Automatic Hierarchical Clustering of Web Pages

Ricardo Campos (ricardo.campos@ipt.pt) and Gaël Dias (ddg@di.ubi.pt)

Centre of Human Language Technology and Bioinformatics , University of Beira Interior, Covilhã - Portugal

Abstract

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.

Introduction

The World Wide Web has become a huge network of information and search engines actually deal with the problem of retrieving and organizing 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, i.e., they do not interpret de documents taking in account language ambiguity and text context;
(2) they present the retrieved information in an unstructured way.

Purposes

Query: Benfica
Football
Nuno Gomes
www.ojogo.pt
www.abola.pt
Eusébio
www.slbenfica.pt
Houses
T1
www.era.pt

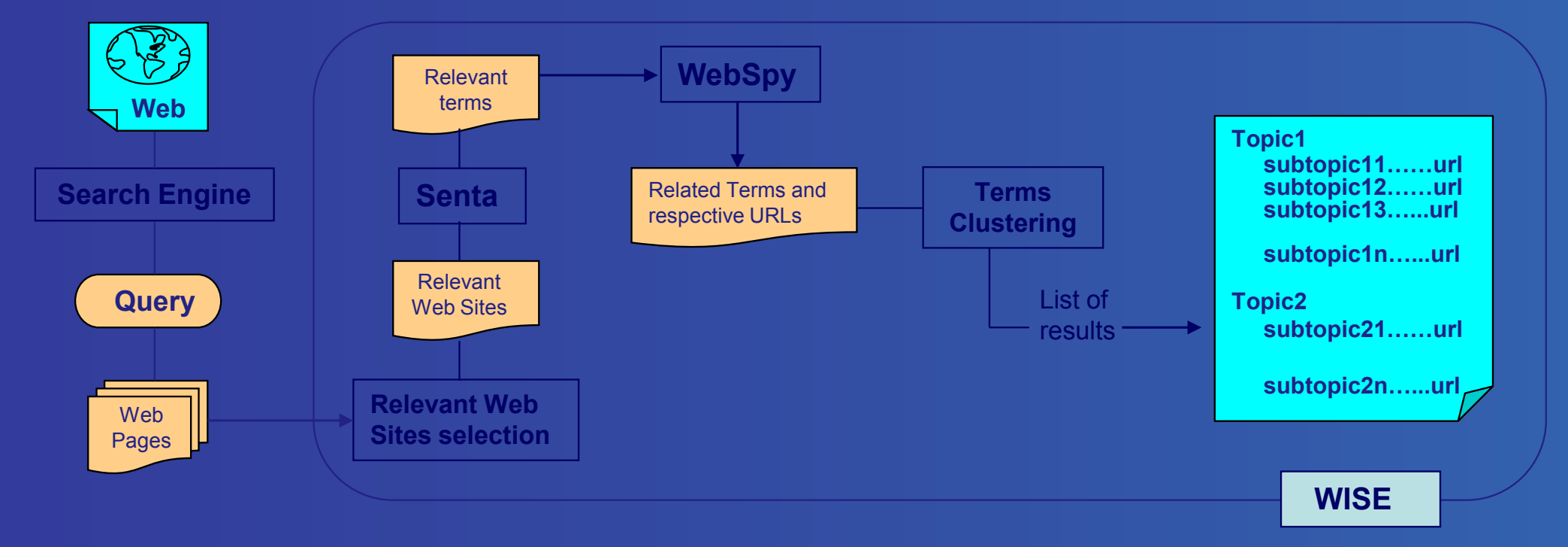
To answer these problems, we propose a meta-search engine named WISE [4] to find, analyze, understand, disambiguate and organize the set of documents returned by any search engine for a given query. The documents should be grouped in clusters, and the information presented hierarchically and concept disambiguated.
As a consequence, the user's search is simplified turning his task automatic and less time-consuming.

Global Architecture

Our architecture called WISE [4] 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 [2] that extracts phrases from raw texts;
(3) The detection of relevant terms that characterize the document, using the WEBSPY software [3] that implements the web content mining techniques;
(4) The presentation of the documents into a hierarchical structure using the PoBoc [1] algorithm.

The overall architecture can be seen in next figure:



The different literature studied in the scope of this work do not use any semantic measure to understand the meaning of each text. This is the reason why clusters potentially not related can be produced.
To go beyond these problems, we developed a flexible architecture for the hierarchical clustering of web pages. In particular, we developed a new method for extracting web relevant pages and a new representation of documents based in a set of web content mining techniques, so the soft clustering algorithm can be applied. Our algorithm follows the next 8 steps:

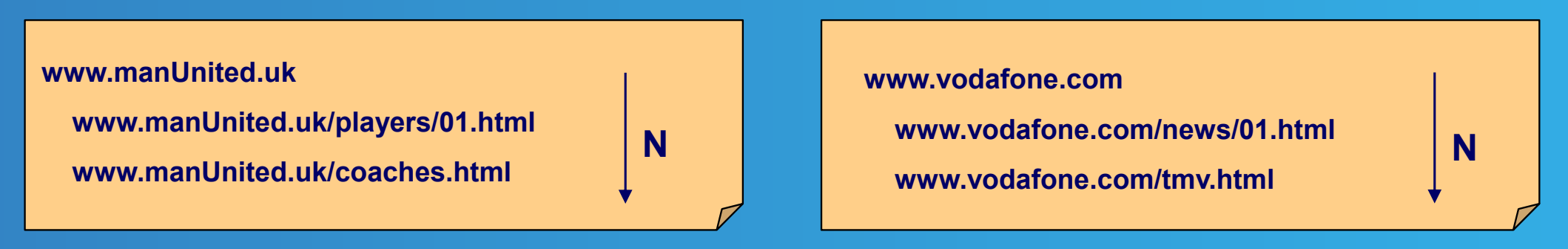
(1) Retrieve the list of results of the search engine for a given query (meta-crawler);
(2) Select the most important results from all the retrieved documents. Producing clusters in many documents of little relevance can reduce the quality of the results, and for that reason we exclude some results, 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 next equation:

$$average\_relevance = \frac{\#\,\text{returned URLs}}{\#\,\text{different absolute URLs}}$$

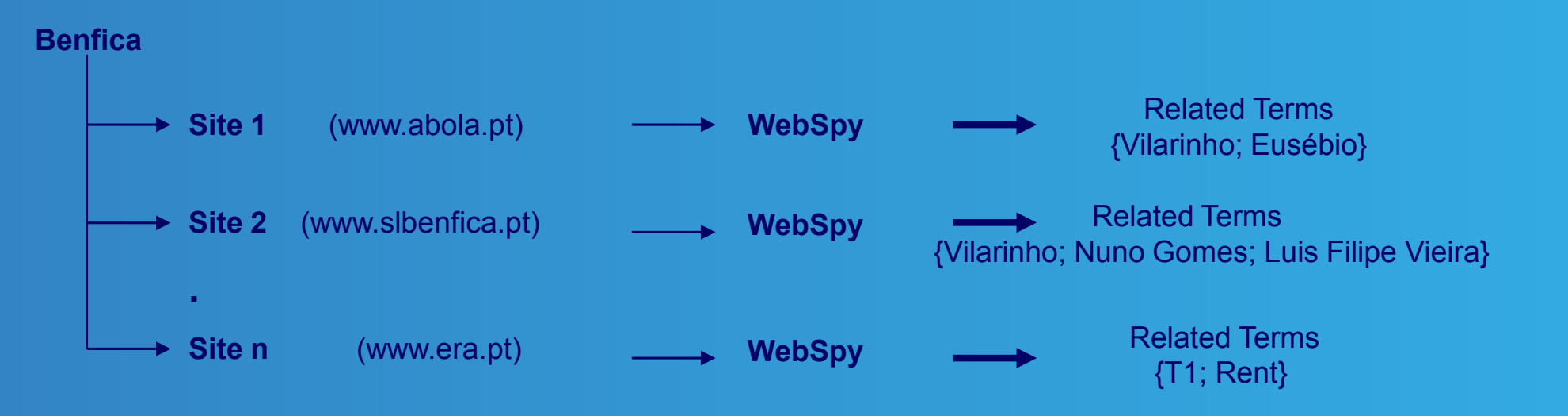
Based on the above equation, we select all the relevant URLs which number of occurrences is greater than the calculated threshold. But, we also take as relevant all the retrieved absolute URLs.
In order to better understand our procedure, the number of occurrences of a URL is the sum of all URLs that share the same absolute URL. For instance, considering next figure, the average relevance threshold would be 4/3 = 1.3.

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

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 (see next figure), we catch its N best pages re-running the search engine over the absolute URL with the same query.


www.manUnited.uk
www.manUnited.uk/players/01.html
www.manUnited.uk/coaches.html
N
www.vodafone.com
www.vodafone.com/news/01.html
www.vodafone.com/tmv.html
N

(3) Identify phrases in the documents in order to increase the knowledge about each document by using the SENTA software [2].;
(4) Calculate the set of relevant terms to the query for each document by applying the WEBSPY software [3] 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 (see next figure). 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;



In the previous step we aim flat clustering. With so, the user feels lost in the middle of so many information. The set of results, shares different concepts, but disposed all together (ambiguity problem unsolved). To aim hierarchical clustering, WISE [4] uses again WebSpy: for each relevant term retrieved from step 4, we apply WebSpy based on the pages where relevant terms occurs. As a consequence, each relevant word/phrase is also represented by a set of related words with a given relevance probability (see figure 5).

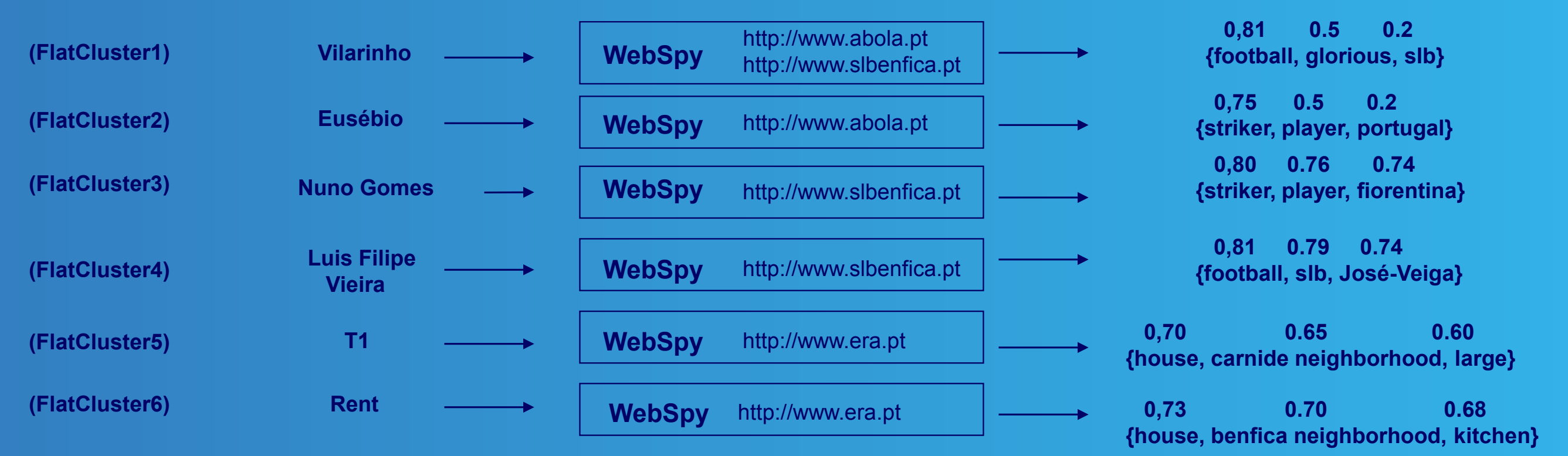


Figure 5: Flat Clustering

(5) The next step is to calculate the similarity between all the flat clusters. The value of each similarity is registered in a matrix (see next figure):

	FlatCluster1	FlatCluster2	FlatCluster3	FlatCluster4	FlatCluster5	FlatCluster6
FlatCluster1		Sim(FC1,FC2)	Sim(FC1,FC3)	Sim(FC1,FC4)	Sim(FC1,FC5)	Sim(FC1,FC6)
FlatCluster2			Sim(FC2,FC3)	Sim(FC2,FC4)	Sim(FC2,FC5)	Sim(FC2,FC6)
FlatCluster3				Sim(FC3,FC4)	Sim(FC3,FC5)	Sim(FC3,FC6)
FlatCluster4					Sim(FC4,FC5)	Sim(FC4,FC6)
FlatCluster5						Sim(FC5,FC6)
FlatCluster6						

Each similarity is calculated under de Cosine measure (see, next equation, when the vectors share equal terms, otherwise the value is zero), which measures the distance between 2 vectors. Those vectors are formed by the related terms and probabilities, obtained in figure 5.

$$Co\sin e(vector_1, vector_2) = \frac{\sum_{i=1}^{n} prob(vector_{1,i}) \times prob(vector_{2,i})}{\sqrt{\sum_{i=1}^{n} prob(vector_{1,i})^2} \times \sqrt{\sum_{i=1}^{n} prob(vector_{2,i})^2}}$$

(6) Group into a hierarchical structure (see next figure) the set of all relevant documents retrieved in step (2). By this clustering step, we aim at disambiguating the sense of the query term. This is done using the Poboc algorithm [1], a soft clustering algorithm that allows a word/phrase to belong to different clusters unknown apriori (its obvious that one word may appear in different contexts with a different semantic content).

Cluster 1

Vilarinho	www.abola.pt
	www.slbenfica.pt
	http://....../
Eusébio	www.abola.pt
	http://....../
Nuno Gomes	www.slbenfica.pt
	http://....../
Luis Filipe Vieira	www.slbenfica.pt
	http://....../

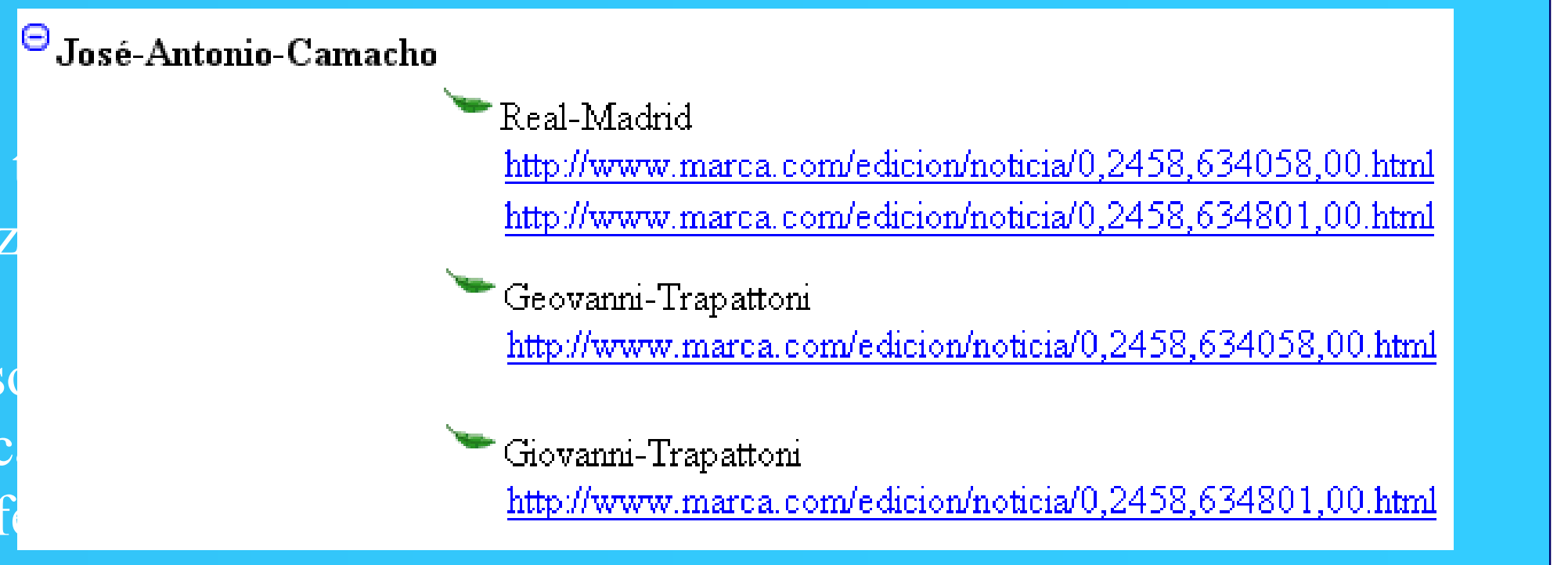
Cluster 2

T1	www.era.pt
	http://....../
Rent	www.era.pt
	http://....../

(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 word/phrases representing relevant terms and taking in account the sum of probabilities in case of ties. Taking as a reference the given example, we will have for Cluster 1 – Football and for Cluster 2 – House;
(8) Present the final results to the user.

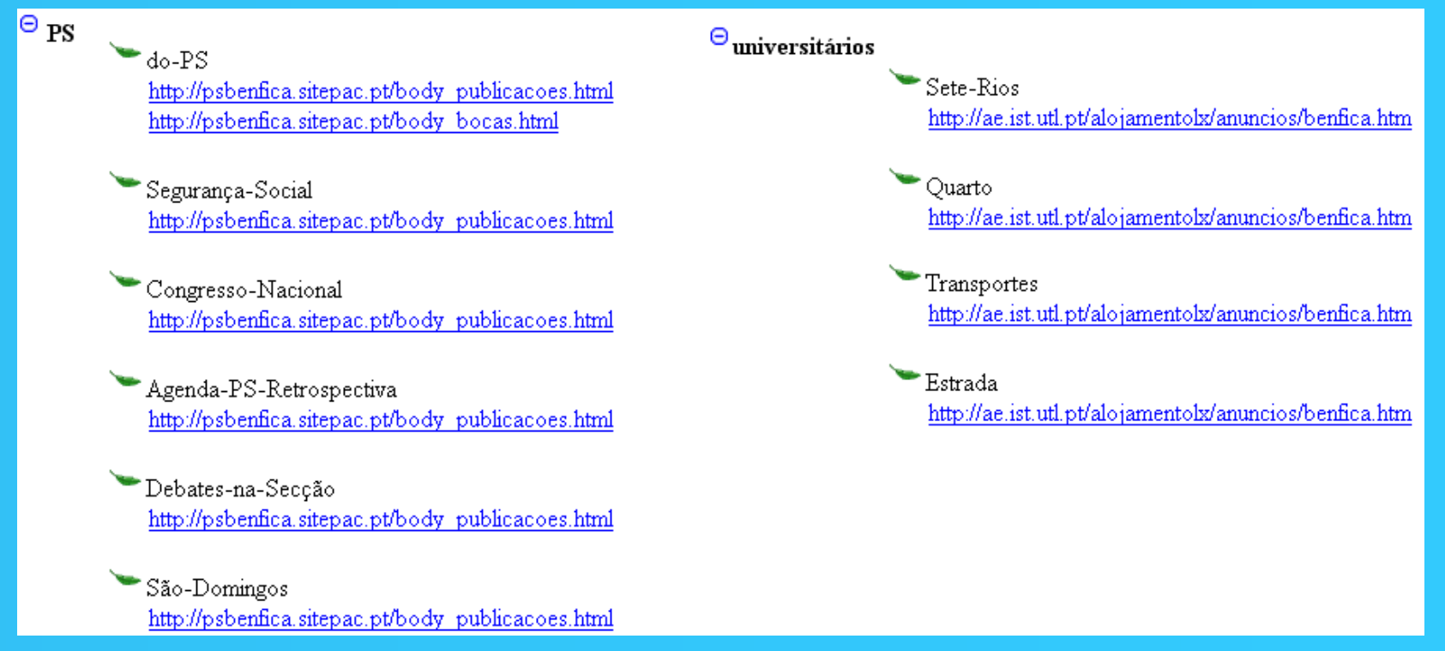
Results

The results shown below are clusters returned by our system WISE [4] from the system query execution *Benfica* on May 31st, 2005 using Google as search engine. The cluster that can be seen in figure 8 refers to the set of URLs related to *José-António-Camacho* 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.


José-Antonio-Camacho
Real-Madrid
http://www.marca.com/edicion/noticia/0,2458,634058,00.html
http://www.marca.com/edicion/noticia/0,2458,634801,00.html
Geovanni-Trapattoni
http://www.marca.com/edicion/noticia/0,2458,634058,00.html
Giovanni-Trapattoni
http://www.marca.com/edicion/noticia/0,2458,634801,00.html

Note: *José-António-Camacho* stands for
António Camacho that has been recogniz
as a phrase
Note: the fact that we have only on abso
the concept *José-António-Camacho* is c
our system is capable of retrieving diff
URls for the same concept.

In next figure, 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).


PS
do-PS
http://psbenfica.sitepac.pt/body_publicacoes.html
http://psbenfica.sitepac.pt/body_bocas.html
Segurança-Social
http://psbenfica.sitepac.pt/body_publicacoes.html
Congresso-Nacional
http://psbenfica.sitepac.pt/body_publicacoes.html
Agenda-PS-Retrospectiva
http://psbenfica.sitepac.pt/body_publicacoes.html
Debates-na-Secção
http://psbenfica.sitepac.pt/body_publicacoes.html
São-Domingos
http://psbenfica.sitepac.pt/body_publicacoes.html
universitários
Sete-Rios
http://ae.ist.utl.pt/alojamentolx/anuncios/benfica.htm
Quarto
http://ae.ist.utl.pt/alojamentolx/anuncios/benfica.htm
Transportes
http://ae.ist.utl.pt/alojamentolx/anuncios/benfica.htm
Estrada
http://ae.ist.utl.pt/alojamentolx/anuncios/benfica.htm

Conclusion

The solution here described allows showing to the user, the information more important in what concerns t a given query, also organized in a structured manner.

To aim these purposes we use:

- an algorithm that ignores non relevant documents and another that increase them;
- phrases to define concepts, with consequences in documents understanding;
- a set of web content mining techniques that allow the extraction of semantic terms, related with the document and the query, understanding facts, that none till now tries to understand;
- document clustering algorithm to present the information in a structured organized and hierarchically manner.

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, maintaining our system flexible and domain/language independent. The architecture and the solution proposed are the solution to one of the biggest problems that search engines actually deal with:

quality search results retrieved, through a flexible structure, automatic, organized and concept disambiguated.

References

[1] Cleuziou, G., Martin, L. and Vrain, C. 2003. PoBOC: na 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.

[2] 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).

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

[4] http://wise.di.ubi.pt