

# A Multi Criteria Function to Concept Extraction in HTML Environment

A. Ribeiro

Industrial Automation Institute (IAI)  
Spanish Council for Scientific Research  
Arganda del Rey  
28500 Madrid. Spain.

V. Fresno

Industrial Automation Institute (IAI)  
Spanish Council for Scientific Research  
Arganda del Rey  
28500 Madrid. Spain.

**Abstract** - The core of Internet and the World Wide Web revolution is the capacity to efficiently share the huge quantity of data. But the rapid and chaotic growth of the Net has extremely complicated the task of share or mining useful information. Each inference process, from Internet information, requires an adequate characterization of the Web pages. The textual part of a page is one of the most important aspects that should be considered to appropriately perform a page characterization. The textual characterization should be made through the extraction of an appropriate set of relevant concepts that represent properly the included text in the Web page. This paper presents a method, essentially based on the extraction of characteristics in the HTML language, to obtain a set of relevant concepts from a Web page. In addition, to prove the validity of the proposed approach a comparative study is shown. It exhibits a higher quality in the representations generated by the proposed method versus a commercial tool.

**Keywords:** Concept extraction, Feature vector in HTML texts, Web characterization. Web page representation.

## 1. Introduction

The classic techniques of Information Retrieval (IR) are habitually used to obtain information from the Web [1]. When these techniques are applied, the problems not only appear due to the enormous number of pages or

the continuous changes in them; but also because Web users are significantly different from the groups that traditionally have used the IR techniques. In addition, there are not standards or style rules in the Web; the page contents are created by a set of very disparate people and in an autonomous way. On the other hand, the inherited IR technology has slowly progressed to take into account the Web necessities. Consequently, the search engines show little scope and a poor accuracy, in other words they recover a small fraction of the total of existent documents and for this fraction only a small portion is significant.

From a general point of view and for any document type, the first step in tasks such as automatic summarization, text classification, information retrieval, information extraction, or text mining is to obtain a data structure for digital processing that represents the text. But, the access to the text information is difficult, since the relationship among the form (usually a sequence of characters) and the meaning is not as clear as in the case of numeric data. However, the results of this first stage are essential since the success of the analysis task is strongly dependent on having an appropriate text representation that captures the most relevant aspects of the document. In fact, without a proper set of features a classifier will not be able to accurately discriminate between different categories.

In few words, texts have been traditionally represented in a vector space model [2] [3] using basically two different methods. The first one represents a document by a vector of binary

attributes indicating if a word occurs or does not occur in the document. In this case the frequency of a word in a document is not captured. The second model represents a document by the set of occurrences of each word in the document. In both cases the order of the words in the document is lost.

Both representations generate vectors with a very high dimensionality (of  $10^4$  to  $10^7$  components) that hinders in many cases the use of knowledge extraction algorithms [4]. This paper presents an approach that allows obtaining a Web page representation with the aim of carrying out classification and mining tasks later on. The approach takes advantage of some Web page characteristics to obtain a significantly reduced vector that contains the most representative words of the Web document as well as an associate number to each word that characterizes the relevance of this word into the document. Additionally, to evaluate the effectiveness of the proposed method a comparative study versus a commercial tool is presented.

## 2. Web page attributes and characterization functions

In the World Wide Web, the HTML language supports documents. Every Web document is built as a combination of tags and text information that Web browsers recognize and visualize.

There are many types of Web tags [5], such as links to other pages, references to images or files and textual attributes. These textual tags are used to assign special properties to the text, therefore if fragments of text are established between two respective tags (for instance `<b>` and `</b>`) the portion of text will assume that tag. With tags, users can indicate which words belong to the Web page title, body, font style, headings, and many other attributes for the Web page. The textual tags or attributes will be the core of the method presented in this paper in contrast with other approaches that consider only *meta* tags [6] and the textual content as a plain text. To consider only *meta* tags might cause irrelevant results when the Web page do not contain this kind of information.

In the proposed method some textual tags are selected to represent the Web page through the words in the textual part of the page and an assigned weight to each word, that computes the relevance of the word in the text.

Among all pages attributes that apparently have information for computing the word significance in the text, the most promising are:

1. Tags that indicates the page title (`<title>...</title>`).
2. Tags like `<b>...</b>`, `<u>...</u>`, `<em>...</em>`, `<i>...</i>`, and `<strong>...</strong>` that allow emphasis to parts of the text and to distinguish these parts from the rest.

It seems obvious that if a word belongs to the page title, this characteristic should be considered when the relevance of the word in the document is computed (the weight component). The same consideration holds for the emphasized sentences in the text. However there is an essential difference between one case and another, while to emphasize is an operation consciously made by the author when he is designing the Web page, the title content could be the result of some automatic process and, for this reason, in some cases not relevant. This fact has been verified in many Web pages [7].

In addition to these two attributes there are other more "classical issues" that could be considered to compute the word relevance in a document: the *word position* into the text and the *word frequency* in the text. The word position is a possible criterion to estimate the relevance of a word in a text, because the users tend to structure their texts in three parts: introduction, body and conclusions. Of course, not all the pages have this structure. In some cases, irrelevant results could be reached, but this also holds for the rest of criteria. This only reinforces the thesis that the analysis of multiple criteria is the most appropriate solution in a heterogeneous domain as the Net.

Other attributes like the *meta* tags could be considered, but right now are not enough used and we have decided to obviate them in this first stage.

At this point it is possible to define a set of functions that evaluates the word relevance through the previously described aspects.

When only the frequency of a word is taken into account, it can be defined the following function.

$$P_f(i) = n_f(i)/N_{tot} \quad (1)$$

Where  $n_f(i)$  is the frequency of the word  $i$  in the text and  $N_{tot}$  is the number of words in the page.

In a similar way can be defined a function related to the number of occurrences of a word in the title.

$$P_t(i) = n_t(i)/N_{tit} \quad (2)$$

Where  $n_t(i)$  are occurrences of the word  $i$  in the title and  $N_{tit}$  the number of words in the title.

If the emphasis of a word is observed, the relevance for this word can be evaluated as:

$$P_e(i) = n_e(i)/N_{enf} \quad (3)$$

In this case  $n_e(i)$  represents times that the word  $i$  is emphasized and  $N_{enf}$  the number of word that are emphasized.

Finally, to compute the relevance of a word from the position aspect, the Web page is split in four parts and the following function is defined:

$$P_p(i) = \frac{3/4 * n_{1,4}(i) + 1/4 * n_{2,3}(i)}{N_{tot}} \quad (4)$$

Where  $n_{1,4}(i)$  are occurrences of the word  $i$  in the first and the fourth quarter of the page and  $n_{2,3}(i)$  are occurrences of the same word in the second and third quarter of the page. Notice that the first and the last positions have more weight than the intermediate ones, that is because, as we said previously, people tend to begin and to finish a text with the more important subjects.

Hereafter these four functions will be called characterization functions.

The question now is how to combine the characterization functions to reduce the dimensionality of the feature vector increasing the quality of the representation. We assume for each word that the relevance is a lineal combination of the previous functions, that is:

$$S(i) = C_1 P_f(i) + C_2 P_t(i) + C_3 P_p(i) + C_4 P_e(i) \quad (5)$$

A statistical analysis was been accomplished with the intention of knowing the most adequate coefficient set. The reduction of the dimensionality of the feature vector was studied for each characterization function. Main results of this statistical study can be summarized along these lines. The frequency and the position have a similar behavior and consequently, they should have similar contributions in the combination of the characteristic function. On the other hand, frequency and position are attributes presents in every Web page, consequently the combined contribution should be higher than 50%. Additionally, it seems that a good combination for the characterization functions should give more weight to those criteria that generate a greater reduction in the feature vector; this is the case for the emphasis and the title criteria. But this is only partially true, because in many cases, although a title exists for the Web page, this does not make sense or is not relevant of the text content. The reader can find more and important aspects of this study in [7].

### 3. Evaluation of the concept extraction system

To evaluate the proposed approach a software system has been developed that generated a concept vector from an HTML file. To accomplish that, the system first removes the HTML tags and some important information, as which word is empathized, is saved. Second, the stop-words are eliminated from the page text, where the stop-words are the words without relevance for our purpose. Articles, prepositions, conjunctions are stop-words and represent 38.8% of the word total in a Web page. Lastly the line number where the word appeared, the position inside the line and

the frequency are taken to compute the value of each characterization function for each word in the current page. The coefficients used in the relevance function (5) are as follows:

$C_1$ (Frequency)	0.30
$C_2$ (Title)	0.15
$C_3$ (Position)	0.30
$C_4$ (Emphasis)	0.25

As a result an ordered bi-dimensional vector is obtained from the HTML text.

To perform the evaluation process a commercial product has been selected [8]. Copernic Summarizer is text-summarizing software that also has concept extraction capabilities. The comparative study has been performed from different sample sets with the intention of covering totally the heterogeneous nature of the World Wide Web. Sample sets have been composed by Web pages that have been randomly selected without a specific topic and with several sizes. The idea is to compare the behavior of both software systems in terms of size and heterogeneity versus homogeneity. Related to the size, three sample sets have been built: 1) pages with less than 50 words; 2) pages with more than 50 words but less than 250 words and 3) pages with more than 250 words. In reference to the heterogeneous and homogeneous content of the page, two sets of samples have been constructed. Sites of newspapers, of buy/sale, or of auctions have been considered as heterogeneous pages. To sum up, Web pages that contains several topics in opposition to Web pages that concern to one topic; these last ones have been considered as homogeneous pages.

In this case the discrimination capability of the concept vector has not been established by applying, for instance, a classification method later on. A good concept vector, from a qualitative point of view, has been fixed by a human expert analysis. In others words, a human expert has performed the evaluation of both applications results. Therefore the evaluation has a qualitative estimation, whereas it has been indispensable to estimate if a word had enough discrimination capability for classification and mining tasks. Consequently to perform this qualitative analysis, before

obtaining the feature vectors of each page, an expert classifies all the words in the sample set in the following categories:

- A: Category formed by the words, which clearly differentiate some specific topic.
- B: Category shaped by ambiguous words in the sense that they can characterize different topics.
- C: Category composed by words that do not belong to any topic.

Additionally, to accomplish a more precise comparison, the developed system (IAI-extractor) generates concepts integrated by more than a word (two and three words), as Copernic Summarize offers this possibility. The combinations of words are built from adjacent words; this is the reason why the position inside the line is saved. The new term weight is computed from the weight combination of the words that form the term. This combination of words allows obtaining very descriptive concepts such as "data mining", "artificial intelligence", etc.

In figure 1 the charts achieved in the comparative analysis are shown. In both applications the dimension of the feature vector was fixed to 20 words. For IAI-extractor, the feature vector was composed by the weightiest concepts: ten concepts of the first level (one word), five concepts of the second level (two words), and five concepts of the third level (three words). All the shown results are the average of those obtained for each page of the sample set. The vertical columns of the charts indicate the percent of words, in the feature vector, that belong to each category. The white columns represent the results of the IAI-extractor and the gray columns the behavior of the Copernic Summarize. As can be seen, the IAI-extractor improves the results achieved by Copernic Summarize, because the feature vector produced by IAI-extractor always has more concepts classified in the A category than the feature vector generated by Copernic Summarize.

## 4. Conclusions and Future work

This paper presents an approach for concept extraction from Web pages. The proposed method generated for a Web page a bi-dimensional vector, where the first component is a word extracted from the page content and the second one is a number, which evaluates how well the word represents the text in the Web page. To calculate this second component the employed function takes into account some characteristics of the HTML text. A clear improvement of the proposed representation versus a commercial software product has been showed through a qualitative analysis performed by a human expert. With the aim of achieving this analysis three word categories, that estimate the relevance of a word in a given Web page, have been defined.

Future work will be conducted in four different ways. First, a classification algorithm will be developed based on this concept vector. Second, the sample set will be increased. Third, a learning algorithm will be applied in order to find the optimal coefficients  $C_1$ ,  $C_2$ ,  $C_3$  and  $C_4$ . Last, others criteria such as *meta* tags will be included in the proposed relevance function.

## 5. Acknowledge

Present work is fully supported by Innovatec S.A.

## 6. References

- [1] V. N. Gudivada, V.V. Raghavan, W.I. Grosky y R. Kasanagottu. "Information Retrieval on the World Wide Web". IEEE Internet Computing. Septiembre-Octubre, pag. 58-68. 1997.
- [2] R. Baeza-Yates and B. Ribeiro-Neto. "Modern Information Retrieval". ACM Press Books, Addison-Wesley. 1999.
- [3] Dunja Mladenic, "Text-Learning and related intelligent agents" Revised version in IEEE Expert special issue on Applications of Intelligent Information Retrieval, July-August 1999.
- [4] D. Koller y M. Sahami. "Toward Optimal Feature Selection". International Conference on Machine Learning. Editor L. Saitta. Volumen 13, Morgan-Kaufmann. 1996.
- [5] C. Musciano and B. Kennedy. "HTML The Complete Guide". McGraw Hill. 1997.
- [6] J. M. Pierre. "On the Automated Classification of Web Sites". Linköping Electronic Articles in Computer and Information Science. Vol. 6(2001): nr 0. Linköping University Electronic Press Linköping, Sweden. 2001.  
<http://www.ep.liu.se/ea/cis/2001/000/>
- [7] V. Fresno and A. Ribeiro. "Feature selection and dimensionality reduction in Web pages representation". International ICSC Congress on Computational Intelligence: Methods & Applications. Bangor, Wales (U.K.). June 2001.
- [8] Copernic Summarizer. 2001.  
<http://www.copernic.com/>

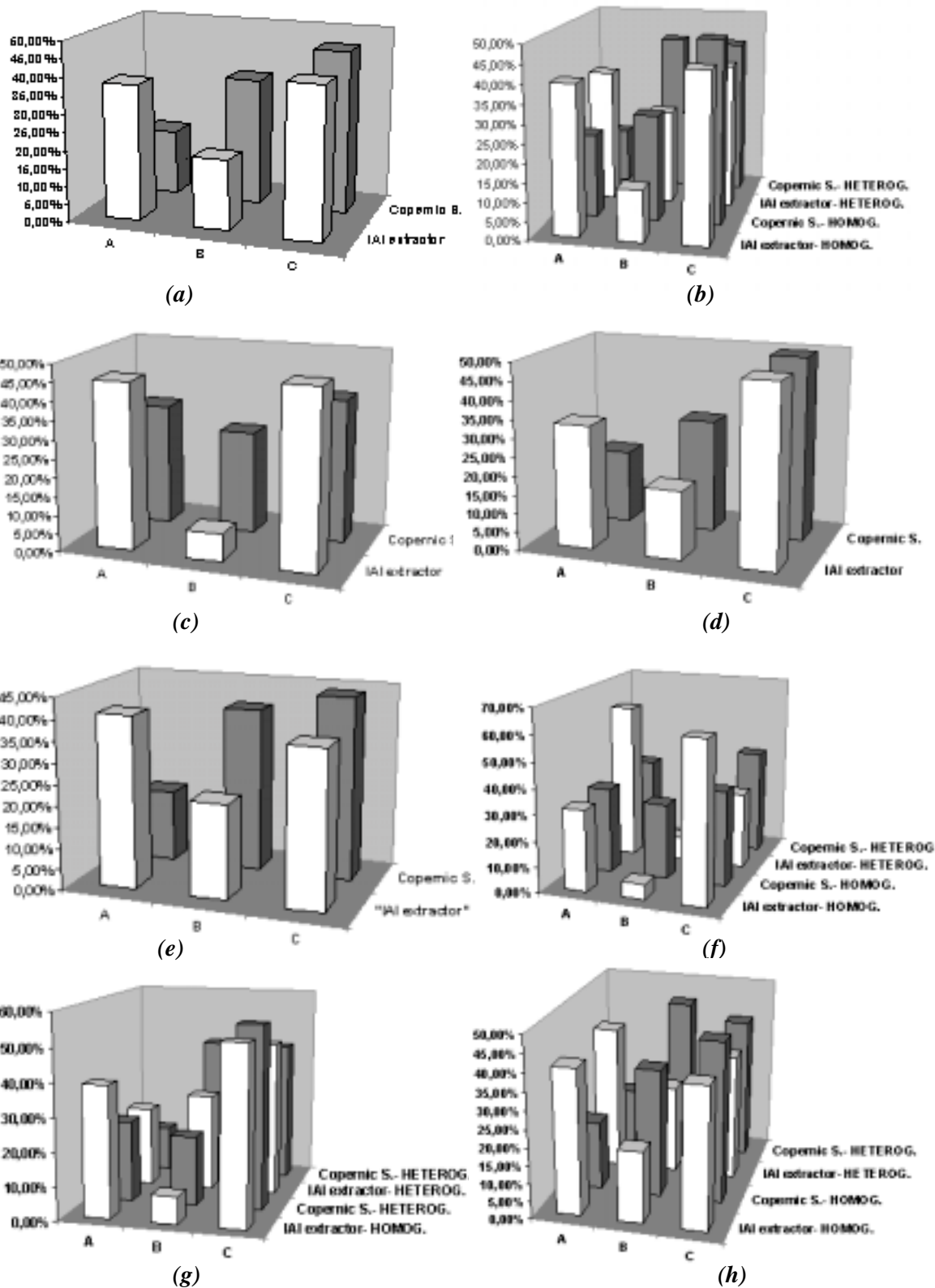


Figure 1: (a) General behavior. (b) Heterogeneous and homogeneous pages. (c) Pages with less than 50 words. (d) Pages with a number of words between 50 and 250 words. (e) Pages with more than 250 words. (f) Heterogeneous and homogeneous pages with less than 50 words. (g) Heterogeneous and homogeneous pages with a number of words between 50 and 250 words. (h) Heterogeneous and homogeneous pages with more than 250.