

# A Fuzzy System for the Web Page Representation

Angela Ribeiro, Víctor Fresno, María C. Garcia-Alegre, and Domingo Guinea<sup>1</sup>

Industrial Automation Institute. Spanish Council for Scientific Research.  
28500 Arganda del Rey. Madrid. Spain.  
angela@iai.csic.es

**Abstract.** This paper addresses the issue of an adequate representation of a web page, to perform further on classification and data mining. The approach focuses the textual part of web pages, which are represented by a two-dimension vector. The vector components are sorted by the relevance of each word in the text. Two approaches, analytical and fuzzy, that take advantage of characteristics of the HTML language are presented to compute the word relevance. Both models are contrasted in learning and classification tasks, to evaluate the suitability of each approach. The experiments show an obvious improvement of fuzzy method versus analytical one. The analytical and fuzzy approaches here presented are general, in the sense that every characteristic of the web pages could be easily integrated without additional cost.

**Keywords.** Concept extraction, feature vector, HTML texts, web characterization, web page representation, fuzzy decision system.

## 1 1 Introduction

The Internet and the World Wide Web are causing a revolution in many aspects of our lives. The core of this alteration is the hypothetical capacity to share efficiently and effectively the huge quantity of information embedded in the Net. But the rapid and chaotic growth of the Net, approximately a million electronic pages were daily created in 1999 [1], has generated a scarce structure and organization in the network information, which extremely complicates sharing or mining useful information.

Every inference process performed on the information needs a characterization of the web page. Among the multiple aspects that should be kept in mind to appropriately perform the page characterization, the textual part is one of the more relevant. 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 numerical data. However, the results of this first stage analysis is essential since the success of further on studies

---

<sup>1</sup> The work is supported by Innovatec S.A company and by a research grant (Ref. CAM 09/0052/1999).

strongly depends on the correctness of a text representation that captures the most pertinent aspects of the document. In fact, without a proper set of features a classifier will not be able to accurately discriminate between different categories.

The textual characterization of a page should be accomplished through the extraction of an appropriate set of relevant concepts; in other words a data structure for digital processing that represents the text included in the web page. The classic techniques of information retrieval [2] or text mining [3] are usually applied to obtain information from the web. When such techniques are utilized in the web, the problems appear not only as a consequence of both the enormous number of pages and their variability but also because web users are significantly different from the groups that traditionally had used information retrieval and text mining techniques. In addition, there are not standards or style rules in the web; the page content is created by a set of very distinct people and in an autonomous way. Furthermore, the inherited information retrieval technology has slowly progressed to take into account the web necessities. Consequently, the search engines, summarize tools, etc. show a short scope and a low accuracy.

It has been confirmed that the current search engines only recover a fraction of the overall documents and only a small portion of that fraction is significant [4] [5] [6].

In a few words, texts have been traditionally represented in a vector space model [7] [8] using basically two different methods. The first one represents a document by a vector of binary attributes indicating whether or not a word occurs in the document. In this case the frequency of a word in the document is not captured. The second model represents a document by the set of occurrences of each word. In both approaches the order of the words in the document is lost. The second method has been widely used in statistical language modeling for speech recognition [9]. Both representations generate vectors with a very high dimensionality (of  $10^4$  to  $10^7$  components) that avoids in many cases the use of knowledge extraction algorithms [10]. There are many reasons that explain that point. First, the time requirement for an induction algorithm often grows dramatically with the number of vector components, also called features. Furthermore, many learning algorithms can be viewed as estimators of the probability of the class label once a set of features is given. In domains with a large number of features, this distribution is very complex and of high dimensionality. Moreover, many algorithms use the Occam's Razor bias to prefer the simplest hypothesis that fits the data [11]. Irrelevant and redundant features are a source of problems in this context since they may confuse the learning algorithm masking the small set of truly.

Present paper describes both analytical and fuzzy based approaches that allow to obtain a web page representation with the final aim of carrying out classification and data mining. Both approaches take advantage of some characteristics of the HTML language to derive a significantly reduced vector

that contains the most representative words of the web document with an associate number that characterizes its relevance in the document. That is, a two-dimension vector which components, in first dimension, are words and in second dimension, numbers. The elements of the former dimension are ordered from the highest to the lowest number. Additionally, a comparative study both learning and classification perspectives is presented.

## 2 2 Web page structure

In the World Wide Web documents are expressed in HTML language. 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 [12], 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 associated tags (for instance `<b>` and `</b>`) the portion of the included text will assume such tags. With tags, users can indicate which words belong to the web page title, body, font style, headings, and many other attributes of the web page.

The textual tags or attributes will be the core of the method presented in this paper. Some textual tags are selected in order to represent the web page through present words in the web page text and a weight is assigned to each word that computes its relevance in the text.

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

1. Tags that indicate the page title (`<title>...</title>`).
2. Tags such as, `<b>...</b>`, `<u>...</u>`, `<em>...</em>`, `<i>...</i>`, and `<strong>... </strong>` that allow to emphasize parts of the text and, consequently, 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 these two cases, while emphasis is an operation consciously performed by the user when he is designing the web page, the title content could be the result of some automatic process and thus irrelevant in some cases [13].

In addition to these two properties there are other "classical" attributes that could be considered to compute the word relevance: the *word position* and the *word frequency* in the text. Other attributes such as, the *meta* tags could be considered, but right now they are not sufficiently used [14] and they have been obviated in the proposed representation.

### 3 3 An analytical approach to a web page representation

Derived from formerly described aspects, the following analytical characterization functions are defined to compute the relevance of a word in a web page. Then the calculation method of the relevance of a word in a page is also described.

#### 3.1 The frequency function of a word in a web page

$$f_f(i) = n_f(i)/N_{tot}$$

Where  $n_f(i)$  is the number of occurrences of a word  $i$  in a page and  $N_{tot}$  is the total number of word in the web page. This definition allows the normalization of the function:  $\sum_1^k f_f(i) = 1$ ;  $k$  is the number of different words in the document.

#### 3.2 The frequency function of a word in the title

$$f_t(i) = n_t(i)/N_{tit}$$

Where  $n_t(i)$  is the occurrence number of a word  $i$  in the title and  $N_{tit}$  is the number of words in the title. As previously  $\sum_1^k f_t(i) = 1$ ;  $k$  represents the different words in the title.

#### 3.3 Word emphasis function

$$f_e(i) = n_e(i)/N_{emph}$$

Where  $n_e(i)$  are the times that a word  $i$  is emphasized and  $N_{emph}$  the number of words that are emphasized in the whole document. As in former cases:  $\sum_1^k f_e(i) = 1$

#### 3.4 Word position function

To compute the relevance of a word from the position criterion, the web page is split into four parts attempting to characterize the fact that often the users structure the text so that the first and the last lines are more relevant than those in the middle. The defined function is:

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

Where  $n_{1,4}(i)$  are the occurrences of the word  $i$  in the first and the fourth quarter of the page and  $n_{2,3}(i)$  are the occurrences of the same word in the second and third quarter of the page.  $N_{tot}$  is the total number of words in the page. Notice that the first and the last parts have a greater weight than the intermediate ones, as in the majority of the cases there is a tendency to begin and to finish the text with the

relevant topics. In this case, since different pieces of the page have different weights:  $\sum_1^k f_i(i) < 1$

### 3.5 Calculation of the word relevance

Now the objective is to represent the web page through a feature vector composed by an ordered list of the words in the page. The relevance will specify the priority of each word in the list. The question addressed now deals with the fusion of the before mentioned functions to calculate an adequate relevance factor for each word in the page. A lineal combination appears as the easiest relation to compose all criteria:

$$r(i) = C_1 f_f(i) + C_2 f_t(i) + C_3 f_e(i) + C_4 f_p(i)$$

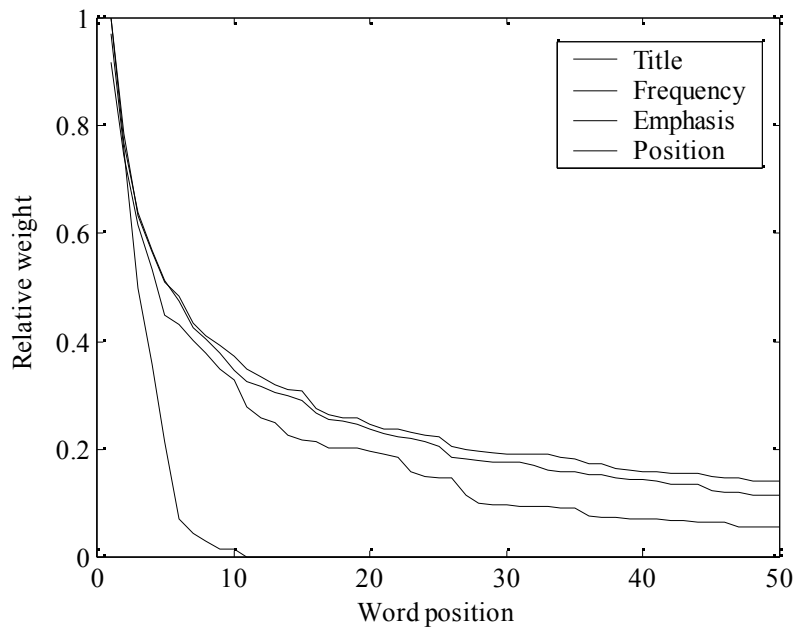
Now, the best coefficients  $C_1$ ,  $C_2$ ,  $C_3$  and  $C_4$  have to be obtained. A statistical study has been initially performed [13]. To this aim a sample set was selected to represent the heterogeneous nature of the World Wide Web. Web pages without a specific topic and with diverse page sizes were selected from the most important Spanish search engine, where pages are classified by a user and not by an automatic process. Once the sample was selected, each page was pre-processed as follows. First, HTML tags were removed and at the same time some important information, such as: which words appear emphasized, was saved. Second, words (stop-words) that do not have any important associated semantics and are irrelevant to express the textual content of a page: articles, prepositions, conjunctions, were eliminated. They represent around 39% of the words in a web page. Lastly, the number of line where the word appeared, its position in the line and its frequency were stored. Then with the total gathered information the value of each characterization function for each word in each page was computed.

The mean distribution of each characterization function of the sample set is displayed in the figure (see Fig. 3.1). The algorithmic process to obtain the graphics progresses as follows: a) Four feature vectors are computed for each element in the sample set one for each criterion. b) Components in each vector are ordered in an increasing way with respect to its relevance in the page. c) Only the first fifty components are considered to calculate the average values for the characterization functions.

Annualizing the graphic representation, individual quantitative characteristics for each function can be extracted. First, functions  $f_f(i)$  and  $f_p(i)$ , frequency and position, always are present and their mean value is higher than the average value in the rest of the functions. The reason for that is that all words has contributions to both criteria. Moreover, all words have a frequency value and a value associated to the position within the text, but a word do not necessarily belongs to the title neither is emphasized. Consequently, the emphasis and title criteria could not be selected alone, as they are not always included in all the possible cases. In fact, in the sample set it has been found emphasized words in the 89.30% of the pages and title tags in

the 97.05%, even though only the 51.97% of the titles were good representatives of the page.

In addition, it can be thought that the criterion of belonging to the title could be a good bias when the objective is to obtain a small feature vector. In fact, a clear reduction in the vector dimension is observed in the average value of the function when it is equal or smaller than the 40% (0.4 in Fig. 3.1).



**Fig. 3.1** The means for the four characterisation function.

Word frequency and position have a similar behavior. Both criteria contribute the same to the weight of each word and consequently they should have similar contributions to the characterization function combination. Furthermore, they are present in 100% of the sample set and so the combined contribution should be higher than 50%. Additionally, the appearance ratio between the emphasis and the title is 1.7, so that this value has to be accounted for in the criterion combination. On the other hand, a good function combination should weight more those criteria that greatly contribute to the reduction of the feature vector. But this is partially true, because the discrimination capability of the qualitative terms in the vector is also significant.

With these premises in mind there is a lot of potential values for the coefficients, the following one is a possible set of values.

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

Then the relevance of a word  $i$  in a page would be expressed as:

$$r(i) = 0.3f_f(i) + 0.15f_t(i) + 0.25f_e(i) + 0.3f_p(i)$$

## 4 4 A fuzzy approach to a web page representation

Up to now the relevance of a word in a web page is obtained by means of a lineal combination of a set of criteria. But all along the different statistical analysis performed [13], the difficulty for finding an optimal analytical function that adequately combines criteria, has emerged. The fundamental cue is that often a criterion evaluates the importance of a word only when it appears combined with another criterion. As an example, it has been demonstrated that the title not always describes the page content. Therefore if a word becomes visible in the title, it will really be relevant only when it also appears emphasized or when it has a high appearance frequency in the text. From the diverse quantitative and qualitative studies we have obtained several heuristics for the combinatory process. Similarly, our experience [15][16][17] is that a rule-based system could be the most appropriate method to properly combine the characterization functions. On the other hand, a fuzzy reasoning system is the most suitable framework to capture the qualitative human expert knowledge to solve the ambiguity inherent to the current reasoning process. Notice that fuzzy sets suitably model the uncertainty inherent to human reasoning processes, by embodying his knowledge and expertise in a set of linguistic expressions that manage words instead of numerical values [18][19][20].

As a consequence a fuzzy rule based system is suggested to derive the relevance factor for each word in a web page and to construct the feature vector of the page.

### 4.1 Linguistic variables

The design of a fuzzy system has nearly infinity degrees of freedom, and aims to provide “good enough” solutions that are biased to subjectivity and interpretation [21]. Additionally, it is important to ensure that the use of linguistic terms yields a more realistic representation of the problem than using its crisp counterpart. Keeping in mind the previously defined criteria to evaluate the word relevance in a text, four linguistic variables are defined: *text-frequency*, *title-frequency*, *emphasis* and *global-position* (see Fig. 4.1 and Fig 4.2). These terms are considered as the inputs of a fuzzy rule based approach. Their linguistic values are set in the antecedent of each IF-THEN rule. The rules consequent has a unique variable: the *relevance*.

In such a context criteria more independent of the page size, are needed. As an example, the frequency of a word in the page, in that case the division by the total

number of words in the page normalizes the criterion, but introduces an inconvenient effect in the fuzzy approach. If the document has a huge number of words, the criterion takes its values in an interval very close to zero. If the page has few words, the criterion takes a value in an interval very close to 1; in both cases the relation among words in the same page is correct and appropriate for an analytical model. But, under a fuzzy approach, it is difficult to decide on the more convenient parameters of the membership functions, since they are very dependent of the page size. The simplest solution is to redefine the criteria to be more independent of the page size without losing the initial meaning.

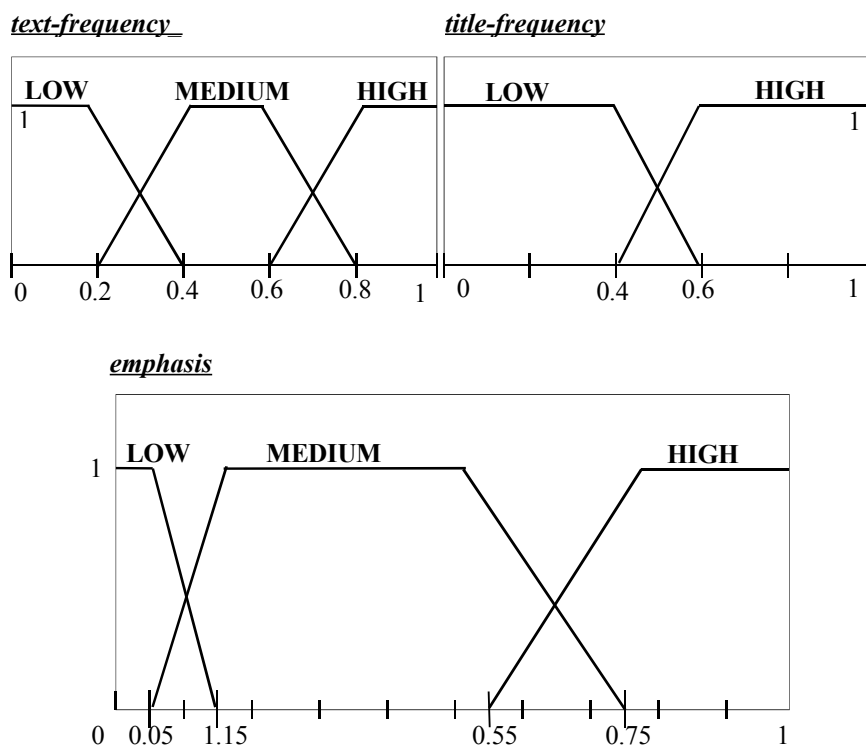


Fig. 4.1 Membership functions: text-frequency, title-frequency, and emphasis.

#### 4.1.1 The frequency criterion of a word in the page

$$c_f(i) = n_f(i) / N_{max}^{page}$$

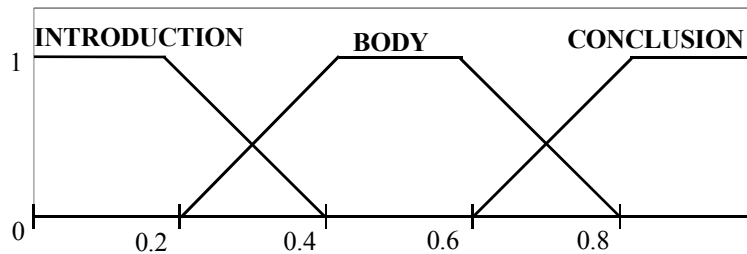
Where  $N_{max}^{page}$  is the occurrence number, in the page, of the most frequent word. Notice that for each word in the page at least there is an occurrence, therefore the universe of discourse of this fuzzy variable is  $(0,1]$ .

#### 4.1.2 The frequency criterion of a word in the title

$$c_t(i) = n_f(i) / N_{max}^{title}$$

Where  $N_{max}^{title}$  is the occurrence number, in the title, of the most frequent word in the title. In this case the word could not be in the title, so that the universe of discourse is  $[0,1]$ .

##### Line-position



##### global-position

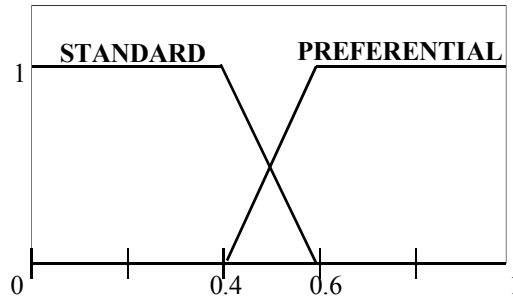


Fig 4.2 Membership functions of the *line-position* and *global-position* variables

#### 4.1.3 Word emphasis criterion

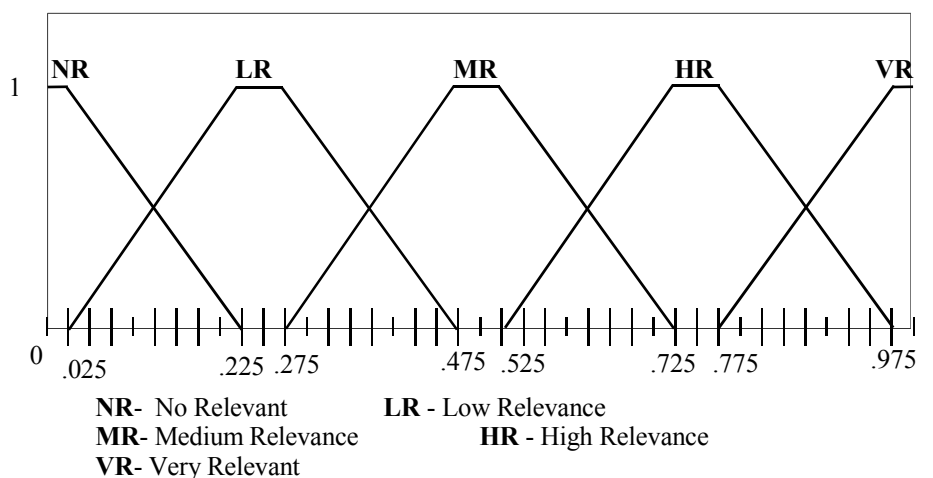
$$c_e(i) = n_e(i) / N_{max}^{emph}$$

Where  $N_{max}^{emph}$  is the emphasized occurrence number of the word the most emphasized. The universe of discourse of this fuzzy variable is  $[0,1]$ . The

membership functions and linguistic labels for the three criteria are shown in (see Fig. 4.1).

#### 4.1.4. Word position criterion

The position criterion of a word must be calculated from the contribution of every relevance



**Fig. 4.3** Membership functions of the output variable *relevance*

position of the word in the page. In other words, a term can appear in different lines of the text and the word position criterion should keep all of them. The contribution of an occurrence  $o$  of a word  $i$  in a text line  $l$  can be expressed as follows:

$$c_p(i, o) = n_p(i, o) / N_{tot}$$

Where  $n_p(i, o)$  is the line number of the occurrence  $o$  of word  $i$  and  $N_{tot}$  is the total number of text lines in the page. To calculate the global position criterion of a word in the totality of the page, another fuzzy system is used. So that the expression (1) is calculated for each occurrence  $o$  of a word  $i$ . Through a fuzzification process, linguistic labels (see Fig. 4.2, INTRODUCTION, BODY and CONCLUSION) are assigned at each occurrence  $o$  of the word  $i$ . This representation allows for capturing a gradual behavior in the transition points of the proposed page partition in the analytical approach. The output of the system are two linguistic labels: STANDARD and PREFERENTIAL that represent whether or not a word  $i$  globally belongs to a favored parts of the text. The set of IF-THEN rules is described in Table 4.1.

**Table 4.1.** Rule Base for obtaining the global position criterion of a word in a page.

<i>line-position</i>	<i>global-position</i>
INTRODUCTION	PREFERENTIAL
BODY	STANDARD
CONCLUSION	PREFERENTIAL

#### 4.1.5 The output value: the word relevance

The relevance output variable of the fuzzy system has five associated labels (see membership functions in Fig 4.3). The linguistic labels set represents the whole range of values, used by a human expert that has to score each text word with its level of relevance in the text.

#### 4.2 Fuzzy IF-THEN Rules

The knowledge acquisition is the first step in the design of a fuzzy system. In current approach, the previous statistical study and the common sense are crucial to define the rules. Following statements are important to understand the IF-THEN rules base:

- a) A non-emphasized word can be because no words are emphasized in the web page.
- b) A word not appearing in the title may indicate that the page has not title or the title has not meaning, i.e. it does not enclose relevant words.
- c) In general, the position is a criterion that weighs the longer pages than shorter ones.
- d) A high frequency of a word in a page could mean that the word is a "joker" in the sense that its meaning is not discriminate and the word can be used in several contexts with different meanings. Notice that the elimination process of stop-word does not remove all words with "unclear" meaning.

All earlier statements in addition to the analysis in former paragraphs has lead to the rules base shows in tables Table 4.2 and Table 4.3.

**Table 4.2.** Rule base I

<i>title-frequency</i>	<b>Antecedents</b>		<i>global-position</i>	<b>Consequent relevance</b>
	<i>text-frequency</i>	<i>emphasis</i>		
HIGH	HIGH	HIGH		VR
HIGH	MEDIUM	HIGH		VR
HIGH	MEDIUM	MEDIUM		HR
HIGH	HIGH	MEDIUM		VR
HIGH	LOW	LOW	PREFERENTIAL	MR
HIGH	LOW	LOW	STANDARD	LR
LOW	LOW	LOW		NR
LOW	HIGH	HIGH	PREFERENTIAL	VR
LOW	HIGH	HIGH	STANDARD	HR
HIGH	LOW	MEDIUM	PREFERENTIAL	HR
HIGH	LOW	MEDIUM	STANDARD	MR
HIGH	LOW	HIGH	PREFERENTIAL	VR
HIGH	LOW	HIGH	STANDARD	HR
HIGH	HIGH	LOW	PREFERENTIAL	VR
HIGH	HIGH	LOW	STANDARD	HR
LOW	LOW	MEDIUM	PREFERENTIAL	MR

**Table 4.3.** Rule Base II

<i>title-frequency</i>	<b>Antecedents</b>		<i>global-position</i>	<b>Consequent relevance</b>
	<i>text-frequency</i>	<i>emphasis</i>		

LOW	LOW	MEDIUM	STANDARD	LR
LOW	LOW	HIGH	PREFERENTIAL	HR
LOW	LOW	HIGH	STANDARD	MR
LOW	MEDIUM	LOW	PREFERENTIAL	LR
LOW	MEDIUM	LOW	STANDARD	NR
LOW	MEDIUM	MEDIUM	PREFERENTIAL	MR
LOW	MEDIUM	MEDIUM	STANDARD	LR
LOW	MEDIUM	HIGH	PREFERENTIAL	VR
LOW	MEDIUM	HIGH	STANDARD	HR
LOW	HIGH	LOW	PREFERENTIAL	MR
LOW	HIGH	LOW	STANDARD	LR
LOW	HIGH	MEDIUM	PREFERENTIAL	HR
LOW	HIGH	MEDIUM	STANDARD	MR
HIGH	MEDIUM	LOW	PREFERENTIAL	MR
HIGH	MEDIUM	LOW	STANDARD	LR

---

Each knowledge base (rules base) is designed with the aid of the FuzzyShell programming environment [22] and is recorded as a text file. The dynamic memory allocation allows the definition of an unbound number of variables, linguistic labels and rules in the text file [23], with a specific syntax interpreted by the fuzzy reasoning system. This tool allows to speed up the design and the tuning of the fuzzy reasoning system, keeping the user at a higher level of abstraction, to focus attention on the application design and not on the low detail of the implementation level.

In both fuzzy systems here presented the fuzzyfication of variables, inference and defuzzyfication processes are performed through a function library. Library function can be called from several diagnosis systems, each one defined by means of a rules base. The inference engine is based on a center of mass algorithm (COM) that weights the output of each rule in the knowledge base with the truth degree of its

antecedent. The output is a linguistic label with an associated number related to the relevance of a specific word in the page.

## 5 5 A comparative study

The obtained representations of a web page by means of the analytical and fuzzy approaches have been used to perform a supervised learning process that permits the extraction of the descriptors for two classes: *Medicine-Pharmacology* and *Technology Aerospace*. The final aim being the comparison of the two possible ways that guide to a representations of a web page; the analytical and the fuzzy criteria fusion approaches. Notice that the selected classes are extremely different from one each other, as within the context of the comparative study, the focus of attention is addressed to the achievements of the two methods and therefore both classes are selected to display the most favorable conditions for learning and classification tasks. Once the learning stage is achieved, a classification process is carried out to test the performance of the obtained class descriptors. In the next paragraphs the main aspects of both the learning and the classification processes are presented, as well as the most relevant results of the proposed test.

### 5.1 The learning phase

The class descriptors are obtained from a supervised learning process. Assuming *central limit theorem*<sup>2</sup>, a vector of three dimensions can represent a class; where the first components are the words that belong to the class and the second and third one hold the  $\mu$  and  $\sigma$  parameters that define the *normal* or *gaussian* distribution

$$f_i(r_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-(r_i - \mu)^2 / 2\sigma^2}$$

The density function  $f_i$  gets the probability that a word  $i$ , with relevance  $r$ , appears in a class. The mean and variance are obtained from the two selected sets of examples for each class by a *maximum likelihood estimator* method.

Assuming that the classes are independent of each other, as well as words in a class (derived from the selected representation); then for each word in each page of the sample set the mean value and the standard deviation can be obtained as

$$\mu_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} s_{ijk} \quad \sigma_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} (s_{ijk} - \mu_{ij})$$

Where  $N_{ij}$  represents the occurrences of the word  $i$  in the training set of the class  $j$ , and  $s_{ijk}$  the normalized relevance of the word  $i$  in the page  $k$ . That is:

$$s_{ijk} = r_{ik} / R_{max}^k$$

---

<sup>2</sup> The mean  $\bar{X}_n$  of a random sample from any distribution with finite variance  $\sigma^2$  and mean  $\mu$  is approximately distributed as a normal random variable with mean  $\mu$  and variance  $\sigma^2/n$ .

$R_{max}^k$  is the maximum relevance for a word in the page  $k$  and the word relevance is derived from was criteria fusion process. With this normalization dependence of the relevance from the page size is avoided.

Finally, a threshold for  $\sigma^2$  is defined to reduce the description of the class, in such a way that only words with a enough probability will be representative of a class. Then the class descriptor contains only the most representative words of all appearing in the pages of the training set.

## 5.2 The classification phase

A Bayesian classifier has been selected as the probabilistic approaches are among the most effective to classify text documents [11]. The optimal classification of a new page  $p_k$ , is the class  $c_j$  for which the probability  $P(c_j | p_k)$  is maximum.  $P(c_j | p_k)$  also referred as posteriori probability of the class  $c_j$ , reflects the confidence that  $c_j$  holds given the page  $p_k$ . Then, applying the Bayes theorem the following expression has to be maximized:

$$\operatorname{argmax}_{c_j \in C} \left( \frac{P(p_k | c_j)P(c_j)}{P(p_k)} \right)$$

Under the hypothesis of independence among the word in the page and assuming that all pages have the same prior probability, former expression can be formulated as:

$$\operatorname{argmax}_{c_j \in C} (P(c_j) \prod_i^{N_k} P(w_i | c_j))$$

Where  $N_k$  is the number of different words in the page  $p_k$ . Here, the expression of the *normalized relevance*  $s_i$  is introduced, so as to ponder the probability of  $w_i$  given  $c_j$  with a normalized factor that reflects the significance of the word  $w_i$  in the page.

$$\operatorname{argmax}_{c_j \in C} (P(c_j) \prod_i^{N_k} s_i P(w_i | c_j))$$

Now then, the probability that  $w_i$  holds given  $c_j$  can be expressed as a normal distribution with variance  $\sigma^2$  and mean  $\mu$ , as far as the former class representation is used. Then:

$$\operatorname{argmax}_{c_j \in C} (P(c_j) \prod_i^{N_k} s_i \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2\sigma^2} (s_i - \mu_{ij})^2})$$

Rather than maximizing this expression its logarithmic transformation is chosen, as the logarithm of  $f$  is a monotonic function of  $f$ . Therefore maximizing the logarithm of  $f$  also implies the maximization the function  $f$ .

$$\operatorname{argmax}_{c_j \in C} (\ln P(c_j) + \sum_i^{N_k} \ln(s_i \frac{1}{\sqrt{2\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2}))$$

Finally  $P(c_j)$  represents the prior probability of the class  $c_j$  and reflects the expert knowledge on  $c_j$  as a correct hypothesis. If such prior knowledge would not exist, then the same prior probability would be assigned to each candidate hypothesis. It means that the term  $\ln P(c_j)$  is a constant independent of the tested class, and can therefore be discarded, yielding

$$\operatorname{argmax}_{c_j \in C} (\sum_i^{N_k} \ln(s_i \frac{1}{\sqrt{2\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2}))$$

Now then, the logarithmic function ( $\ln x$ ) presents a discontinuity for  $x = 0$ . To avoid such effect the function is shifted

$$c_{ML} = \operatorname{argmax}_{c_j \in C} (\sum_i^{N_k} \ln(s_i \frac{1}{\sqrt{2\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2} + 1))$$

The maximum likelihood hypothesis  $c_{ML}$  for classifying a given page  $p_k$  can be derived from this final expression, which is used to test the performance of both analytical and fuzzy approaches. Table 5.1 displays the results obtained for the two classes with a sample set of 200 examples selected from the World Wide Web.

The choice of the sample set was made picking up pages of different sizes from different Web Servers. Then, a manual classification was carried out avoiding the automatic classifications that could bias the classification process of the initial sample sets.

On the other hand, the 65% of the sample set has been used in the learning phase and the rest (35%) is been used in the classification stage. Data in the Table 5.1 shows a clear improvement of the fuzzy approach versus the analytical one in the classification.

**Table 5.1.** Main results for the comparative study

		6 Medicine-Pharmacology		7 Technology-Aerospace		8 Mean values	
		Successes (%)	Faults (%)	Successes (%)	Faults (%)	Successes (%)	Faults (%)
9	ANALYTICAL	82.45	17.55	71.93	28.07	<b>77.19</b>	<b>23.61</b>

## 11 6 Conclusions

This paper presents a new approach to the web page representation and two different methods to obtain it. The proposed representation is a two-dimension vector, where the first component is a word extracted from the textual component of the page and the second one is a number, which evaluates the word representativeness in the web page text. To calculate this second component some characteristics of the HTML language and others attributes more “classical”, often used in information retrieval, have been taken into account.

To know the behavior of the suggested function/criteria a statistical study has been accomplished. This study shows the major cues to achieve a proper fusion that gives a relevance factor for each word in the web page, namely the second component of the vector. Two approaches have been suggested and implemented to solve the fusion.

In the second phase, learning and classification algorithms have been developed and demonstrated to test the performance of the two fusion methods. As a result of the test, a clear improvement of the fuzzy fusion technique versus the analytical one is derived. The method using a fuzzy variable becomes more practical and cost-effective than the one that operate with crisp variables. Also it is important to stand out that the analytical approach to fuse functions/criteria has been compared with commercial tools [24]; displaying a clear improvement in extract relevant concepts/words in a web page with respect to the commercial software available. Notice that the representation of a web page can be seen as an extraction of the most relevant concepts/words.

On the other hand, the achievement of an adequate representation of a web page, more precisely of the textual part, is fundamental issue with many applications, as it is the first task that must be performed, in the pre-processing step, before carrying out a mining/analysis. Moreover, this representation step, that can be view as a concept extraction phase, is profitable by its-self without any later on associated task, such as a clustering, classification, etc. This information can be useful in order to know if a page could be or not interesting without reading the full page or as the first step of an automatic organization of the bookmark directory.

## 12 References

1. Members of Clever Project: Hypersearching the Web, (1999) *Scientific American*.
2. Gudivada V. N., Raghavan V.V., Grosky W.I. and Kasanagottu R., (1997) Information Retrieval on the World Wide Web. *IEEE Internet Computing*, 58-68.

3. Merkl D., (1998). Text Data Mining. *A Handbook of Natural Language Processing Techniques and Applications for the Processing of Languages as Text*. Dale R., Moisl H. y Somers H. (Eds). NewYork: Marcel Dekker.
4. Lawrence S. And Giles C., (1999). Accessibility of information on the Web. *Nature*, **400**, 107-109.
5. Meng X. and Chen Z., (2001). The Architecture of Yarrow: A Real-Time Intelligent Meta-Search Engine. *International Conference on Internet Computing 2001(IC'2001)*. Las Vegas (U.S.A.), **1**, 7-13.
6. Kerschberg L., Kim W., and Scime A., (2001). WebSifter II: A Personalizable Meta-Search Agent Based on Weighted Semantic Taxonomy Tree. *International Conference on Internet Computing 2001(IC'2001)*. Las Vegas (U.S.A.),**1**, 14-20.
7. Baeza-Yates R. ans Ribeiro-Neto B., (1999). *Modern Information Retrieval*. ACM Press Books, Addison-Wesley.
8. Mladenic D., (1999). Text-Learning and related intelligent agents. *IEEE Expert special issue on Applications of Intelligent Information Retrieval*.
9. McCallum A. and Nigam K., (1998). A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of Workshop on Learning for Text Categorization, AAAI/ICML-98*. AAAI Press, 41-48.
10. Koller D. and Sahami M., (1996). Toward Optimal Feature Selection *International Conference on Machine Learning*. Editor Saitta L., Morgan-Kaufmann., **13**.
11. Mitchell T. M., (1997). *Machine Learning*. McGraw-Hill International Editions.
12. Musciano C. and Kennedy B., (1997). *HTML: The Complete Guide*. McGraw Hill.
13. Fresno V. and Ribeiro A., (2001). Feature selection and dimensionality reduction in Web pages representation. *International ICSC Congress on Computational Intelligence: Methods & Applications*. Bangor, Wales (U.K.), 416-421.
14. Pierre M., (2001). On the Automated Classification of Web Sites. *Linköping Electronic Articles in Computer and Information Science*, **6**. Linköping University Electronic Press Linköping, Sweden..
15. García-Pérez L., Marchant J., Hague T. and García-Alegre M.C., (2000). Fuzzy Decision System for Threshold Selection to Cluster Cauliflower Plant Blobs from Field Visual Images. *SCI2000*, Orlando, 23-28.
16. Garcia-Alegre M.C., Ribeiro A., Gasós J., Salido J., (1993). Optimization of fuzzy behavior-based robots navigation in partially known industrial environments. *IEEE Inter.Conf.on Industrial Fuzzy Control & Intell.Syst*. Houston,TX, 50-54.
17. García-Pérez L., Cañas J. M., García-Alegre M.C., Yáñez P., Guinea D., (2000). Fuzzy Control of an Electropneumatic Actuator. *STYLF2000*, Sevilla, 133-138.
18. Garcia-Alegre M.C., (1991). Artificial Intelligence in Process Control: Fuzzy Controllers. *Mundo Electrónico*, **214** , 42-49.
19. Isermann R., (1998). On Fuzzy Logic Applications for Automatic Control Supervision and Fault Diagnosis. *IEEE Trans.Syst.Man and Cybern*, **28**, 221-235.

20. Hansen B. K., (2000). Analog forecasting of ceiling and visibility using fuzzy sets. *AMS2000*.
21. Berkan R.C. and Trubatch S. L., (1997). Fuzzy Systems Design Principles: Building Fuzzy IF-THEN Rules Bases. *IEEE Press*.
22. Gasós J., Fernández P.D., García-Alegre M.C., Garcia Rosa R., (1990). Environment for the development of fuzzy controllers. *Proc. Intern.Conf. on A.I. Applications & N.N.*, 121-124,.
23. Cañas, J.M, Garcia-Alegre, M.C., (1999). Modulated agents for autonomous robot piloting. *Proc. 8<sup>th</sup> A.I. Spanish Conf.*, Murcia, 98-106.
24. Ribeiro A. and Fresno V., (2001). A Multi Criteria Function to Concept Extraction in HTML Environment. *International Conference on Internet Computing 2001(IC'2001)*. Las Vegas (U.S.A.), **1**, 1-6