



An Analytical Approach to Concept Extraction in HTML Environments

VICTOR FRESNO*

v.fresno@escet.urjc.es

Escuela Superior de Ciencias Experimentales y Tecnología, Rey Juan Carlos University, 28933 Mostoles, Madrid, Spain

ANGELA RIBEIRO

angela@iai.csic.es

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

Abstract. The core of the Internet and World Wide Web revolution comes from their capacity to efficiently share the huge quantity of data, but the rapid and chaotic growth of the Net has extremely complicated the task of sharing 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 properly represent the text included in the Web page. This paper presents a method to obtain such a set of relevant concepts from a Web page, essentially based on a relevance estimation of each word in the text of a Web page. The word-relevance is defined by a combination of criteria that take into account characteristics of the HTML language as well as more classical measures such as the frequency and the position of a word in a document. Besides, heuristic rules to obtain the most suitable fusion of criteria is achieved via a statistical study. Several experiments are conducted to test the performance of the proposed concept extraction method compared to other approaches including a commercial tool. The results obtained here exhibit a greater success in the concept extraction by the proposed technique against other tested methods.

Keywords: concept extraction, feature vector in HTML texts, Web page characterization, Web page representation, Web page classification

1. Introduction

Due to the rapidly expanding size of the Web, locating useful information on it has become increasingly difficult. Additionally, Internet users numbered 30% more at the end of 2002 than at the end of 2001 in spite of the economic crisis (UNCTAD, 2002). In these circumstances the need for tools that adequately and efficiently extract information from the Web is becoming important and urgent.

The classical techniques of Information Retrieval (IR) are normally used to obtain information from Internet (Gudivada et al., 1997; Dunham, 2002). Frequently, when these techniques are applied, problems appear not only due to the enormous number of pages or the continual changes in them but also because Web users are significantly different from

*Work carried out while at IAI-CSIC.

the groups that traditionally used the IR techniques. In the Web, there are no standards or style rules; the page content is created by a set of very heterogeneous people and in an autonomous way. The inherited IR technology has only slowly progressed to take into account these Web factors. Consequently, the search engines display a limited scope and poor accuracy; in other words, they only retrieve a small fraction of the total number of existent documents, and of this fraction only a small portion is significant. The most well-known Web search services do return a ranked list of Web pages in response to a user's search request. However, Web pages on different topics or different aspects of the same topic are mixed together in the returned list. Chen and Dumais (2000) argue that users often prefer to have a search result organized into a hierarchical category structure. A category-based view of retrieved documents enables users to find the most relevant information more efficiently.

Several techniques have been proposed to classify a Web page. Many research works are based on the Web page content (the intra-document structure), and others on the structure of the hyperlinks within the Web itself (inter-document structure). In the latter case, research works are inspired by the study of social networks and citation analysis (Henzinger, 2000). Furthermore, to extract rich and predictive features from both sources some research works combine hyperlinks and textual information. What is important is to be aware that for some real-world cases the only useful place to look for information about the class label of a document is the document itself. In other words, looking outside the document can degrade the classification performance. When there are no hyperlink regularities, no benefit from using hyperlinks can be expected and in this case it is advantageous to use standard text classifiers on the text of the document itself (Yang et al., 2002). In Yang et al. (2002), a study of approaches to hypertext categorization that explores various hypotheses about the structure of the hypertext is made. Some relevant conclusions of this study that motivated our approach are:

1. The identification of hypertext regularities in the data and the selection of appropriate representations for hypertext are crucial for the optimal design of a classification system; and
2. Meta data about the Web pages can be extremely useful for improving the classification performance. This suggests the importance of exploring information extraction techniques for automated acquisition of meta data. Recognizing useful HTML fields in hypertext pages and using them together can improve the classification performance.

It has to be noted that in this context the meta data are contained in the Web pages (i.e. HTML tags as META and TITLE tags) and they are technically not meta data because they are internal and not external to the Web page. Nevertheless, these tagged fields can be treated differently from other parts of the Web documents and for this reason they are called meta data.

From a general point of view and for any document type, the first step for 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. However, access to the information content of the text is difficult, since the relationship between the form (usually a sequence of characters) and the meaning is not as clear for text as it is in the case of numeric data. Nevertheless, 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. Without a proper set of features, a classifier will not be able to accurately discriminate between different categories.

Texts or unstructured documents have been traditionally represented by a vector space model (Salton et al., 1975; Mladenic, 1999). The vector representation (Baeza-Yates and Ribeiro-Neto, 1999) takes single words, found in the training corpus, as features. This representation ignores the sequence in which the words occur and is based on the statistics about single independent words. The feature could be Boolean (a word either occurs or does not occur in the document), or frequency based (frequency of the words in the document) known as “bag-word” approach. Variations of the feature selection include removing the case, punctuation, infrequent words, and stop words. In Kosala and Blockeel (2000), can be found a complete list of less traditional methods. In all cases of the vector space approach, the information about the order of the words in the document is lost. Moreover, both representations generate vectors with very high dimensionality (of 10^4 to 10^7 components) which hinders, in many cases, the use of knowledge extraction algorithms (Koller and Sahami, 1996). This paper describes an approach that allows for obtaining a Web page representation with the final aim of carrying out classification and data mining. The proposed method takes advantage of an adequate combination of traditional text representation methods with some characteristics of the HTML language, deriving a significantly reduced-dimensional vector that contains the most representative words/concepts of the web document with an associated number that characterizes its relevance to the document. It produces two related vectors with representative words (vector 1) and the associated numbers (vector 2). The elements of the word vector are ordered according to the values (from highest to lowest) of the elements of the associated number vector.

The remainder of this paper is organized as follows. In Section 2 some HTML language characteristics relevant to the proposed approach are outlined. In Section 3 the characterization functions to calculate the relevance of a word in a Web page are defined. Section 4 is devoted to the statistical study of the defined characterization functions in order to obtain the relevance of a word via a linear combination of the characterization functions. In Section 5 a comparative study is presented to show the improvement of the proposed Web page representation or information extraction over that of a commercial tool. In Section 6 a Naive-Bayes classifier adapted to the proposed representation is introduced as well as a study of the classifier performance. Finally, some conclusions are discussed in Section 7.

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 display.

There are many types of Web tags (Musciano and Kennedy, 1997), 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 positioned between two related corresponding tags (for instance `` and ``) the portion of text will assume that tag.

With tags, authors 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 focus of the method presented in this paper, in addition to the textual content regarded as plain text. Some textual tags are selected in order to represent the Web page through words present in the web page text and a weight is assigned to each word that evaluates its relevance in the text.

The most promising attributes of a page which contain information about the significance of a word in the text are:

1. Tags that indicate the page title (`<title>` `</title>`).
2. Tags such as, `` ``, `<u>` `</u>`, `` ``, `<i>` `</i>`, or `` `` that emphasize parts of the text and, consequently, 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 by assigning it a weight. The same consideration holds for the emphasized sentences in the text. However, there is an essential difference between these two cases, because while emphasis is an operation consciously performed by the author in the Web page design, the title content could be the result of some automatic process and thus irrelevant in some cases (Fresno and Ribeiro, 2001).

In addition to these two attributes, there are other more “classical” issues that could also be considered to compute the word relevance in a document: the word position in 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 authors often 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 the criteria. Such possible irrelevance only reinforces the thesis that the analysis of multiple features application of multiple criteria is the most appropriate solution in a heterogeneous domain such as the Internet.

At this point it is appropriate to define a set of functions which evaluate the word relevance through the previously described aspects.

3. The characterization functions

Based on the above discussion, the following analytical characterization functions are defined to compute the relevance of a word in a Web page.

3.1. The frequency function of a word in a Web page

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

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

3.2. The frequency function of a word in the title

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

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

3.3. Word emphasis function

$$f_e(i) = n_e(i)/N_{\text{emph}} \quad (3)$$

Here $n_e(i)$ is the number of times that a word i is emphasized and N_{emph} the total 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 plain text in the Web page is split into four quarters recognizing the fact that often the authors structure the text so that the first and the last quarters are more relevant than those in the middle. We can define two preferential quarters and two other standard quarters in the four parts that form the plain text of the page, and then we can assign a weight of $3/4$ to preferred quarters and a weight of $1/4$ to the standard quarters of the text. The defined function is:

$$f_p(i) = \frac{\frac{3}{4}n_{1,4}(i) + \frac{1}{4}n_{2,3}(i)}{\sum_{l=1}^k (\frac{3}{4}n_{1,4}(l) + \frac{1}{4}n_{2,3}(l))} \quad (4)$$

Here $n_{1,4}(i)$ is the number of occurrences of the word i in the first and the fourth quarter of the page (preferential parts) and $n_{2,3}(i)$ are the occurrences of the same word in the second and third quarter of the page (standard parts). In this case, since different pieces of the page have different weights: $\sum_1^k f_p(i) < 1$. The defined denominator overcomes this problem of normalizing the function; k is the number of different words in the page. The word position function can be transformed to achieve an expression analogous to former functions as follows:

$$\begin{aligned} f_p(i) &= \frac{2n_{1,4}(i) + n_t(i)}{\sum_{l=1}^k 2n_{1,4}(l) + n_t(l)} = \frac{2n_{1,4}(i) + n_t(i)}{\sum_{l=1}^k 2n_{1,4}(l) + \sum_{l=1}^k n_t(l)} \\ &= \frac{2n_{1,4}(i) + n_t(i)}{2n_{\text{tot}-1,4} + N_{\text{tot}}} \end{aligned} \quad (5)$$

where $n_t(i) = n_{1,4}(i) + n_{2,3}(i)$ and it represents the occurrences of a word i weighted according to the word frequency in appropriate quarter (preferential or standard) of the

document. On the other hand, $n_{\text{tot}-1,4}$ represents the total number of words in the preferential parts of web page and N_{tot} is the total number of words in the page.

Finally, the previous equation can be expressed, for each different word i , in a general form as

$$f_p(i) = \frac{(a-b)n_{1,4}(i) + bn_t(i)}{(a-b)n_{\text{tot}-1,4} + bN_{\text{tot}}} \quad (6)$$

with $a + b = 1$ and where a and b represent the weights given to each one of the parts (preferential and standard respectively) of the text.

4. Calculating the relevance of a word in a Web page

For calculating the relevance of a word, the first objective is to represent the Web page through a feature vector composed of an ordered list of the words in the page. The relevance will then specify the priority of each word in the list. The question now addressed deals with the fusion of the functions defined in Section 3 to calculate an adequate relevance factor $r(i) = r_i$ for each word in the page. A linear combination appears to be the easiest relationship to include all criteria, that is,

$$r_i = C_f f_f(i) + C_t f_t(i) + C_e f_e(i) + C_p f_p(i) \quad (7)$$

Notice that each coefficient C_f , C_t , C_e , and C_p can take values into $[0, 1]$ interval, and $C_f + C_t + C_e + C_p = 1$. The values for the coefficients C_f , C_t , C_e , and C_p can be obtained by statistical analysis (Fresno and Ribeiro, 2001) as follows.

A sample set of pages 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 a major Spanish search engine, where pages are classified by the author and not by an automatic process.

After selecting the sample pages, 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 or in the title, was saved. Next, words (stop-words) that do not have important associated semantics and are irrelevant to expressing the textual content of a page (such as articles, prepositions and conjunctions) were eliminated. They represent around 39% of the words in the Web pages of the sample set. Lastly, the line number where each considered word appeared, its position in the line and its frequency, were stored. With all this collected information four feature vectors are computed for each element in the sample set, one for each criterion. Then, components in each vector are ordered with highest to lowest values of their relevance in the page. The result was four feature vectors for each Web page in the sample set. Then only the fifty components with the highest relevance are considered in calculating the mean for each characterization function as shown in figure 1. The chart shows on the y-axis the mean of relevance factor values for the fifty first components of the feature vector, which are plotted on the x-axis.

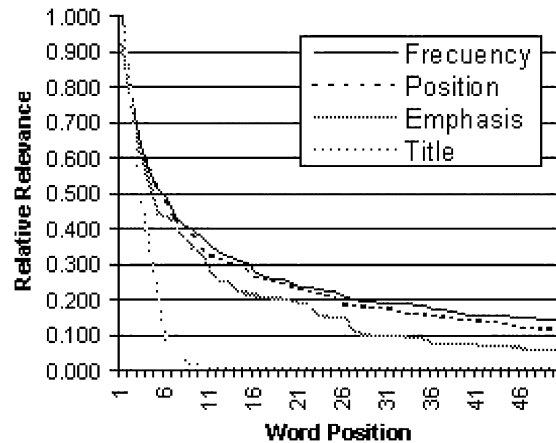


Figure 1. Statistical analysis of the Web pages: The mean distribution of the relative relevance for first fifty components of the feature vector. Results for four characterization functions are showed.

An analysis of figure 1 allows the extraction of the following quantitative facts. First, the frequency and position functions are always present and therefore their means are higher than the mean values of the other functions. This is because all words have a frequency and a position in the text, but a word does not necessarily belong to the title, neither does it have to be emphasized. Consequently, the emphasis and title criteria cannot always be estimated for every word in the Web page. In fact, in the sample set it was found that emphasized words were present in 89.30% of the pages and title tags in 97.05%.

In addition, it may be assumed that the criterion of belonging to the title could be a good basis when the objective is to obtain a short feature vector. In fact, a clear reduction in the vector dimension is observed in the mean value of the relevance when it is equal to or less than 60% (0.6 in figure 1). However, in the sample set, only 51.97% of the titles were really representative of the page content, therefore this criterion should not be used alone.

In figure 2 the behavior of different criteria combination, that are built using diverse values for coefficients C_f , C_t , C_e , and C_p is displayed.

This analysis shows that the frequency and the position have a similar influence on relevance. Both properties contribute to raise the relative relevance of a word and consequently, they give rise to similar contributions in different criteria combinations. Furthermore, they are present in 100% of the sample page set and so their combined contribution should be higher than 50%. Additionally, words with emphasis appear 1.7 time more frequently than title words, and we must take into account this value in combining of the functions. On the other hand, it seems that a good combination for the criteria would give more relevance to those features that lead to a greater reduction in the size of feature vector. But this is only partially true, because the discrimination capability of the words in the vector is also important (Fresno and Ribeiro, 2000). Based on these considerations, a good selection of values for the linear combination coefficients could be: $C_f = 0.30, C_t = 0.15,$

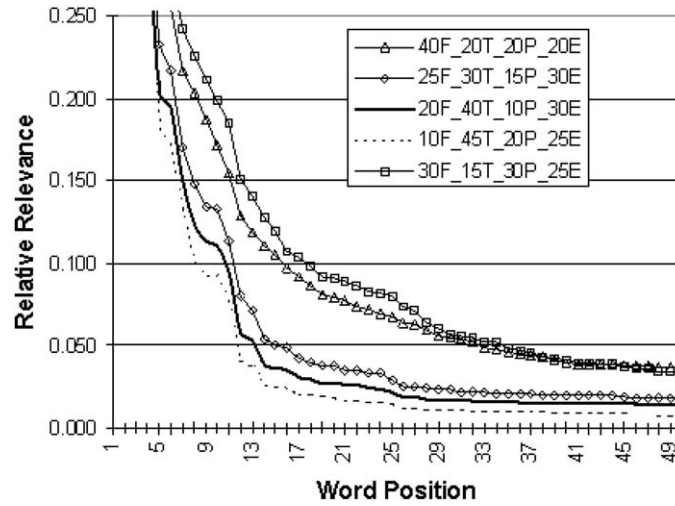


Figure 2. Statistical analysis of the Web pages: The mean distribution of the relative relevance for first fifty components of the feature vector. Results for different coefficient combinations.

$C_e = 0.25$, and $C_p = 0.30$. Then the relevance factor of a word in a page would be expressed as:

$$r_i = 0.3f_f(i) + 0.15f_t(i) + 0.25f_e(i) + 0.3f_p(i) \quad (8)$$

The behavior of the relevance factor with these coefficients is shown in figure 3. In addition, the relevance factor that only considers the frequency (the classical approach) is also represented; this allows comparison of the behavior of the different combination functions

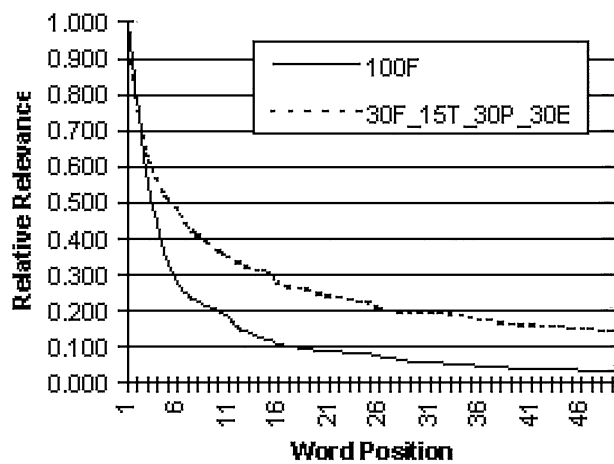


Figure 3. Statistical analysis of the Web pages: The mean distribution of the relative relevance for first fifty components of the feature vector. Propose approached versus classical word frequency approach.

versus a classical approach. An obvious reduction in the feature vector dimensionality is observed when a threshold equal or smaller than 50% is selected.

5. The proposed approach versus a commercial tool

To study the performance of the proposed approach in a concept extraction task, a comparative study of the proposed method versus a commercial product, namely Copernic Summarizer,¹ has been made. Copernic Summarizer is a text-summarizing software tool that also has concept extraction capabilities. It can analyze a text of any length, on any subject, in any one of four languages, and create a summary as short or as long as is needed. It can summarize Web pages among other documents. This commercial product has been selected mainly for two reasons: (a) It works pretty well with various text formats. That means that it uses general text techniques, mainly statistical methods. Therefore better performance by the proposed approach would establish the advantage of using concept extraction methods that take advantage of the HTML environment. (b) It is a widely used product and therefore it is presumably fault-free.

The approach proposed in this paper is the core of a developed tool, namely IAI-Extract,² that can extract the most important concepts from both English and Spanish Web pages for building the feature vector of a Web page. IAI-Extractor gives a concept list where the terms are ranked by relevance. Additionally, to accomplish a more precise comparison, the developed system (IAI-Extractor) generates concepts integrated by more than a single word (two and three words), as does Copernic Summarizer. The combinations of words (terms) are built from adjacent words; this is the reason why the information about the position of the word in a line is saved. The relevance of a term is computed from the weights of the words that form the term. This combination of words allows for obtaining very descriptive concepts such as “data mining”, “artificial intelligence”, etc.

The comparative study has been performed on different sample sets with the intention of adequately covering the heterogeneous nature of the World Wide Web. Sample sets have been composed from Web pages randomly selected without a specific topic and in several sizes. The idea is to compare the performance of both software systems in terms of their ability to deal with Web pages of different size and heterogeneity versus homogeneity. Related to the size, three sample sets have been built: (1) pages with fewer than 50 words; (2) pages with more than 50 but fewer 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/sell, or of auctions have been considered as heterogeneous pages. In summary, Web pages that contain several topics are considered heterogeneous as opposed to Web pages that deal only with one topic, which are considered as homogeneous.

For the comparative evaluation of performance, the discrimination capability of the concept vector has been established through a process that first builds the intersection of words contained in all the Web pages of the training set (Corpus). Second, all the words in the Corpus has been manually classified in the following categories:

A Category formed by the words that clearly differentiate some specific topic. This characteristic is associated with monosemous words (e.g. hotel). These words are clearly

discriminatory words that can help to achieve a correct classification of the Web page.

- B Category shaped by ambiguous words in the sense that they can characterize different topics because they have several meanings (e.g. cycle could mean life cycle or bicycle); i.e. polysemous words (Mannig and Schtze, 2001). Many efforts have been made in word sense disambiguation (Hovy and Lin, 1999). For example, the Bayes approach proposed by Gales et al. (1992) treats the context of occurrence as a bag of words without structure, but it integrates information from many words in the context; it looks at the words around an ambiguous word to extract the sense of the last one. Therefore, an adequate Web classification probably requires the exploration of other words in the page. In this sense, the words that belong to category B can not be very useful in the classification tasks, when they are considered alone, but the situation could radically change when they are analyzed together with words of category A.
- C Category composed of words that do not belong to any topic and obviously are not useful in the classification task (e.g. site, welcome).

This categorization permits the evaluation of discrimination capability of a word when a classification or mining task is conducted using some automated method. As the discrimination capability of a word is directly related to the quantity of meanings of that word. Of course, the classification of a word into these defined types is the result of a subjective evaluation by an expert. Because the expert carries out the evaluation before the software does its task, the expert's evaluation is unbiased and affects the 'goodness of fit' equally for the results produced by either software package.

In both applications, the dimension of the feature vector was fixed at 20 words. For IAI-Extractor, the feature vector was composed of 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 results shown correspond to the average of those obtained for each page of the sample set. The vertical columns of the charts indicate the percentage of words, in the feature vector, that belong to each category. The gray columns represent the results of the IAI-extractor and the black columns the behavior of the Copernic Summarizer. Figure 4 clearly shows that the IAI-Extractor improves over the results achieved by Copernic Summarizer, because the feature vector generated by IAI-Extractor

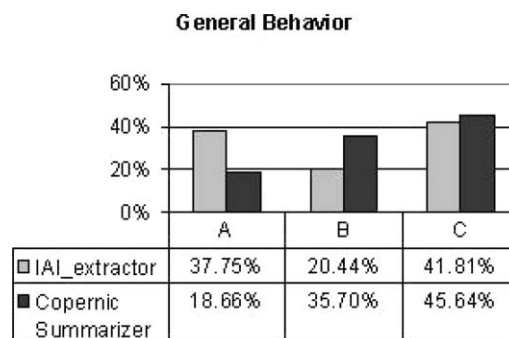


Figure 4. IAI-Extractor vs. Copernic Summarizer: General behaviour.

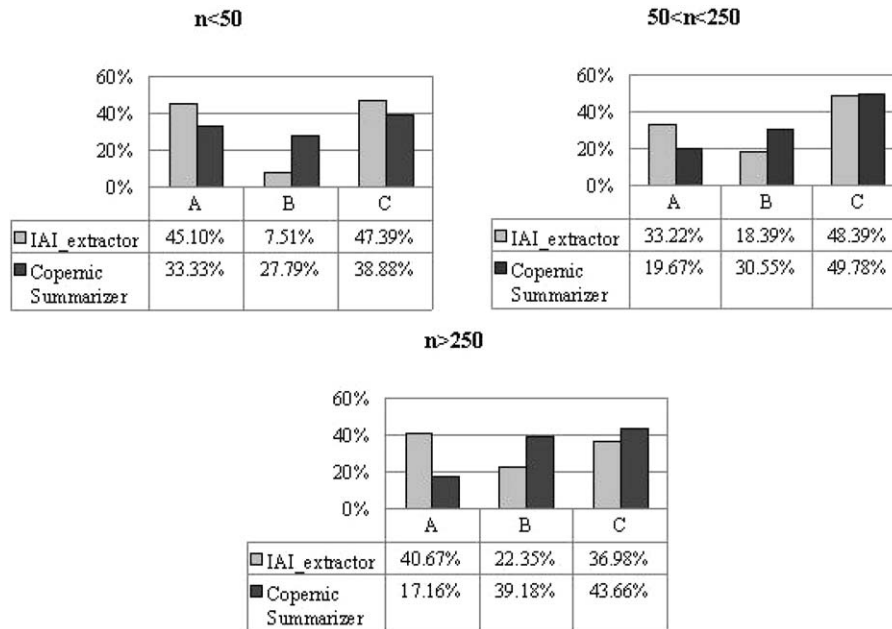


Figure 5. IAI-Extractor vs. Copernic Summarizer: Page size behavior; n is the number of words in a page.

always has more concepts classified in the A category than the feature vector generated by Copernic Summarizer. The result of adding the elements in category A and category B gives the highest result to IAI-Extractor, but with a very slight difference (3.83%) against Copernic Summarizer.

Both tools generated a balanced number of concepts in category C, in fact about 46% for the Copernic Summarizer and about 42% for IAI-Extractor. Next a deeper analysis is performed to highlight other aspects of the behavior of each tool.

In figure 5 the performance of the IAI-Extractor vs. Copernic Summarizer, based on the page size, is shown. IAI-Extractor on average (39.66%) obtains more concepts belonging to the category A than Copernic Summarizer (23.39%). Moreover in Copernic Summarizer, the number of words grouped in A tends to decrease when the page size increases, while IAI-Extractor shows a more stable behavior. The sum of categories A and B behaves differently in IAI-Extractor and Copernic Summarizer. The range of numbers for $A + B$ in figure 5, for $n < 50$, $50 < n < 250$, and $n > 250$, is IAI-Extractor: 52.61%, 51.61% and 63.02% and for Copernic Summarizer: 61.12%, 50.22% and 56.34%. Consequently, it can be derived that the performance of IAI-Extractor increases when the page size increases whilst the performance of Copernic Summarizer seems to decrease if the size page increase.

When the focus is the homogeneity or heterogeneity of the pages, the performance behavior of each tool is shown in figure 6. Here again IAI-Extractor presents more words belonging to A category than Copernic Summarizer. In addition, the performance of Copernic is worse in heterogeneous pages than in homogeneous ones. This behavior persists when the addition of categories A and B is considered. In this case, IAI-Extractor is an improvement over the

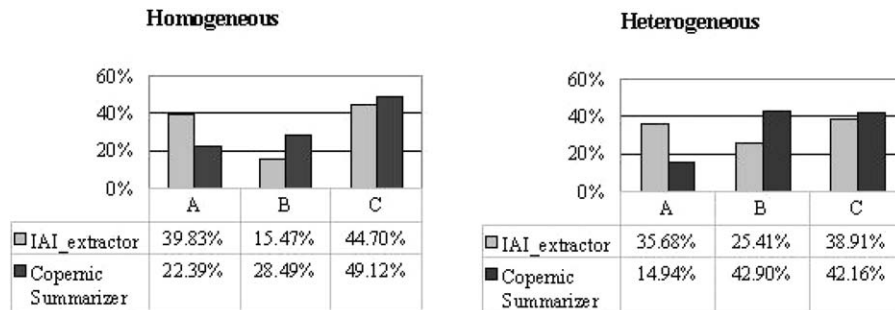


Figure 6. IAI-Extractor vs. Copernic Summarizer: The performance behavior in homogeneous and heterogeneous Web page.

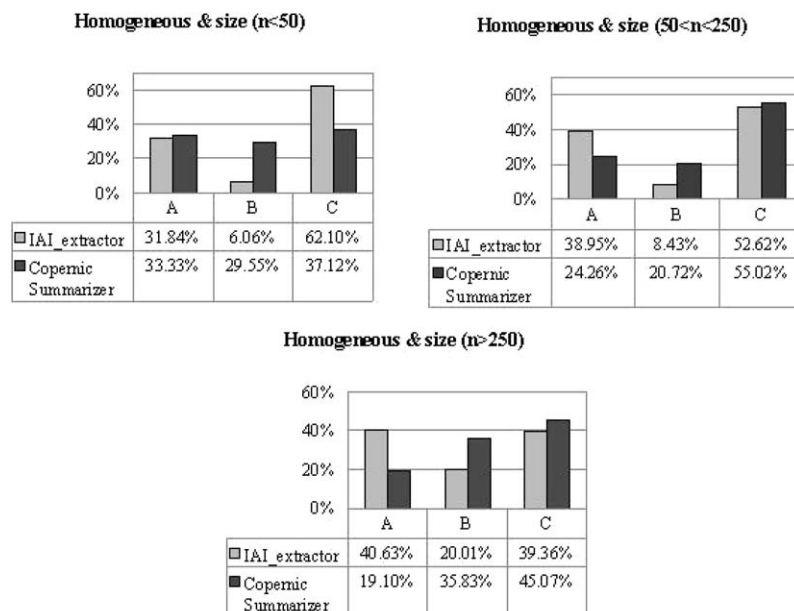


Figure 7. IAI-Extractor vs. Copernic Summarizer: The performance in homogeneous pages for different sizes.

Copernic Summarizer results and works even better with heterogeneous pages than with homogeneous ones.

The performance of each studied tool is displayed in figure 7, focusing on the homogeneity and size aspects of a page together. Copernic Summarizer generates 1.49% more concepts belonging to category A than IAI-Extractor in pages with fewer than 50 words. Moreover, in this kind of page the addition of categories A and B gives better results for Copernic Summarizer (62.88%) than for IAI-Extractor (37.9%). When other page sizes are investigated the situation is inverted and the IAI-Extractor behavior out-performs Copernic Summarizer for both category A and the sum of categories A and B.

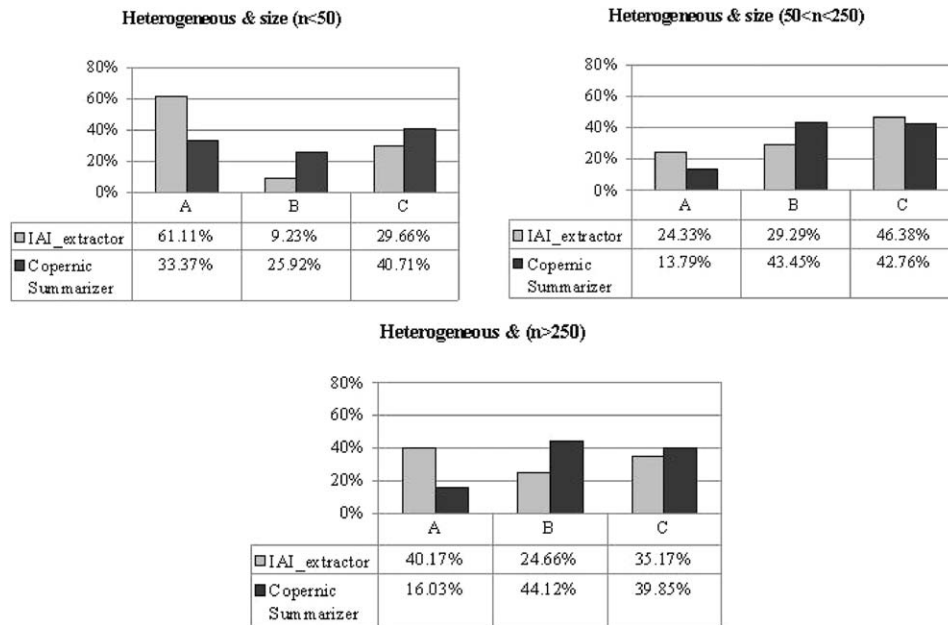


Figure 8. IAI-Extractor vs. Copernic Summarizer: The performance in heterogeneous pages for different sizes.

Finally, figure 8 compares the performance for heterogeneous pages considering only the size of the pages. Here also, IAI-Extractor displays much better accuracy than Copernic Summarizer in pages with a size of fewer than 50 words, with 61.11% of the extracted concepts belonging to category A whereas only 33.37% for Copernic Summarizer. Similar performances can be observed for pages with more words. The great quantity of tags to emphasize some parts of the text, usually present in heterogeneous pages, could explain the performance observed in the experiment.

6. Behavior of the proposed approach in classifications tasks

The objective in this section is to study the performance of the proposed method in a classification task. To achieve this aim the developed IAI-Extractor tool extracts the most relevant concepts in a Web page. From these concepts a representation of a Web page can be obtained and can be used in a supervised learning process to learn the descriptor of a class. In this study two classes: Medicine-Pharmacology and Aerospace-Technology are considered. The objective is to compare the behavior of the fusion of criteria versus the situation where each criterion is separately considered. Notice that the selected classes are extremely different from one another, as within the context of a comparative study, the focus of attention is on the achievements of the proposed methods and therefore both classes are selected to display the most favorable conditions for learning and classification tasks. With this selection it is pursued a lack of overlapping between the classes.

The Web text classification involves a training phase, where the class descriptors are obtained from Web pages with known category labels; followed by a testing phase, where the previously trained classifier is used to categorize new Web pages. In the following, the main aspects of both, the learning and the classification processes are presented, as well as the most relevant results of the proposed tests.

6.1. The learning phase

The class descriptors are obtained from a supervised learning process. Assuming *central limit theorem*,³ a $3 \times n$ matrix represents a class; where the first row are the n words that belong to the class and the second and third one hold the μ and σ parameters that define the *normal* or *Gaussian* distribution.

$$f_i(r_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r_i - \mu)^2 / 2\sigma^2} \quad (9)$$

The density function $f_i(r_i, \mu, \sigma)$ represents the probability that a word i , with relevance, r_i 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; then for each word in each page of the sample set the mean value and the standard deviation can be obtained as

$$\mu_{ij} = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} s_{ijk} \quad (10)$$

$$\sigma_{ij}^2 = \frac{1}{N_{ij}} \sum_{k=1}^{N_{ij}} (s_{ijk} - \mu_{ij})^2 \quad (11)$$

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 page k . That is:

$$s_{ijk} = r_{ik} / R_{\max}^k \quad (12)$$

where R_{\max}^k is the maximum relevance for a word in page k and the word-relevance is derived from a criteria-fusion process, to obtain the relevance function (7). With this normalization, the dependence of relevance on the page size is eliminated.

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

6.2. The classification phase

A Bayesian classifier has been selected as this probabilistic approach is one of the most effective for classification of text documents (Mitchell, 1997). The optimal classification of a new page p_k , is the class c_j for which the probability $P(c_j | p_k)$ shows a maximum. The conditional probability $P(c_j | p_k)$, also referred to as posteriori probability of the class c_j , reflects the confidence that c_j holds, given the page p_k . Then, applying the Bayes's theorem, the following expression has to be maximized:

$$\arg \max_{c_j \in C} \left(\frac{P(p_k | c_j)P(c_j)}{P(p_k)} \right) \quad (13)$$

Here C is the set of all candidate classes. Under the hypothesis of independence among the words in the page and assuming that all pages have the same prior probability, the former expression can be formulated as:

$$\arg \max_{c_j \in C} \left(P(c_j) \prod_{i=1}^{N_k} P(w_i | c_j) \right) \quad (14)$$

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 consider the probability of w_i given c_j with a normalized factor that reflects the significance of the word w_i in the page.

$$\arg \max_{c_j \in C} \left(P(c_j) \prod_{i=1}^{N_k} s_i P(w_i | c_j) \right) \quad (15)$$

Now, the probability that w_i holds given c_j can be expressed as a normal distribution with variance σ^2 and mean μ , as long as the previously defined class representation is used. Then:

$$\arg \max_{c_j \in C} \left(P(c_j) \prod_{i=1}^{N_k} s_i \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2} \right) \quad (16)$$

Rather than maximizing this expression its logarithm is chosen for maximization, because the logarithm of f is a monotonic function of f . Therefore maximizing the logarithm of f also implies the maximization of function f .

$$\arg \max_{c_j \in C} \left(\ln P(c_j) + \sum_{i=1}^{N_k} \ln \left(s_i \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2} \right) \right) \quad (17)$$

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 does 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

$$\arg \max_{c_j \in C} \left(\sum_{i=1}^{N_k} \ln \left(s_i \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2} \right) \right) \quad (18)$$

The logarithmic function $\ln x$ has a discontinuity at $x = 0$. To avoid this difficulty, the function is shifted

$$c_{ML} = \arg \max_{c_j \in C} \left(\sum_{i=1}^{N_k} \ln \left(s_i \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2\sigma_{ij}^2}(s_i - \mu_{ij})^2} + 1 \right) \right) \quad (19)$$

The maximum likelihood hypothesis c_{ML} for classifying a given page p_k can be derived from this final expression.

6.3. Analysis of the results

This section explains the classification 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 by 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.

Sixty-five percent of the sample set was used in the learning phase and the rest (35%) was used in the classification/verification stage. The classification accuracy, when the most relevant concepts of the Web pages are selected by individually considering each criterion or heuristic combinations, is shown in Tables 1 and 2. In these experiments, the following aspects are tested:

- (a) The performance when concept relevance is calculated by the word emphasis function (100E); in other words when $Cf = Ct = Cp = 0$ and $Ce = 100$.
- (b) The classification behavior when the concept relevance is obtained by the frequency function (100F); that is $Ce = Ct = Cp = 0$ and $Cf = 100$.
- (c) The performance in classification for the concept relevance calculated from the frequency function of a word in the title (100T); that is $Ce = Cf = Cp = 0$ and $Ct = 100$.
- (d) The behavior in classification when the relevance of the concepts is computed by the previously defined position function (5); in other words $Ce = Ct = Cf = 0$ and $Cp = 100$. The results are identified in the tables by label 100P.
- (e) The performance when the concept relevance is computed by an alternative position function (6) that weighs the preferential parts twice as highly as the standard ones; that is $a = 2/3$, $b = 1/3$. The results of this option are identified in the tables by label 100P2.

Table 1. Classification accuracy of the proposed approach versus other approximations (Part I).

Classification criteria		Medicine-pharmacology (%)			Aerospace-technology (%)		
		Successes	Faults	Not classified	Successes	Faults	Not classified
100F	DCred	85.84	11.15	3.01	57.03	39.76	3.21
	DC	91.02	6.93	2.05	47.99	48.80	3.21
100T	DCred	25.82	1.01	73.17	25.66	0.00	74.34
	DC	28.68	0.00	71.32	30.63	0.00	69.37
100E	DCred	–	–	–	–	–	–
	DC	39.05	4.88	56.07	29.42	8.92	61.66
100P	DCred	87.83	10.12	2.05	39.27	57.51	3.22
	DC	77.38	21.07	1.55	49.21	47.58	3.21
100P2	DCred	86.73	11.22	2.05	45.99	50.80	3.21
	DC	91.02	6.93	2.05	51.41	45.37	3.22
30F-15T-25E-30P	DCred	87.69	10.26	2.05	70.17	26.62	3.21
	DC	91.88	5.02	3.1	58.25	22.76	18.99
35F-10T-20E-35P	DCred	88.67	9.28	2.05	57.95	38.84	3.21
	DC	90.84	6.07	3.09	70.90	25.89	3.21

Table 2. Classification accuracy of the proposed approach versus other approximations (Part II).

Classification criteria		Average (%)		
		Successes	Faults	Not classified
100F	DCred	72.74	24.02	3.24
	DC	71.99	25.39	2.62
100T	DCred	25.62	0.51	73.87
	DC	29.65	0.00	70.35
100E	DCred	–	–	–
	DC	34.80	7.66	57.54
100P	DCred	66.32	31.05	2.63
	DC	71.95	25.43	2.62
100P2	DCred	68.68	28.70	2.62
	DC	72.88	24.49	2.63
30F-15T-25E-30P	DCred	80.02	17.36	2.62
	DC	80.99	15.83	3.18
35F-10T-20E-35P	DCred	74.70	22.67	2.63
	DC	79.77	17.13	3.10

- (f) The behavior in classification when the concept relevance is estimated from a criteria-combination (8) where $C_f = 30$, $C_t = 15$, $C_e = 25$, and $C_p = 30$; it is labeled 30F-15T-25E-30P.
- (g) Finally the classification performance when the concept relevance is computed by a criteria-combination (7) where $C_f = 35$, $C_t = 10$, $C_e = 20$, and $C_p = 35$; which is labeled 35F-10T-20E-35P.

A complete class descriptor (DC) is obtained for each class through a learning phase. This descriptor includes the intersection of words contained in all the Web pages of the training set (Corpus) whose average relevance (μ) was different from zero. In this description terms/concepts that appear only once or twice in the whole of the documents can be extracted, giving rise to a high-dimension descriptor. Moreover, they are not sufficiently representative to describe the class. As a consequence, a reduced class descriptor (DCred) could be obtained applying a threshold that preserves only the terms with frequencies higher than a minimum. This reduction allows us to handle a smaller vector with a higher statistical consistency.

It is important to be aware that dimensions of DC and DCred are completely different. In the selected test set, the complete class descriptors (DCs) for title (100T) and position (100P) had a size of about 4000 terms, while the reduced descriptors (DCred) was around 60 concepts. These two facts, similar rates of classification success and size reduction, imply that the dimension of the DC could be oversized. Obviously, the classification algorithm requires more computation time when a higher-dimensional class descriptor is used and consequently, to reduce computer time the DCred could be more appropriate to accomplish an effective real world classification.

More conclusions can be extracted if the behavior of each criterion is separately studied. The highest classification rate is achieved when the relevance is computed by the frequency function. The classification accuracy is around 72%, quite similar to that obtained when the position criterion is considered, around 65-70%. These two criteria, frequency and position, share the main property in obtaining the best results: in a web page with textual content all terms have a position and a frequency and therefore these criteria have more information than the others.

The title and emphasis criteria, taken individually, offer poor rates, around 25–35%, in spite of intrinsically containing information about the author's intentions. The sizes of the descriptors obtained by these criteria are smaller compared to those from the position and frequency criteria. This fact is crucial in explaining the poor classification rate when relevance is calculated through the frequency in the title function or in the emphasis function. The situation when the DCred is considered is even worse because of the minimal dimension of the descriptors, producing only one term in some descriptors and none in others. This indicates that the intersection between terms in the Web page and concepts in the class descriptor must not be a null set in order to achieve a good classification performance.

When the criteria-fusion is used to calculate the concept relevance, experiments show better results than the previously analyzed cases. The best of the criteria-combination displays a classification rate around 80% by using the smallest descriptor (DCred).

Focusing on partial results for each class, Aerospace-Technology presents better results than Medicine-Pharmacology. Here again, frequency and position give rise to the best performance, while the emphasis and title produce the worst results. The fact that words in

the Medicine-Pharmacology selected pages are more subclass representative and words in the Aerospace-Technology selected pages are more class representative is able to explain the results. Finally, the Naive-Bayes classification algorithm is a statistical method that returns the most probable class, but sometimes this algorithm is unable to find the most probable class, as is the case where the concepts included in the Web page text do not appear in the training set Corpus.

7. Conclusion

This paper presents an approach for concept extraction from Web pages. The proposed method generates for a Web page an array of two vectors, where the first vector is a collection of concepts extracted from the page content and the second vector is a collection of the corresponding numbers, called relevance factors, which evaluate the appropriateness of the word to represent the text in the Web page. To calculate this second component vector, an analytical function is proposed which takes into account some characteristics of the HTML text as well as some "classical" Information Retrieval criteria such as, the frequency and the position of a word in the text. Next, to achieve good fusion of criteria, a statistical study was performed over a Web page set selected to represent the heterogeneous nature of the World Wide Web. The Web pages were obtained from a major Spanish search engine, where pages are manually categorized. The main conclusions of this statistical analysis can be summarized as follows:

1. Frequency and position criteria can always be defined for each word in the Web page and accordingly their mean value has a higher value than those derived from the other criteria functions. Therefore, their combined contribution is higher than 50% of the total.
2. Title and emphasis criteria are not always present, as a word does not necessarily belong to the title neither does it have to be emphasized. On the other hand, the title introduces a bias when the objective is to obtain a small vector that characterizes the Web page. With automatically generated titles, only 50% of the tested page titles were really representative of the page text, whereas the author consciously defines the emphasized words in the Web page design stage. Consequently, the weight of emphasis criterion should be higher than the weight of the title criterion.

A criteria-combination function has been proposed and diverse experiments and comparative studies have been conducted to analyze the performance of the proposed method. Based on these proposals and studies, a software tool has been implemented, namely IAI-Extractor that extracts the most important concepts from both English and Spanish Web pages.

Experimental investigations and their main results can be summarized as follows:

1. Superior performance of the proposed approach over that of a commercial software product (Copernic Summarizer) has been clearly demonstrated through an analysis that compares discrimination capabilities of the representations generated for a sample set of Web pages by IAI-Extractor versus Copernic Summarizer. The comparative study has

been performed on different sample sets with the intention of adequately covering the heterogeneous nature of the World Wide Web.

2. Text classification has been accomplished by comparing different Web page representations, using different functions to compute the relevance of a concept in a page. These experiments show an improved performance of the proposed fusion approaches compared to other approximations that consider each criterion separately.

Finally, every Web content mining process needs to represent the text part of a Web page. A good and accurate representation will improve the mining tasks. In fact, we have verified this by using the novel proposed approach in a “classical” mining task, a learning process and a posterior classification.

Acknowledgments

Present work was partial supported by the company Innovatec S.A.

Notes

1. <http://www.copernic.com/>
2. IAI-Extractor is a tool, based in the ideas presented in this paper, developed in the Industrial Automation Institute (IAI) under the partial support of Innovatec S.A. company (copyright pending).
3. The mean \bar{X}_n of a random sample from any distribution with finite variance σ^2 and mean μ is approximately distributed as a normal random variable with mean μ and variance σ^2/n .

References

- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley: ACM Press Books.
- Chen, H. and Dumais, S.T. (2000). Bringing Order to the Web: Automatically Categorizing Search Results. In *Proc. Of CHI'00, Human Factor in Computing Systems* (pp. 145–152). Den Haag, New York, US: ACM Press.
- Dunham, M.H. (2002). *Data Mining. Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall.
- Fresno, V. and Ribeiro, A. (2001). Feature Selection and Dimensionality Reduction in Web Pages Representation. In *International ICSC Congress on Computational Intelligence: Methods & Applications* (pp. 416–421). Bangor, Wales, U.K.
- Gales, W., Kenneth, W.C., and Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, 26, 415–439.
- Gudivada, V.N., Raghavan, V.V., Grosky, W.I., and Kasanagottu, R. (1997). Information Retrieval on the World Wide Web, *IEEE Internet Computing*. Sept.–Oct., 58–68.
- Henzinger, M. (2000). Link Analysis in Web Information Retrieval. *Bulletin of the Technical Committee on Data Engineering*, 23, 3–8.
- Hovy, E. and Lin, C.Y. (1999). Automated Text Summarization in SUMMARIST. In I. Mani and M.T. Maybury (Eds.), *Advances in Automatic Text Summarization*. Cambridge, MA: The MIT Press.
- Koller, D. and Sahami, M. (1996). Toward Optimal Feature Selection. In *ICML-96: Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 284–292). San Francisco, CA: Morgan Kaufmann.
- Kosala, R. and Blockeel, H. (2000). Web Mining Research: A Survey. *ACM SIGKDD Explorations*, 2(1), 1–15.
- Manning, C.D. and Schtze, H. (2001). *Foundations of Statistical Natural Language Processing*, Cambridge, MA: The MIT Press.
- Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill International Editions.

- Mladenic, D. (1999). Text-Learning and Related Intelligent Agents. *IEEE Expert* Special issue on Applications of Intelligent Information Retrieval. July–August.
- Musciano, C. and Kennedy, B. (1997). *HTML The Complete Guide*. McGraw Hill.
- Salton, G., Wong, A., and Yang, C.S. (1975). A Vector Space Model for Information Retrieval. *Communications of the ACM* 18(11), 613–620.
- UNCTAD (2002). *E-Commerce and Development Report 2002*. Report of the United Nations Conference on Trade and Development. United Nations, New York and Geneva.
- Yang, Y., Slattery, S., and Ghani, R. (2002). A Study of Approaches to Hypertext Categorization. *Journal of Intelligent Information Systems*, 18(2), 219–241.