DAEDALUS at WebPS-3 2010: k-Medoids Clustering using a Cost Function Minimization

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3], José Carlos González-Cristóbal[1,3]

[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.
slana@diatel.upm.es, jvillena@it.uc3m.es,
josecarlos.gonzalez@upm.es

Abstract. This paper describes the participation of DAEDALUS team at the WebPS-3 Task 1, regarding Web People Search. The focus of our research is to evaluate and compare the computational requirements and results achieved by different solutions based on the minimization of cost functions applied to clustering algorithms. Our clustering technique is based on an implementation of k-Medoids algorithm, run over a sparse term-document matrix built with the terms of the pages that are associated to each of the person names. We define an empty-cluster that holds all the individuals that are not part of any other cluster. Based on the results obtained, we can conclude that although clustering techniques play a very relevant role in the resolution of the problem of name homonymy in a set of web pages, there is a previous challenge still to solve: how to determine which contents are relevant for describing the person in that webpage, thus which are not part of the other navigational information contained in the webpage.

Keywords: Web People Search, Word Sense Disambiguation, Cross document coreference resolution, text clustering, k-medoids.

1 Introduction

The Web People Search (WePS) [1] [2], one of the tracks in CLEF 2010, is divided into two tasks: Task 1 is related to Web People Search (WPS) and focuses on person name ambiguity and person attribute extraction on Web pages, and Task 2 is related to Online Reputation Management (ORM) for organizations and focuses on the problem of ambiguity for organization names and the relevance of Web data for reputation management purposes. In this paper, we will focus on Task 1, which addresses the problem of name homonymy. The basic goal of this task is to cluster a set of web pages, which are the result of a Web search for a person name, in as many groups as entities sharing that name.

Our research group is led by and named after DAEDALUS, a small private company in the field of Information and Telecommunication Technologies and a leading provider of language-based solutions in Spain, and research groups of two universities, Universidad Politécnica de Madrid and Universidad Carlos III de Madrid. We have taken part in CLEF since 2003 in many different tracks and tasks, as part of the MIRACLE team till last year. This paper describes our participation at the WebPS Web People Search (Task 1).

Traditionally, the solutions to solve the problem of name disambiguation have been typically based on techniques such as Hierarchical Agglomerative Clustering (HAC) or Incremental Vector Space clustering algorithm, or some of their variants. The idea behind the set of experiments that we have carried out this year is to evaluate the computational cost and the precision values that are achieved when another type of solution based on the minimization of cost function is applied. Specifically, our research has focused on the clustering algorithm based on k-Medoids [3].

2 System Description

Our system is modular and built from of a set of small components that are easily combined in different configurations and executed sequentially to build the final result set. A common baseline algorithm was used in all experiments to process the data collection, following these steps:

1. **Text Extraction:** Ad-hoc scripts are run on web pages to extract the relevant information.
2. **Tokenization:** This process extracts basic textual components. Some basic entities are also detected, such as numbers, initials, abbreviations, and years. So far, compounds, proper nouns, acronyms or other types of entity are not specifically considered. The outcomes of this process are single words, multi-words and years in numbers.
3. **Conversion to lowercase:** All document terms are normalized by changing all letters to lowercase.
4. **Filtering:** All words recognized as stopwords are filtered out. Stopwords in the target languages were initially obtained from the University of Neuchatel's resources page [4] and afterwards extended using our own developed resources.
5. **Stemming:** This process is applied to each one of the words to be indexed or used for retrieval. Standard Porter stemmers [5] for each considered language have been used.
6. **Sparse matrix generation:** For each corpus associated to each person name, a sparse term-document matrix is built. This matrix stores the set of terms that are contained in each corpus, considering all documents corresponding to the same person name as a whole.
7. **Clustering:** This process generates the final results by applying a clustering algorithm over the information stored in the sparse matrix. A set of configuration parameters, depending on the algorithm itself, allow to define different experiment settings.

3 Experiments and Results

As mentioned before, the objective of our experiments was to make an exhaustive evaluation and comparison of the computational costs and results achieved when solutions based on the minimization of cost functions are used as a basic clustering algorithm. For our experiments, we have specifically focused on an implementation of k-Medoids clustering algorithm [3], due to time constraints and lack of resources, although we plan to run the same experiments with other clustering algorithms such as Fuzzy C-Means.

In our settings, we use k as the result of the minimization of the cost function (see Equation 1).

$$\frac{cIntraCluster}{k} + \frac{cInterCluster}{k} * \frac{\#UnaryCluster}{k} + \alpha * k * R * \ln R \qquad (1)$$

where:

k: number of clusters (2 – R/8).

cIntraCluster: summation of the intra-cluster distances.

cInterCluster: summation of the inter-cluster distances.

#UnaryCluster: number of clusters with only one element (the medoid).

R: size of the simple space once removed the vectors without significant components.

α: dispersion coefficient (0.1 - 0.9).

This algorithm is continuously run over the sparse term-document matrix that was built with the terms of the pages that are associated to each person name. The maximum number of epochs has been set to 10. The algorithm was designed allowing no cluster overlapping, i.e., the membership function returns just one value for each individual, or, in other words, each individual belongs to one cluster maximum.

Different experiments have been defined, using different combinations of the textual information present in different components of the web pages with different values of the dispersion coefficient in the cost function to minimize in Equation 1. After studying some strategies, we finally decided to define a cluster (*empty-cluster*) that groups all those individuals that have no significant term after the linguistic processing. Finally we submitted 4 experiments to be evaluated, listed in **Table 1**.

Table 1. Description of the experiment set.

Run Identifier	Component in page	Dispersion coefficient (α)
daedalus_1	Body	0.3
daedalus_2	Metadata + Title	0.3
daedalus_3	Metadata + Title + Body	0.3
daedalus_4	Metadata + Title + Body	0.6

Table 2 summarizes the main features of those experiments. The *Pages* column shows the average number of pages associated to each person name; *k* is the average number of cluster per person name; *UnaryC* is the average number of clusters that have only just one member (thus the medoid); *EmptyCPop* is the average number of individuals belonging to the empty-cluster; *Population* is the average value of the population considering only those clusters whose population is greater than 1; and finally *maxPop* is the average value of the population associated to the top populated cluster for each person name.

All submitted experiments show a cluster that tends to accumulate a high percentage of individuals (87% on average) and also a high proportion of clusters having just one element (61,83% on average).

Table 2. Evaluation of run features.

Run Identifier	Pages	k	UnaryC	EmptyCPop	Population	maxPop
daedalus_1	**191.19**	13.59	8.48	20.51	40.65	144.60
daedalus_2	**191.19**	12.62	7.34	**21.42**	39.34	147.00
daedalus_3	**191.19**	14.79	10.18	2.02	**47.25**	168.81
daedalus_4	**191.19**	**14.82**	**10.20**	2.02	45.72	**169.07**

Table 3 shows the results of applying the Unanimous Improvement Ratio (UIR) of A (rows) over B (columns), assuming as scores F-measure for α equal to 0.5. As shown in the table, the *daedalus_3* experiment exhibits a higher robustness to variations of the α parameter than the rest of the submitted experiments.

Table 3. Unanimous Improvement Ratio (UIR).

Run Id	all_in_one	daedalus_1	daedalus_2	daedalus_3	daedalus_4	one_in_one
all_in_one	0.00	-0.14	-0.12	-0.24	-0.23	-0.01
daedalus_1	0.14	0.00	0.03	-0.10	-0.16	-0.03
daedalus_2	0.12	-0.03	0.00	-0.12	-0.14	-0.03
daedalus_3	**0.24**	**0.10**	**0.12**	0.00	**0.01**	-0.02
daedalus_4	0.23	0.16	0.14	-0.01	0.00	-0.02
one_in_one	0.01	0.03	0.03	**0.02**	0.02	**0.00**

Table 4 shows the evaluation results using B-Cubed measures. As inferred from that metric, the achieved results present, in general, a behaviour closer to *all_in_one_baseline* than to *one_in_one_baseline*. This result is consistent and agrees with the figures shown in **Table 2**. This may be explained that, although some clusters have been identified, the main set of individuals are assigned to a medoid and the rest of the individuals (a few of them) are distributed along clusters whose population is only one or two individuals.

Table 4. Results of experiments using B-Cubed Metrics.

Run Identifier	Avg. BCubed Precision	Avg. BCubed Recall	Avg. F-measure
all_in_one	0,22	1,00	0,32
daedalus_1	0,30	0,71	0,37
daedalus_2	**0,30**	0,72	0,38
daedalus_3	0,29	0,84	**0,39**
daedalus_4	0,28	**0,85**	0,38
one_in_one	1,00	0,23	0,35

4 Conclusions and Future Work

Based on the results obtained, we can conclude that although clustering techniques play a very relevant role in the resolution of the problem of name homonymy in a set of web pages, there is a previous challenge still to solve before studying whether it is better to generate one cluster with all the individuals that have nothing in common with the others, rather than creating one cluster for each of them.

This challenge has a greater impact in the success of the experiments, and it is related to the extraction of relevant information: given a web page whose structure is unknown, it is essential to determine which contents in this page are useful for the resolution of the problem (i.e., are relevant for describing the person to whom the page is assumed to refer) and which contents should be filtered out (such as navigational information, titles, headers and footers, disclaimers, etc.). And moreover, regarding the first ones, it is necessary to determine if those contents refer exactly to the expected person or else is referring to other different entities or concepts (such as lists of similar authors, for instance). In our humble opinion, these should be the challenges on which we should focus in future editions of WebPS.

Acknowledgements

This work has been partially supported by the Spanish Center for Industry Technological Development (CDTI, Ministry of Industry, Tourism and Trade), through the CONTENIDOS A LA CARTA Project, INGENIO 2010 Programme, AVANZA I+D 2008. Other partners in the project are Agencia EFE, Germinus XXI, 11870.com and Universidad Politécnica de Madrid.

References

1. Overview of the WebPS 3 task at ImageCLEF 2010. Working Notes of CLEF 2010. Padova. Italy. 2010.

2. Artiles, J.; Gonzalo, J. and Sekine, S. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. 2009.
3. Park. Hae-sang; Lee. Jong-seok; Jun. Chi-hyuck. A K-means-like Algorithm for K-medoids Clustering and Its Performance. Proceedings of the 36th CIE Conference on Computers & Industrial Engineering. pp.1222-1231. Taipei. Taiwan. Jun. 20-23 (2006).
4. University of Neuchatel. IR Multilingual Resources at UniNE. http://members.unine.ch/jacques.savoy/clef/index.html
5. Porter. M. Snowball stemmers and resources page. http://www.snowball.tartarus.org