

Features selection and dimensionality reduction in Web pages representation

V. Fresno and A. Ribeiro

Instituto de Automática Industrial, CSIC
Carretera Campo Real Km 0,200
La Poveda, Arganda del Rey - Madrid, Spain
{vfresno,angela}@iai.csic.es

Abstract

The rapid and chaotic growth of the World Wide Web has generated a poor level of structure and organisation into the information network. At this point it would be convenient to build more and better tools to extract significant information of the Web in an appropriate way for user. The first step in any task such as automatic summarisation, text classification, information retrieval, information extraction, or text mining is to obtain a data structure for digital processing that represents the text. This paper presents a new method to obtain a significant reduction in the Web pages representation for carrying out both classification and data mining tasks later on. This method is based on a bi-dimensional representation, where the first component is a word extracted from the page and the second one is a number, which evaluates how representative of the page the word is. To show the performance of the proposed representation method, an analysis is introduced from two points of view: a quantitative and a qualitative. These analyses are supported by two evaluation functions and they show an evident reduction in the feature vector dimensionality and an improvement in the quality of the representation.

Introduction

Every day the World Wide Web grows by roughly a million of electronics pages. It is the first time in history that millions of people have virtually instant access from their homes and offices to the output generated by a significant fraction of the planet's population. Additionally the disorganised fast increase of the Web has generated a weakly structured information network. In fact, the Web has evolved into a global mess of never imagined proportions. In this context a question raises, how people obtain useful information from the Web? In fact, more and better tools are needed to extract, in an appropriate way for the user, significant information from the Web.

Any task such as automatic summarisation, text classification, information retrieval, information

extraction, or text mining requires an adequate data representation 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. On the other hand, the success of further on analysis is dependent on having an appropriate text representation that captures the most relevant aspects of the document.

In document retrieval, texts have been traditionally represented in a vector space model [1] [2]. The different methods that have been used so far can be basically grouped in two models. The first one represents, from a fixed word set, a document by a vector of binary attributes indicating which word occurs and 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 word occurrences in itself. Here also, the order of the words is lost, however, the number of occurrences of each word in the document is got. Numerous people have been using this second approach for text classification although it is more traditional in statistical language modelling for speech recognition [3].

Even if these models are very rudimentary, as they do not account for many aspects of the language and of the semantics, frequently work well in classification tasks, although the vector's high dimensionality (of 10^4 to 10^7 components) hinders the use of most of the classic learning algorithms. There are many reasons [4] for this. First, the time requirements for an induction algorithm often grow dramatically with the number of vectors components, also called features. Furthermore, many learning algorithms can be viewed as performing estimation of the probability of the class label given a set of features. In domains with a large number of features, this distribution is very complex and of high dimension. Moreover, many algorithms employ the Occam's Razor bias to prefer the simplest hypothesis that fits the data [5]. Irrelevant and redundant features also cause problems in this context as they may confuse the learning algorithm masking the small set of truly

relevant features. In summary, the appropriated reduction of the feature vector may have two effects: a decrease in the running time of the induction algorithm and an increase in the accuracy of the resulting model.

This paper presents a new method to obtain a significant reduction in the Web pages representation for accomplishing classification and mining tasks later on. The proposed approach takes advantage of some Web page characteristics that do not have other documents, with the aim of obtaining a small set of relevant features, the most representative in the text. As a final result a bidimensional vector is obtained, which represents the Web page, $((w_1, s_1), (w_2, s_2), \dots, (w_n, s_n))$, where each component is composed of two terms: a qualitative term, a extracted word from the text, and a quantitative term that characterises the word relevance into the Web page.

This paper is organised as follow. In next section some important characteristics of Web pages, from the representation point of view, are analysed and a weighting function for each word is proposed. After that, a Web page statistical study is presented with the purpose of finding appropriate coefficients for the weighting function in order to obtain a good and small feature vector for a Web page. Finally, in the last sections, some results are given to show the performance of the proposed algorithm.

Web page structure and feature vector

In the World Wide Web documents are supported by HTML language. Every Web document is built as a combination of tags and text information that Web browsers recognise and visualise.

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 `` and ``) 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. 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 the relevance of the word 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 like `...`, `<u>...</u>`, `...`, `<i>...</i>`, and `...` that allow to emphasise 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 emphasised sentences in the text. However there is an essential difference between one case and another, while to emphasise is an operation consciously made by the user 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.

In addition to these two attributes there are other more "classical issues" that could be considered to compute the word relevance: the *word position* into the text and the *word frequency* in the text.

As previously stated, a bi-dimensional vector represents the Web page, where the first component is a word, a textual feature (qualitative), and the second one corresponds to an associated weight for that feature (quantitative). To calculate the associated weight for each word a lineal combination of previous factors is proposed. In other words, for each word belonging to the Web page text, a weight is calculated by mean of the following expression:

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

Where:

1. $P_f(i) = \frac{n_f(i)}{N_{tot}}$

$n_f(i)$ - Frequency of the word i in the text.

N_{tot} - Number of words in each page

2. $P_t(i) = \frac{n_t(i)}{N_{tit}}$

$n_t(i)$ - Occurrences of the word i in the title

N_{tit} - Number of words in the title

$$3. P_e(i) = \frac{n_e(i)}{N_{enf}}$$

$n_e(i)$ - Times that the word i is emphasised

N_{enf} - Number of emphasised words

$$4. P_p(i) = \left(\frac{3}{4} \times n_{1,4}(i) + \frac{1}{4} \times n_{2,3}(i) \right) / N_{tot}$$

To compute this value the Web page is split in four parts, so that:

$n_{1,4}(i)$ - Occurrences of the word i in the first and the fourth quarter of the page.

$n_{2,3}(i)$ - Occurrences of the word i in the second and third quarter of the page.

N - Number of words in each page

From now on these four functions will be called characterisation functions.

Statistical Analysis

The question now is how to combine the previous function to reduce the dimensionality of the feature vector increasing the quality of the representation. In other words, how to find the best coefficients C_1 , C_2 , C_3 and C_4 . To obtain these coefficients a statistical study has been carried out.

The sample set for the statistical study was selected to represent the heterogeneous nature of the World Wide Web. For this reason a set of Web pages without a specific topic and with divers measures was selected. The pages were chosen from the most important Spanish search engine, where page are classified by a user and not by an automatic process.

Once selected the sample set, it was necessary to pre-process each page in the set in the following way. First, HTML tags were removed and some important information, as which words were empathised, was saved. Second, stop-words were eliminated from pages, where stop-words are words that do not have semantics and they are not important to express the textual content of a page. Articles, prepositions, conjunctions are stop-words. This kind of words represents around 38.8% of the words in a Web page. Last the line number where the word appeared, the position inside the line and the frequency were taken to compute the value of each characterisation function for each word in each page.

Figure 1 shows the mean value distribution of each characterisation function in the first fifty components of the feature vector, for the sample set. Hence four feature vectors are computed for each element in the sample set. Components in each vector are ordered in increasing way and only the first fifty components are considered to calculate the average values for the functions.

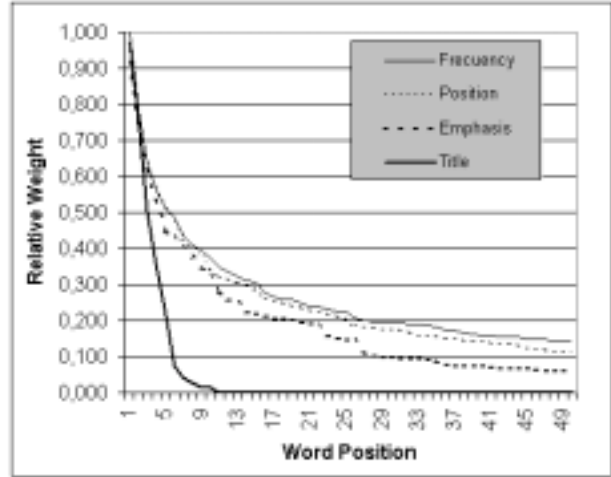


Figure 1: The mean value distribution of each characterisation function

Analysing the graph, individual quantitative characteristics for each function can be extracted. First, functions $P_f(i)$ and $P_p(i)$, frequency and position, always are present and their mean value is higher than the average value in the other functions. The reason way is that all words have contributions to these criteria. All words have a frequency value and a value associate to the position within the text, but a word do not necessarily belongs to the title neither is emphasised. Consequently, the emphasis and title criteria could not only be selected, as they are not always included in all the possible cases. In fact, in the sample set it has been found emphasised words in the 89.30% of the pages and title tags in the 97.05% (but only the 51.97% of the titles were representative titles 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 60% (0.6 in figure 1). But we have found, in the sample set, that only the 51.97% of the titles really were representative of the page content, therefore this criterion should not be used alone.

Alternatively, if the feature vector is reduced taking into account only the components that have a weight value equal or higher than the 10% of the highest weight value

in the vector for each characterisation function, the following average results are found:

Title	6 words
Emphasis	28 words
Position	>50 words
Frequency	>50 words

If the threshold is 20%, the results are:

Title	5 words
Emphasis	18 words
Position	26 words
Frequency	27 words

In figure 2 the behaviour of different weight functions, that are built using the combination of diverse coefficients, is displayed.

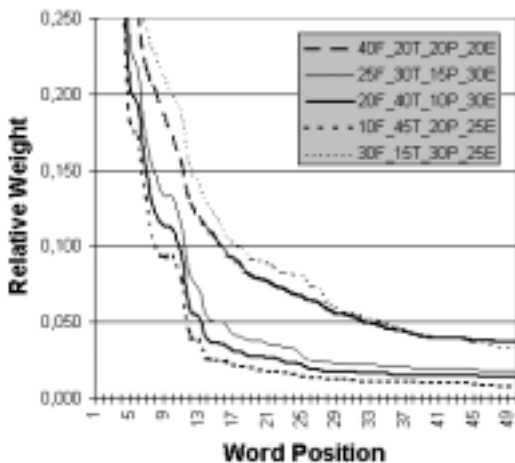


Figure 2: Weight function with different coefficient combinations

By analysing those results it can be seen that the frequency and the position have a similar behaviour. Both criterions contribute to raise the relative weight of each word and consequently, they should have similar contributions in the characteristic functions combination. Furthermore they are present in the 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 and we must take into account this value in the criterion combination. On the other hand it seems that a good combination for the characterisation functions should give more weight to those criteria that generate a greater reduction in the feature vector. But this is only partially true, because the discrimination capability of the qualitative terms in the vector is also important, as

we will see in the following section. With those premises the following linear combination coefficients is proposed:

C ₁ (Frequency)	0.30
C ₂ (Title)	0.15
C ₃ (Position)	0.30
C ₄ (Emphasis)	0.25

The behaviour of the weight function with these coefficients is shown in the figure 3. In addition in the picture the weight function that only considers the

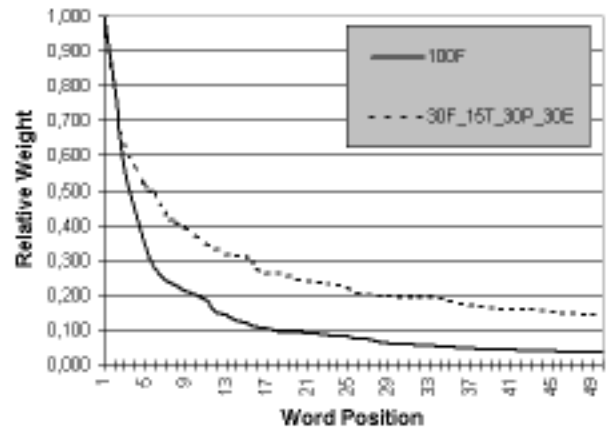


Figure3: The proposed weight function versus classical weight function

frequency is also represented. An obvious reduction in the feature vector dimensionality is observed when a threshold equal or smaller than the 50% is selected.

Results

To show the performance of the proposed representation method an analysis has been done from two points of view: a qualitative study, for the first component of the bi-dimensional vector, and a quantitative study, for the second one. The definition of the benefit of the representation in quantitative and qualitative terms is a major issue, because in this case the representation improvement can not be established by applying for instance a classification method later on. A good representation, from a quantitative view, will be that wherein a small set of words has a relative weight higher than the rest of features in the set. On the other hand, a good representation, from a qualitative point of view, will be fixed by a human expert analysis.

Quantitative Analysis

A function to evaluate the quantitative improvement in terms of an estimation of the reduction obtained with

the proposed method versus the classical representation, where the weight function is the frequency, has been defined. Its expression is as follows:

$$R(i) = \frac{n(i) - n'(i)}{n(i) + n'(i) - 2} \quad \forall i \quad | \quad n(i) \neq n'(i)$$

$$R(i) = 0 \quad \forall i \quad | \quad n(i) = n'(i)$$

Where:

$n(i)$ - Vector dimension in classical representation.

$n'(i)$ - Vector dimension in the proposed representation.

The function takes values between 1 and -1. $R=1$ represents the maximum reduction, that is $n'=1$. And $R=-1$ represents the minimum reduction, that is $n=1$. When $n=n'$ the function is equal to zero value.

The average values for sample set of the reduction function are presented in table 1. This function evaluates if there is reduction on the average between the proposed method and the classic representations.

Qualitative Analysis

A human expert must do all the qualitative evaluation, whereas it is indispensable to estimate if a word has enough discrimination capability for classification and mining tasks.

With the intention of performing the qualitative analysis, words in the sample set have been classified in three categories:

- A** : words which clearly belong to a topic.
- B** : words which belong to various topics.
- C** : words which do not belong to any topic.

The function defined to estimate the quality improvement is as follows:

$$Q(i) = \frac{n_A(i)}{n_A(i) + n_C(i)} \quad \forall i \quad | \quad n_A(i) \neq n_C(i)$$

$$Q(i) = 0.5 \quad \forall i \quad | \quad n_A(i) = n_C(i)$$

Where:

$n_A(i)$ - Number of words that belong to A category in the features vector i .

$n_C(i)$ - Number of words that belong to the C category in the features vector i .

As it can be observed, the defined qualitative evaluation function only takes into account A and C categories, because they are well delimited set. The B set contains words, which in some cases can improve the classification and in others they worsen it, therefore it should not be kept in mind. The function takes values between 0 and 1. $Q(i)$ is equally to 1 if all the words that compose the features vector belong to category A. $Q(i)$ is equally to 0 if all words belong to category C, and finally $Q(i)$ is equally to 0.5 if the words in the features vector belong to both categories in the same proportion.

With the help of this function, the quality evaluation of the classical representation versus the proposed one can be done. The obtained average results, for the sample set, are displayed in table 1.

Threshold	Reduction (R)	Classic Represent. (Q)	Proposed Represent. (Q)
10%	0.5767	0.1926	0.3382
20%	0.6078	0.2409	0.4252

Table 1

The results show a clear increase in the representation quality of the features vector that has been computed with the proposed method.

Summary, Conclusions and Future work

This paper presents a new approach to the Web page representation. The method proposes 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 in account some characteristics of the HTML text. A clear improvement of the proposed representation versus classical one has been showed through the definition of a function that evaluates the obtained reduction and another one that calculates the representation quality. Therefore, this paper describes two evaluation functions that could be used to test the performance of other methods.

Future work will be conducted in three different ways. First, a classification algorithm will be developed based on this representation. Second, the sample set will be

increased. Last, a learning algorithm will be applied in order to find the optimal coefficients C_1 , C_2 , C_3 and C_4 .

Acknowledge

Present work was fully supported by Innovatec S.A.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. "Modern Information Retrieval". ACM Press Books, Addison-Wesley. 1999.
2. Dunja Mladenic, "Text-Learning and related intelligent agents" Revised version in IEEE Expert special issue on Applications of Intelligent Information Retrieval, July-August 1999
3. McCallum and K. Nigam. "A Comparison of Event Models for Naive Bayes Text Classification". In Proceedings of Workshop on Learning for Text Categorization, AAAI/ICML-98. AAAI Press, pp. 41-48. 1998.
4. D. Koller y M. Sahami. "Toward Optimal Feature Selection". International Conference on Machine Learning. Editor L. Saitta. Volumen 13, Morgan-Kaufmann. 1996.
5. T. M. Mitchell. "Machine Learning". McGraw-Hill International Editions. 1997.
6. C. Musciano and B. Kennedy. "HTML The Complete Guide". McGraw Hill. 1997.