Automatic Hierarchical Clustering of Web Pages

Ricardo Campos
Centre of Human Language Technology and Bioinformatics
University of Beira Interior
+351 275 319 891
ricardo.campos@ipt.pt

Gaël Dias
Centre of Human Language Technology and Bioinformatics
University of Beira Interior
+351 275 319 891
ddg@di.ubi.pt

ABSTRACT
In this paper, we propose a system that clusters web pages and presents them as a hierarchical structure instead of a classical search ordered list retrieved from any search engine. The organization of the results based on this concept makes easier the user's search navigation between the results of the search engine. In particular, we use web content mining techniques to represent texts, based on their most relevant terms, which can be simple words or phrases. A soft clustering algorithm is then applied to group documents into clusters hierarchically linked. Finally, each cluster is labelled with its most relevant term based on a simple heuristics.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering.*

General Terms
Algorithms, Measurement, Experimentation.

Keywords
Web page automatic clustering, Hierarchical soft clustering, Web content mining.

1. INTRODUCTION
The World Wide Web has become a huge network of information and search engines actually deal with the problem of retrieving relevant documents. One of their main problems is that the induced relevance may not satisfy the user intents. This is manly due to two problems: (1) search engines interpret the content of documents in a basic way and (2) they present the retrieved information in an unstructured way. In fact, systems are not capable of understanding completely what users are looking for due to small queries and on the other side, they keep retrieving a huge set of unstructured information.
To answer these problems, we propose a meta-search engine named WISE that uses web content mining techniques introduced by [1] associated to a soft clustering algorithm called PoBoc [2] to find, analyze, understand, disambiguate and organize the set of documents returned by any search engine for a given query. As a consequence, the user's search is simplified turning his task less time-consuming.

ELECTRA Workshop -Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications Salvador, Brazil, August 19, 2005, in association with SIGIR 2005

2. RELATED WORK
The Information Retrieval community has suggested in scientific published literature different solutions to the problem of organizing web search results. But all these works [3] [4] [5] [6] [8] [9] [10] have in common the fact that they mainly consider the titles and the snippets of each document retrieved from a search query. However, [3] refer that the results are obviously inferior when compared to the use of overall text. For that purpose, [4] enriches[1] the snippets with two existing knowledge bases. All these works also use lists of stop-words and stemming algorithms which make their solution language-dependent.
In order to represent the documents retrieved by the search engines, [4] [5] [6] use the well-known space vector model [7] considering only the snippets and not the overall text. Some other works show another text representation. [2] [8] [9] [10] use the concept of shared n-grams between snippets. But none uses web content mining techniques to represent their documents as we will show in the next section.
Once documents are represented into a given structure, clustering techniques must be applied to produce a structured list of results. The proposed algorithms proposed so far in the literature distinguish themselves by the use of (1) simple words or phrases and (2) by implementing flat clustering or hierarchical clustering. In particular, the work done by [11] which was not tested in web environment and [3] propose flat clustering with simple words, while [8] and [10] do it with phrases. [9] are the first to introduce hierarchical clustering with phrases, followed by [4] and [5]. [6] also propose hierarchical clustering but just considering simple words. In our work, we use a soft clustering algorithm called PoBoc [2] that has shown successful results within the analysis of textual data and allows words to belong to different clusters. It is used in association with phrases that are extracted previously from texts based on the SENTA software [12].
Our solution differs from all previous work as it proposes a deep analysis of text content using a multiword extractor that produces relevant phrases. Compared to existing methodologies that elect frequent strings as phrases, we use a more sophisticated language-independent phrase extractor [12]. Based on the extraction of these phrases, we then apply web content mining techniques to extract as deep knowledge as possible from texts to finally

[1] To our best knowledge, they are the only ones.

produce a set of soft clusters based on a soft clustering algorithm called PoBoc [2].
In this paper, we will first show the architecture of our solution. In a second part, we will show some first results and finally draw some conclusions.

3. GLOBAL ARCHITECTURE

Our architecture called WISE is composed of 4 main parts:

(1) The selection of relevant pages from the set of all retrieved documents by the search engine;
(2) The integration of the SENTA software [12] that extracts phrases from raw texts;
(3) The detection of relevant terms that characterize the document, using the WEBSPY software [1] that implements the web content mining techniques;
(4) The presentation of the documents into a hierarchical structure using the PoBoc [2] algorithm.

The overall architecture can be seen in Figure 1.

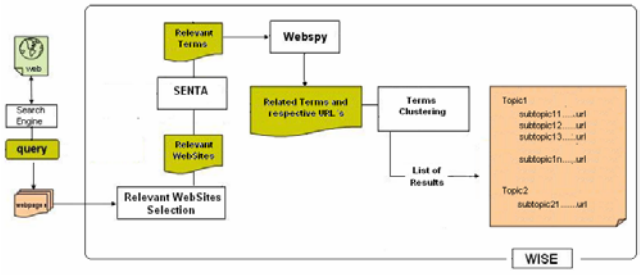

Figure 1. The architecture of WISE.

Our algorithm proposes a new method for extracting relevant pages and a new representation of documents, based on relevant terms to apply a clustering algorithm. Our algorithm follows the next 8 steps:

(1) Retrieve the list of results of the search engine for a given query (meta-crawler);
(2) Select the most important results from all the retrieved documents. All the literature specified above treats each document as equals, but they are not. Each one has different relevance to the query, which decreases as more and more documents are retrieved. Producing clusters in many documents of little relevance can reduce the quality of the results, and for that reason we exclude some results, benefiting precision upon recall. For that purpose, we applied a function that chooses from the returned documents, the best ones i.e. the ones that overpass a given threshold calculated by Equation 1.

$$average\_relevance = \frac{\#\,\text{returned URLs}}{\#\,\text{different absolute URLs}} \qquad (1)$$

Based on the above equation, we select all the relevant URLs which number of occurrences is greater than the calculated threshold. But, we also take as relevant all the retrieved absolute URLs.
In order to better understand our procedure, the number of occurrences of a URL is the sum of all URLs that share the same absolute URL. For instance, considering Figure 2, the average relevance threshold would be 4/3 = 1.3. As a consequence, we would not consider the webpage which absolute URL is http://geocities.com being its number of occurrences below the threshold.

Site	N.º of Occurrences
http://www.vodafone.com	2
http://www.vodafone.com/news/01.html	2
http://geocities/mobiles/test.html	1
http://www.manUnited.uk	1

Figure 2. List of results related to mobile phones query.

In addition, we extend the number of relevant pages given by the search engine by considering a set of pages not caught by the system, but related with the query. For that purpose, for any absolute retrieved URL, we catch its N best pages re-running the search engine over the absolute URL with the same query.
(3) Identify phrases in the documents in order to increase the knowledge about each document by using the SENTA software [12];
(4) Calculate the set of relevant terms to the query for each document by applying the WEBSPY software [1] that implements a set of decision trees C5.0 based on 12 characteristics between all the words/phrases in the document and the query terms. This step retrieves a set of related terms with a probability of relevance. Our purpose is to use the overall text and all the relations between words/phrases and the query term to represent as best as possible the semantic content of each text;
(5) Calculate the similarity matrix. In order to prepare the clustering step, the WEBSPY software is applied again. For each relevant term retrieved from step (4) we apply WEBSPY based on the pages where relevant term occurs. As a consequence, each relevant word/phrase is also represented by a set of related words with a given relevance probability. In order to build the similarity matrix, we then apply the Cosine measure between all the pairs of relevant word/phrase.
(6) Group into a hierarchical structure the set of all relevant documents retrieved in step (2). By this clustering step, we aim at disambiguating the sense of the query term. Thus, the user is helped in his search for information. This step is done using the PoBoc algorithm [1], a soft clustering algorithm that allows a word/phrase to belong to different cluster. This characteristic is fundamental for text analysis as it is obvious that one word may appear in different contexts with a different semantic content.
(7) Label each cluster with its most relevant word/phrase based on a simple heuristic that chooses the word/phrase that occurs more often in the set of words/phrases representing relevant terms and taking into account the sum of probabilities in case of ties.
(8) Present the final results to the user.

Our solution uses the overall text information, not only the titles and the snippets, turning the solution more robust in terms of semantic ambiguity. Also, we do not use lists of stop-words neither we use stemming algorithms, which makes our solution flexible, and domain/language-independent. To our best knowledge, we are the first using web content mining techniques to understand documents, using afterwards this knowledge as a base to form structured hierarchical soft clusters.

4. RESULTS

The results shown below are clusters returned by our system WISE from the system query execution *Benfica* on may 31st, 2005 using Google™ as search engine. The cluster that can be seen in Figure 3 refers to the set of URLs related to *José-António-*

Camacho[2], the former football coach of *Real-Madrid*, spoken at that time to be a possible successor of *Giovanni-Trapattoni* as the new coach of *Benfica*. We can notice that the labels show some degree of quality and semantic description of the content of the cluster due to the identification of relevant phrases. One other interesting issue is the capacity of the system to deal with word mistakes (*Giovanni*, not *Geovanni*). This result is due to the fact that no restriction is made over term frequency.

Figure 3. Monothetic labels and phrases[3].

In Figure 4, we can see the ability of our system to deal with term disambiguation. As we have already seen, *José-António-Camacho* is related to *Benfica* (i.e., *Benfica* is related to Benfica Football Club), but the system also retrieves a cluster with label *PS* which refers to the politic socialist party located in the *Benfica* neighborhood (i.e. *Benfica* is related to politics through the fact that *Benfica* is also related to a famous neighborhood of Lisbon). Moreover, the label *Universitários* refers to student life like housing, transports and roads (i.e. *Benfica* is also related to a privileged neighborhood for student housing).

Figure 4. Word sense disambiguation.

5. CONCLUSION

Our paper proposes to organize the flat ranked search lists returned by current search engines to produce a topic hierarchical structure that will help the user in his search for information. Our main contribution to the field is the use of web content mining techniques applied to the overall information within texts which allows deep semantic analysis of web documents, understanding facts that, till now, no search engine understands. Moreover, the identification of phrases to define key concepts in texts allows a greater document content understanding. The architecture and the proposed algorithms are the answer to one of the biggest problems of search engines: returning quality results through an organized and disambiguated structure of concepts. This paper shows an on-going work and further improvements will be made, especially in terms of merging clusters as data sparseness usually produce too many clusters. The WISE software will be soon freely available at http://wise.di.ubi.pt under GPL license.

[2] *José-António-Camacho* stands for the name *José António Camacho* that has been recognized by SENTA as a phrase.

[3] The fact that we only have one absolute URL for the concept *José-António-Camacho* is casual. In fact, our system is capable of retrieving different absolute URLs for the same concept.

6. REFERENCES

[1] Veiga, H., Madeira, S. and Dias, G. 2004. *Webspy*. Technical Report no 1/2004. http://webspy.di.ubi.pt

[2] Cleuziou, G., Martin, L. and Vrain, C. 2003. PoBOC: an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. In Proceedings of the 16th biennial European Conference on Artificial Intelligence (ECAI'04), Valencia, Spain, August, 440-444.

[3] Jiang, Z., Joshi, A., Krishnapuram, R. and Yi, L. 2002. Retriever: Improving web search engine results using clustering. In Managing Business with Electronic Commerce.

[4] Ferragina, P. and Gulli, A. 2005. A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering. In the Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, ISBN: 1-59593-051-5. 801-810.

[5] Martins, B. and Silva, M. 2003. Web Information Retrieval with Result Set Clustering. In Proceedings of NLTR 2003 - Natural Language and Text Retrieval Workshop associated to EPIA'03, December.

[6] Fung, B., Wang, K. & Ester, M. (2003). *Large hierarchical document clustering using frequent itemsets*. In Proceedings of the SIAM International Conference on Data Mining, Cathedral Hill Hotel, San Francisco, CA, May 1-3.

[7] Salton, G., Yang, C.S., and Yu, C.T. 1975. A theory of term importance in automatic text analysis. American Society of Information Science 26, 1, 33-44.

[8] Zamir, O. and Etzioni, O. 1998. Web Document Clustering: A Feasibility Demonstration. In Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR'98), 46-54.

[9] Zhang, D. and Dong, Y. 2001. Semantic, Hierarchical, Online Clustering of Web Search Results. In Proceedings of the 6th Asia Pacific Web Conference (APWEB), Hangzhou, China, April.

[10] Zeng, H., He, Q., Chen, Z. and Ma, W. 2004. Learning to cluster web search results. In the Proceedings of the 27th annual international conference on Research and development in information retrieval, Sheffield, UK, ISBN: 1-58113-881-4, 210-217.

[11] Hearst, M. and Pedersen, J. 1996. Re-examining the Cluster Hypothesis: Scatter/ Gather on Retrieval Results. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), Zurich, Switzerland, August, 76-84.

[12] Dias, G. (2002). Extraction Automatique d'Associations Lexicales à partir de Corpora. PhD Thesis. DI/FCT New University of Lisbon (Portugal) and LIFO University of Orléans (France).