19–23, 2009, Boston, Massachusetts, USA.  
ACM 978-1-60558-483-6/09/07.

## 2. RETRIEVAL-BASED CLUSTERING

There are a number of well-established clustering approaches that have been used in various machine learning tasks, with K-Means clustering and agglomerative hierarchical clustering being the most prominent ones. Our experiments confirmed the findings of Balog et al. [4] that agglomerative hierarchical clustering performs best in the context of person disambiguation and we focus our attention on the various strategies within agglomerative hierarchical clustering.

Agglomerative hierarchical clustering is an unsupervised, greedy machine learning approach that iteratively groups items together. Starting with documents being their own cluster (i.e. a singleton cluster), the two clusters with the highest similarity are clustered together, replacing the original two clusters. This step is applied iteratively until only one cluster remains, or the stopping criterion is fulfilled, which can be a fixed number of iterations, or—as we do here—a minimum similarity threshold [7].

At the core of agglomerative hierarchical clustering lies the definition of similarity. The approaches we investigate here are centered around cosine similarity, derivations of which are commonly used in IR. While several of the previous top-performing approaches exploit richer features, including named entity recognition [5, 6], base phrase structure [5], and document structure information [5, 6], we are particularly interested in finding out how far standard retrieval-based measures can take us.

## 3. EXPERIMENTS AND RESULTS

The approaches discussed here are evaluated against the different WePS corpora: WePS-1 dev-test (49 names), WePS-1 test (30 names), and WePS-2 test (30 names). WePS-1 dev-test was distributed before WePS-1 to allow participants to develop their systems, while the latter two were the test sets of the respective WePS evaluations.

The quality of a clustering is typically measured with respect to purity and inverse purity, but recently Amigó et al. [1] have shown that B-cubed precision and recall have most of the desirable properties of a clustering evaluation metric. As B-cubed precision and recall are also the main metrics for the WePS-2 evaluation we use them for comparing our different approaches. B-cubed precision (B<sup>3</sup>-P) and recall (B<sup>3</sup>-R) are defined as [1]:

$$B^3-P = \text{avg}_d \text{avg}_{d' \in c(d)} [d' \in t(d)] \quad (1)$$

$$B^3-R = \text{avg}_d \text{avg}_{d' \in t(d)} [d' \in c(d)] \quad (2)$$

where  $c(d)$  is the set of documents put by the system into the same cluster as  $d$  and  $t(d)$  is the ground truth, i.e. the docu-

**Table 1: Experimental results for the different clustering approaches broken down by the different test sets. The metrics used are B-cubed precision (B<sup>3</sup>-P) and recall (B<sup>3</sup>-R) and their F-measure combination (B<sup>3</sup>-F).**

Run	WePS-1 dev-test			WePS-1 test			WePS-2 test			All		
	B <sup>3</sup> -P	B <sup>3</sup> -R	B <sup>3</sup> -F	B <sup>3</sup> -P	B <sup>3</sup> -R	B <sup>3</sup> -F	B <sup>3</sup> -P	B <sup>3</sup> -R	B <sup>3</sup> -F	B <sup>3</sup> -P	B <sup>3</sup> -R	B <sup>3</sup> -F
a. max-surface-tf_nor	0.931	0.494	0.565	0.910	0.635	0.711	0.973	0.349	0.456	0.937	0.493	0.576
b. centr-surface-tf_nor	0.926	0.622	0.693	0.883	0.739	0.781	0.823	0.666	0.704	0.886	0.666	0.720
c. min-surface-tf_nor	0.888	0.672	0.721	0.781	0.793	0.772	0.893	0.740	0.790	0.860	0.724	0.754
d. min-stm-tf_nor	0.865	0.732	0.756	0.723	0.828	0.755	0.852	0.798	0.807	0.822	0.777	0.770
e. min-stm-hyper-tf_nor	0.893	0.644	0.696	0.801	0.781	0.775	0.864	0.731	0.771	0.860	0.706	0.739
f. min-stm-tf_abs	0.726	0.850	0.742	0.415	0.905	0.547	0.503	0.951	0.608	0.579	0.893	0.651
g. min-stm-tf_nor-loc	0.825	0.653	0.677	0.629	0.810	0.683	0.818	0.638	0.678	0.769	0.692	0.679
h. min-stm-tf_nor-crawl	0.748	0.840	0.752	0.448	0.898	0.569	0.540	0.936	0.644	0.608	0.883	0.672
i. min-stm-tf_nor-wiki	0.648	0.906	0.707	0.274	0.937	0.390	0.466	0.978	0.568	0.495	0.934	0.581
j. min-stm-tf_nor-win	0.901	0.741	0.778	0.785	0.818	0.781	0.726	0.829	0.754	0.821	0.787	0.772
k. min-tf_nor_names	0.843	0.678	0.702	0.770	0.800	0.771	0.606	0.872	0.655	0.758	0.765	0.708

ments with which  $d$  should be clustered together.  $[\cdot]$  returns 1 if the argument statement is true and 0 otherwise. We also combine both metrics in the macro-averaged F-Score (with  $\alpha = 0.5$ ).

Below we describe the dimensions along which we compared the different approaches. Names between parentheses refer to the runs in Table 1.

**Cluster Representation.** The similarity between clusters can be based on the similarity between the two closest/farthest documents (min/max) or the centroid of all documents within the clusters (centr). Overall, the min strategy (c) outperforms the other two ((a) and (b)).

**Term Expansion/Normalization.** The index of the documents can be based on the surface words (surface) or the stemmed terms (stm). Additionally, the document representation can be expanded by WordNet hypernyms (hyper). Stemming improves performance ((d) outperforms (c)), but adding hypernyms does hurt; see (e) vs. (d).

**Term Frequencies.** All the runs so far used normalized term within-document frequencies (tf\_nor), where the tf-score is computed as:

$$tf(t, d) = \frac{1 + \log(\text{freq}_{t,d})}{1 + \log(\text{avg}_{t' \in d} \text{freq}_{t',d})}$$

Using the absolute frequencies instead (abs) again hurts performance substantially; see (f) vs. (d).

**Collection Frequencies.** When computing the idf-scores, some approaches use local (loc) document frequencies, considering only the documents retrieved for the name to be disambiguated. Using the document frequencies for entire collections, i.e. all names, does lead to more reliable counts and better performance; see (g) vs. (d). On the other hand, including documents frequencies from additional background collections such as web crawls (crawl) and Wikipedia (wiki) do again hurt performance; see (h) and (i) vs. (d).

**Term Window.** The improvements of using normalized term frequencies show that clustering is very sensitive to the set of terms that mainly represents a document. Continuing along this line we experimented with varying window sizes (win) and indexed only terms that occurred within  $n$  words from

a mention of the search name (here  $n = 50$ ). This approach lead to further improvements ((j) vs. (d)), and to the best overall system.

**Named Entities.** As some well-performing approaches have focused on indexing named entities only, we include this approach as well (names). For the WePS-1 test collection the performance of (k) comes close to the best simple term-based measure (j), but overall, it falls clearly behind.

## 4. CONCLUSIONS

The approach described here shows that simple methods, mainly using frequency-based statistics can lead to high performance in the person disambiguation task. At the same time, our comparisons show that small variations in the definition of the similarity function can have a substantial impact on performance, warranting careful evaluation.

## 5. REFERENCES

- [1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, to appear.
- [2] J. Artiles, J. Gonzalo, and S. Sekine. The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. In *Proceedings of the International Workshop on Semantic Evaluations (SemEval 2007)*, 2007.
- [3] J. Artiles, J. Gonzalo, and S. Sekine. Weps 2 evaluation campaign: overview of the web people search clustering task. In *2nd Web People Search Evaluation Workshop (WePS 2009)*, *18th WWW Conference*, 2009.
- [4] K. Balog, L. A. Azzopardi, and M. de Rijke. Resolving person names in web people search. In R. Baeza-Yates and I. King, editors, *Weaving Services, Location, and People on the WWW*. Springer, 2009.
- [5] Y. Chen and J. H. Martin. CU-COMSEM: Exploring rich features for unsupervised web personal name disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 125–128, 2007.
- [6] E. Elmacioglu, Y. F. Tan, S. Yan, M.-Y. Kan, and D. Lee. PSNUS: Web people name disambiguation by simple clustering with rich features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 268–271, 2007.
- [7] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.