

Text Content Approaches in Web Content Mining

Víctor Fresno Fernández*

Departamento de Informática, Estadística y Telemática
Escuela Superior de Ciencias Experimentales y Tecnología
Universidad Rey Juan Carlos
c/ Tulipán s/n
28933, Móstoles
Madrid, Spain
voice: +34 91 488 7180
fax: +34 91 488 7049
email: v.fresno@escet.urjc.es

Luis Magdalena Layos

Departamento de Matemática Aplicada a las TT. II.
ETSI. Telecomunicación
Universidad Politécnica de Madrid
Ciudad Universitaria
28040, Madrid
Spain
voice: +34 336 7287
fax: +34 91 336 7289
email: llayos@mat.upm.es

(* Corresponding author)

Text Content Approaches in Web Content Mining*

Víctor Fresno Fernández**, Rey Juan Carlos University, Spain

Luis Magdalena Layos, Technical University of Madrid, Spain

INTRODUCTION

Since the creation of the Web until now, Internet has become the greatest source of information available in the world. The Web is defined as a global information system that connects several sources of information by hyperlinks, providing a simple media to publish electronic information and being available to all the connected people.

In this context, data mining researchers have a fertile area to develop different systems, using Internet as a knowledge base or personalizing web information. The combination of Internet and Data Mining has been typically referred as Web Mining, defined by Kosala and Blockeel (2000) as “a converging research area from several research communities, such as DataBase (DB), Information Retrieval (IR) and Artificial Intelligent (AI), especially from machine learning and Natural Language Processing (NLP)”.

Web Mining is the use of data mining techniques to automatically discover and extract information from web documents and services; traditionally focused in three distinct ways, based on which part of the Web to mine: *web content*, *web structure* and *web usage*. Brief descriptions of these categories are summarized below.

Web Content Mining - The Web contents consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks. Web Content Mining describes the process of information discovery from millions of sources across the World Wide Web. From an IR point of view, web sites consist of collections of hypertext documents for unstructured documents (Turney, 2002); and from a DB point of view, web sites consist of collections of semi-structured documents, (Jeh & Widom, 2004).

Web Structure Mining - This approach is interested in the structure of the hyperlinks within the Web itself: the inter-document structure. The Web structure is inspired by the study of social network and citation analysis (Chakrabarti, 2002). Some algorithms have been proposed to model the Web topology such as PageRank (Brin & Page, 1998) from Google and other approaches that add content information to the link structure (Getoor, 2003).

Web Usage Mining - Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. A first approach, maps the usage data of the Web server into relational tables for a later analysis. A second approach uses the log data directly by using special pre-processing techniques (Borges & Levene, 2004).

BACKGROUND

In the Web there are not standards or style rules; the contents are created by a set of very heterogeneous people in an autonomous way. In this sense, the Web can be seen as a huge amount of online unstructured information. Due to this inherent chaos, it has emerged the necessity of developing systems that aid us in the processes of searching and efficient accessing to information.

When we want to find information in the Web, we usually access to it by search services, such as Google (<http://www.google.com>) or AllTheWeb (<http://www.alltheweb.com>), which return a ranked list of web pages in response to our request. A recent study (Gonzalo, 2004) showed that

this way to find information works well when we want to retrieve homepages, websites related to corporations, institutions or specific events, and to find quality portals. However, when we want to explore several pages, relating information from several sources, this way has some lacks: the ranked lists are not conceptually ordered and information in different sources is not related. The Google model has the following features: crawling the Web, the application of a simple Boolean search, the PageRank algorithm and an efficient implementation. This model directs us to a web page and then we are abandoned with the local server search tools once the page is reached. Nowadays, these tools are very simple and the search results are poor.

Other way to find information is using web directories organized by categories, such as Yahoo (<http://www.yahoo.com>) or Open Directory Project (<http://www.dmoz.org>). However, the manual nature of this categorization makes too arduous the directories maintenance if machine processes do not assist it.

Future and present research tends to the visualization and organization of results, the information extraction over the retrieved pages or the development of efficient local servers search tools. Below, we summarize some of the technologies that can be explored in Web Content Mining and a brief description of their main features.

Web Mining and Information Retrieval

These systems retrieve contents, as much text as multimedia, and the main feature is that the access to information is accomplished in response to an user's request (Wang et al., 2003; Fan et al., 2004). Techniques inherited from NLP are added to these systems.

Text Categorization in the web

The main goal of these methods is to find the nearest category, from a pre-classified categories hierarchy to a specific web page content. Some relevant works in this approach can be found in (Chakrabarti, 2003; Kwon & Lee, 2003).

Web Document Clustering

Clustering involves dividing a set of n documents into a specific number of clusters k , so that some documents are similar to other documents into the same cluster, and different from those in other clusters. Some examples in this context are (Carey et al., 2003; Liu et al., 2002).

MAIN THRUST OF THE CHAPTER

In general, web mining systems can be decomposed into different stages which can be grouped in four main phases: *Resource Access*, the task of capturing intended web documents; *Information-Preprocessing*, the automatic selection of specific information from the captured resources; *Generalization*, where machine learning or data mining processes discover general patterns in individual web pages or across multiple sites; and finally the *Analysis phase*, or validation and interpretation of the mined patterns. We think that improving each of the phases, the final system behaviour can also be improved.

In this work we focus our efforts to web pages representation, which can be associated to the information-preprocessing phase in a general web mining system. Several hypertext representations have been introduced in literature, in different web mining categories, and they will depend on the later use and application that will be given. Here, we restrict our analysis to web content mining and in addition, hyperlinks and multimedia data are not considered. The main reason to select only the tagged text is to look for the existence of special features emerging from the HTML tags with the aim to develop web content mining systems with greater scope and better

performance as local server search tools. In this case, the representation of web pages is similar to the representation of any text.

A model of text must build a machine representation of the world knowledge and must, therefore, involve a natural language grammar. Since we restrict our scope to statistical analyses for web page classification, we need to find suitable representations for hypertext that will suffice for our learning applications.

We carry out a comparison between different representations using the vector space model (Salton et al., 1975), where documents are tokenized using simple rules, such as whitespace delimiters in English, and tokens stemmed to canonical form (eg. 'reading' to 'read'). Each canonical token represents an axis in the Euclidean space. This representation ignores the sequence in which words occur and is based on the statistical about single independent words. This Independence Principle between the words that co-appear in a text, or appear as multiword terms, is a certain error but reduce the complexity of our problem without loss of efficiency. The different representations are obtained using different functions to assign the value of each component in the vector representation. We used a subset of the BankSearch Dataset as the web document collection (Sinka & Corne, 2002).

First, we obtained five representations using well-known functions in the IR environment. All these are only based on the term frequency in the web page that we want to represent, and on the term frequency in the pages of the collection. Below, we summarize the different evaluated representations and a brief explanation.

(1) *Binary* - This is the most straightforward model, which is called “set of words”. The relevance or weight of a feature is a binary value $\{0,1\}$ depending on whether the feature appears in the document or not.

(2) *Term Frequency (TF)* - Each term is assumed to have an importance proportional to the number of times it occurs in the text (Luhn, 1957). The weight of a term t in a document d is given by: $W(d;t)=TF(d;t)$; where $TF(d;t)$ is the term frequency of the term t in d .

(3) *Inverse Document Frequency (IDF)* - The importance of each term is assumed to be inversely proportional to the number of documents that contain the term. The IDF factor of a term t is given by: $IDF(t)=\log Nxdf(t)$; where N is the number of documents in the collection and $df(t)$ is the number of documents that contain the term t .

(4) *TF-IDF* - Salton (1988) proposed to combine TF and IDF to weight terms. Then, the TF-IDF weight of a term t in a document d is given by: $W(d;t)=TF(d;t) \times IDF(t)$.

(5) *WIDF* - It is an extension of IDF to incorporate the term frequency over the collection of documents. The WIDF weight is given by: $W(d,t)=TF(d,t)/\sum diTF(i,t)$.

In addition to the five representations, we obtained other two representations which combine several criteria extracted from some tagged text and that can be treated differently from other parts of the web page document. Both representations consider more elements than the term frequency to obtain the term relevance in the web page content. These two representations are: the *Analytical Combination of Criteria (ACC)* and *Fuzzy Combination of Criteria (FCC)*. The difference between them is the way they evaluate and combine the criteria. The first one (Fresno & Ribeiro, 2004) uses a lineal combination of those criteria, whereas the second one (Ribeiro et al., 2002) combines them by using a fuzzy system. A fuzzy reasoning system is a suitable framework to capture the qualitative human expert knowledge to solve the ambiguity inherent to the current reasoning process, embodying knowledge and expertise in a set of linguistic expressions that manage words instead of numerical values. The fundamental cue is that often a criterion evaluates the importance of a word only when it appears combined with another criterion. Some web pages representation methods that use HTML tags in different ways can be found in (Molinari et al., 2003; Pierre, 2001; Yang et al., 2002). The combined criteria in ACC and FCC are summarized below.

(i) *Word Frequency in the text* - Luhn (1957) showed that a statistical analysis of the words in the document provides some clues of its contents. This is the most used heuristic in the text representation field.

(ii) *Words Appearance in the text* - The words appearance in the title of the web page, considering that in many cases the document title can be a summary about the content.

(iii) *The positions all along the text* - In automatic text summarization, a well-known heuristic to extract sentences that contain important information to the summary is selecting those that appear at the beginning and the end in the document (Edmunson, 1969).

(iv) *Words appearance in emphasis tags* - Whether or not the word appears into emphasis tags. For this criterion, several HTML tags were selected because they capture the author's intention. The hypothesis is that if a word appears emphasized is because the author wants to stand out.

To compare the quality of the representations, a web page binary classification system was implemented in three stages: *Representation*, *Learning* and *Classification* algorithm. The selected classes are very different one from other to display favorable conditions for learning and classification stages and to show clearly the achievements of the different representation methods.

The representation stage was achieved as follows. The Corpus, a set of documents that generate the vocabulary, was created from 700 pages for each selected classes. All the different stemmed words, found in these documents, generated the vocabulary as axes in the Euclidean space. We fixed the maximum length of a stemmed word to 30 characters and the minimum length to 3 characters. In order to calculate the values of the vector components for each document we followed the next steps: *a)* we eliminated all the punctuation marks except some special marks that are used in URLs, e-mail address, and multiword terms; *b)* the words in a stoplist used in IR were eliminated from the web pages; *c)* we obtained the stem of each term by using the well known Porter's stemming algorithm; *d)* we counted the number of times that each term appeared in each web page, and the number of pages where the term was present; and *e)* in order to calculate the

ACC and FCC representations, we memorized the position of each term all along the web page and whether or not the feature appears into emphasis and title tags. In addition, other 300 pages for each class were represented in the same vocabulary to evaluate the system.

In the learning stage, the class descriptors (information common to a particular class, but extrinsic to an instance of that class) were obtained from a supervised learning process. Considering the central limit theorem, the word relevance (the value of each component in the vector representation) in the text content for each class will be distributed as a Gaussian function with the parameters mean μ and variance σ . Then, the density function:

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

gets the probability that a word i , with relevance r , appears in a class (Fresno & Ribeiro, 2004). The mean and variance are obtained from the two selected sets of examples for each class by a maximum likelihood estimator method.

Once the learning stage is achieved, a Bayesian classification process is carried out to test the performance of the obtained class descriptors. The optimal classification of a new page d is the class $c_j \in C$ where the probability $P(c_j | d)$ is maximum, where C is the set of considered classes.

$P(c_j | d)$ reflects the confidence that c_j holds given the page d . Then, the Bayes' theorem states:

$$P(c_j | d) = \frac{P(d | c_j)P(c_j)}{P(d)}$$

Considering the hypothesis of the independence principle, assuming that all pages and classes have the same prior probability, applying logarithms because it is a non-decreasing monotonic function and shifting the argument one unit to avoid the discontinuity in $x=0$, finally, the Most Likelihood class is given by:

$$c_{ML} = \underset{c_j \in C}{\operatorname{argmax}} \left(\sum_i^N \ln \left(\frac{r_i}{\sqrt{2\pi \sigma_j^2}} e^{-\frac{1}{2\sigma_j^2}(r_i - \mu_j)^2} + 1 \right) \right)$$

where N is the vocabulary dimension. We accomplished an external evaluation by means of the *F-measure*, which combines the *Precision* and *Recall* measures:

$$F(i,j) = 2 \times \text{Recall}(i,j) \times \text{Precision}(i,j) / (\text{Precision}(i,j) + \text{Recall}(i,j))$$

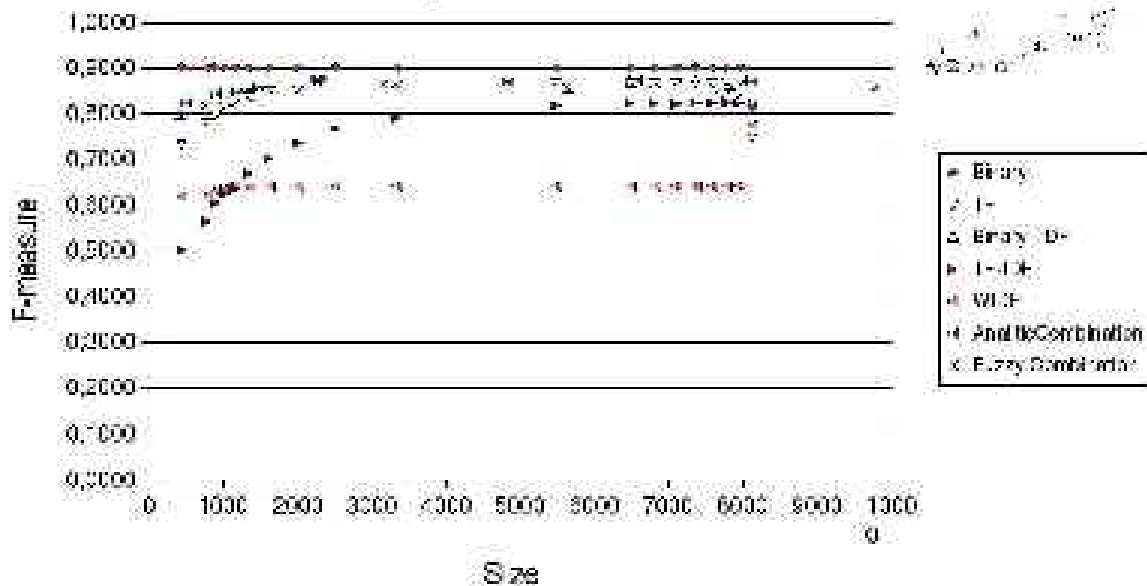
$$\text{Recall}(i,j) = n_{ij} / n_i$$

$$\text{Precision}(i,j) = n_{ij} / n_j$$

where n_{ij} is the number of pages of class i classified as j , n_j is the number of pages classified as j , and n_i the number of pages of the i class. To obtain the different representation sizes, reductions were carried out by the document frequency term selection (Sebastiani, 2002) in binary, TF, binary-IDF, TFIDF and WIDF representations; and thus, for ACC and FCC we have used the proper weighting function of each one as a reduction function, selecting the n most relevant features in each web page. In Figure 1 we show the obtained experimental results in a binary classification, with each representation and with different representation sizes.

Figure 1: Comparison between representations in a binary classification.

Binary Classification



FUTURE TRENDS

Nowadays, the main lack in systems that aid us to the search and access to information process is revealed when we want to explore several pages, relating information from several sources. Future trends must find regularities in HTML vocabulary to improve the response of the local server search tools, combining with other aspects such as hyperlink regularities.

CONCLUSION

The results in a web page representation comparison are very dependent on the selected collection and the classification process. In a binary classification with the proposed learning and classification algorithms, the best representation was the binary because it obtained the best F-measures in all the cases. We can expect that when the classes number will be increased this F-measure values will decrease and the rest of representations will increase their global results.

Finally, apart from binary function, the ACC and FCC representations have best F-measure values than the rest, inherited from the IR field, when the sizes are the smallest. This fact can be result of considering the tagged text in a different way, depending on the tag semantic, and capturing most information than when only the frequency is considered. A most deep exploration must be accomplished to find hidden information behind this Hypertext Markup Language vocabulary.

REFERENCES

- Borges, J. & Levene, M. (2004). An average linear time algorithm for web data mining. To appear in the *International Journal of Information Technology and Decision Making*, 3.
- Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7),107-117.
- Carey, M., Heesch, D. & Rüger, S. (2003). Info Navigator: A visualization tool for document searching and browsing. *International Conference on Distributed Multimedia Systems*
- Chakrabarti S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kaufmann Publishers.
- Chakrabarti S., Roy, S. & Soundalgekar, M. (2003). Fast and accurate text classification via multiple linear discriminant projections. *VLDB Journal*, 12(2),170-185.
- Edmunson, H.(1969). New methods in automatic extracting. *Journal of the ACM*16(2),264-285.
- Fan, W., Fox, E.A., Pathak, P. & Wu, H. (2004). The effects of fitness functions on genetic programming-based ranking discovery for web search. *Journal of the American Society for Information Science and Technology*, 55(7),628-636.
- Fresno, V. & Ribeiro, A. (2004). An Analytical Approach to Concept Extraction in HTML Environments. *Journal of Intelligent Information Systems* 22(3).215-235.
- Getoor, L. (2003). Link Mining: A New Data Mining Challenge. *ACM SIGKDD Explorations Newsletter*, 5(1),84-89.

- Gonzalo, J. (2004). Hay vida después de Google?. *Software and Computing System seminars*. ESCET, URJC. <<http://sensei.lsi.uned.es/~julio/>>.
- Jeh, G. & Widom, J. (2004). Mining the Space of Graph Properties. *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*
- Kosala & H. Blockeel. (2000). Web mining research: A Survey. *ACM SIGKDD Explorations Newsletter*, 2(1),1-15.
- Kwon, O. & Lee, J. (2003). Text categorization based on k-nearest neighbor approach for web site classification. *Information Processing and Management: an International Journal*, 39(1), 25-44.
- Liu, B., Zhao, K. & Yi, L. (2002). Visualizing Web Site Comparisons. *11th International Conference on World Wide Web*.
- Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* (4),309-317.
- Molinari, A., Pasi, G. & Marques, R.A. (2003). An indexing model of HTML documents. *ACM Symposium on Applied Computing*.
- Pierre, J.M. (2001). On the Automated Classification of Web Sites. *Linköping Electronic Articles in Computer and Information Science*, 6(1).
- Ribeiro, A., Fresno, V., García-Alegre, M. & Guinea, D. (2003). A Fuzzy System for the Web page Representation. *Intelligent Exploration of the Web, Series: Studies in Fuzzyness and Soft Computing*, 111, 19-38.
- Salton, G., Wong, A. & Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11),613-620.
- Salton, G. (1988). *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley Editors.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), 1-47.

Sinka, M.P., & Corne, D.W. (2002). A Large Benchmark Dataset for Web Document Clustering. *2nd Hybrid Intelligent Systems Conference*.

Turney P. (2002). Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labeled and Unlabeled Data. *NRC Technical Report ERB-1096*, Institute for Information Technology, National Research Council Canada.

Yang, Y., Slattery, S. & Ghani, R. (2002). A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18(2).

TERMS AND DEFINITIONS

Central Limit Theorem: When an infinite number of successive random samples are taken from a population, the distribution of sample means calculated for each sample will become approximately normally distributed with mean μ and standard deviation σ/\sqrt{N} ($\sim N(\mu, \sigma/\sqrt{N})$).

Crawler: Program that downloads and stores web pages. A crawler starts off with the Uniform Resource Locator (URL) for an initial page and extracts any URLs in it, adds them to a queue to scan recursively.

Information Retrieval (IR): Interdisciplinary science of searching for information, given an user query, in document repositories. The emphasis is on the retrieval of information as opposed to the retrieval of data.

Machine Learning: The study of computer algorithms that improve automatically through experience.

Natural Language Processing (NLP): Computer understanding, analysis, manipulation, and/or generation of natural language. This can refer to simple string-manipulation like [stemming](#) or higher-level tasks such as processing user queries in [natural language](#).

Stoplist: Specific collection of so-called ‘noise’ words, which tend to appear frequently in documents.

Supervised Learning: A machine learning technique for creating a function from training data. The training data consists of pairs of input objects and desired outputs.

Unsupervised Learning: A machine learning technique that typically treats input objects as a set of random variables. A joint density model is then built for the data set.

* This work was supported by the Madrid Research Agency, under project 07T/0030/2003-1.

** Author want to thank to Raquel Martínez Unanue for suggesting many valuable new ideas and Rodrigo Montúfar for its review.