

# Improving Web Page Clustering Through Selecting Appropriate Term Weighting Functions

Víctor Fresno

Escuela Superior de Ciencias  
Experimentales y Tecnología  
Universidad Rey Juan Carlos  
28933, Móstoles (Madrid), Spain  
Email: victor.fresno@urjc.es

Raquel Martínez

Escuela Técnica Superior de  
Ingeniería Informática  
U.N.E.D.  
28040, Madrid, Spain  
Email: raquel@lsi.uned.es

Soto Montalvo

Escuela Superior de Ciencias  
Experimentales y Tecnología  
Universidad Rey Juan Carlos  
28933, Móstoles (Madrid), Spain  
Email: soto.montalvo@urjc.es

**Abstract**— Web page clustering is useful for taxonomy design, information extraction, similarity search, search results visualization, and it can assist to the evaluation of the results of search engines. Therefore, an accurate clustering is a goal in web mining and web information extraction. Besides the particular clustering algorithm, the different term weighting functions applied to the selected features to represent web pages is a main aspect in clustering task. This paper presents the evaluation of the performance of seven different features weighting functions of web pages, by means of a partitioning clustering algorithm results. Five of them are well-known term weighting functions from text content analysis: Binary, Term Frequency (TF), Inverse Document Frequency (Binary-IDF), TF-IDF, and WIDF; the other two are based on a heuristic criteria combination, which consider HTML mark-up information: ACC and FCC. The first of these two later combinations is linear, whereas the other uses a fuzzy system. Besides, two reduction classes have been applied: (1) the proper function, and (2) removing all features occurring more times than upper thresholds in page and collection, and occurring less times than lower thresholds in page and collection. By means of the experimentation with a collection of web documents utilized in clustering research, we have determined that the best results are obtained when the term weighting function based on a fuzzy criteria combination is used.

## I. INTRODUCTION

Internet is a great source of information in digital format whose increase is unstoppable. The usefulness of this resource depends on the ability of the tools extracting information, and discovering previously unknown relationships among the data of the web pages. The later, is the main objective of web mining. Web page clustering is useful for task such as: taxonomy design, similarity search, and search results visualization. Therefore, an accurate clustering is a goal in web mining ([30], [4], [8]).

The web information usually is acceded by search engines and by thematic web directories. Search engines, as *Google*<sup>1</sup>, return to us a sorted list which is not conceptually sorted and it does not connect information extracted from several web pages. Nevertheless, there are search engines, for example *Vivísimo*<sup>2</sup>, which besides the list of relevant documents they show us a cluster hierarchy. When thematic web directories

are used, the documents are showed classified in taxonomies and the search process uses that taxonomy.

In this context, the document clustering algorithms are very useful to apply to tasks such as: automatic grouping before and after the search, search by similarity, and search results visualization on a structured way ([16], [6], [29], and [11]).

Two aspects are very important in order to obtain good web page clustering results: the clustering algorithm, and the term weighting function applied to the selected features of the web pages. In this paper, we fix the clustering algorithm, and we focus on how term weighting functions influence clustering results.

This paper evaluates the adaptation of seven different feature weighting functions to web page clustering, applying two classes of feature reduction. It continues a previous work described in [1] in which neither such reductions nor the clustering performance for different number of features were studied.

The evaluation is carried out by means of two web collections created from the Benchmark Dataset [24]. One of our aims is to determine whether the use of HTML mark-up information as part of the representation of a web page can improve clustering results. Although most HTML tags refer to visualization (except those for metadata), some can be used to calculate the relevance of the textual elements concerned.

Five of the seven term weighting functions are based exclusively on textual information: Binary, Term Frequency, Binary Frequency-Inverse Document Frequency, Term Frequency-Inverse Document Frequency, Weighted Inverse Document Frequency; so no information from HTML mark-up is taken into account. However, two of them: the analytical combination of criteria and the fuzzy combination of criteria, use information from textual HTML tags. These term weight functions, (such as that presented in [14], [15] and [28]), take into account some HTML tags, which can reflect the author's intention in connection to the importance of some words (for instance, to emphasize a word). The combination of criteria is a heuristic selection of textual HTML elements, which are combined to calculate the relevance of each feature of a web page.

The remainder of the paper is organized as follows: Sec-

<sup>1</sup><http://www.google.com>

<sup>2</sup><http://vivisimo.com>

tion 2 summarizes the main approaches in web document representation. Section 3 shows the different term weighting functions we study. Section 4 describes the collection used for evaluation. Section 5 presents the two types of feature reduction, the experiments, the clustering algorithm, the evaluation method, and the results. Finally, Section 6 includes the main conclusions.

## II. WEB PAGE REPRESENTATION

The representation of a web page can be approached by different points of view, depending on which elements are taken into account. Some of the elements which can be distinguished in a web document are: plain text, text enriched with HTML tags, metacontent, text associated with hyperlinks, set of statistical parameters (media types, size, number of images or links, etc.). Thus, the representation of a web page can be defined according to these (and other) elements. The elements can be taken into account separately or can be combined. The studies in web page representation have focused mainly on two approaches: context and content.

In representation by context, the underlying information of the links is explored, and multimedia components are treated [2], [5], [7]. Some studies analyze the Universal Resource Locations (URLs), for instance [13]. In this case, the main aim is to extract information about the document structure. The analysis of meta-tags is also used as a preliminary task prior to representation. However, in [17] a study is carried out that shows the number of web pages with meta-tags is no more than 30%, so this approach is not applicable to most of the web pages.

In representation by content, information is extracted from text and neither document structure nor topology is studied. In this group we can find representations based on concept identification, exploring techniques in Neural Networks or semantic and linguistic analysis [27].

This work is focused on web page representation by text content. Thus, the features which represent web pages will be extracted from their text. In addition, some information from HTML tags will be taken into account. We evaluate five feature weighting functions of the features based solely on the plain text of the web page, and two more which, in addition to plain text, use information from HTML tags (emphasis and the “title” tags) and feature position.

## III. TERM WEIGHTING FUNCTIONS

We represent web pages by using the vector space model [21]. In this model, each web page is represented by means of a vector, where each component is the weight of a feature in the page. In this context, different representations are obtained by using different functions to assign the weight of the features.

If only the plain text is taken into account, the representation of web pages is like the representation of any other text. In this case, the simplest way to represent a web page is to select a set of words from the text as features. Next, a term weighting function calculates the relevance of each feature in each web

page. If a web page consists of frames, each of them will be considered like one page.

First, we used five well-known term weighting functions which are only based on the text plain of the web pages. These functions are:

- **Binary (B)**. This is the most straightforward model, called the “set of words” or binary vector space model. The relevance or weight of a feature is a binary value  $\{0,1\}$  depending on whether the feature is in the document or not. Then, the weight of a feature  $t$  in a document  $d$  is given by:

$$W(d, t) = B(d, t) \quad (1)$$

- **Term Frequency (TF)**. Each term or feature is assumed to have importance proportional to the number of times it occurs in the document [12]. The weight of a feature  $t$  in a document  $d$  is given by:

$$W(d, t) = TF(d, t) \quad (2)$$

where  $TF(d, t)$  is the feature frequency of the feature  $t$  in the document  $d$ .

- **Inverse Document Frequency (Binary-IDF)**. This looks at feature occurrence across a collection of documents. The importance of each feature is assumed to be inversely proportional to the number of documents that contain the feature [9]. The IDF factor of a feature  $t$  is given by:

$$IDF(t) = \log \frac{N}{df(t)} \quad (3)$$

where  $N$  is the number of documents in the collection and  $df(t)$  is the number of documents that contain the feature  $t$  (this definition follows [22]). Then, the weight of a feature in B-IDF representation is:

$$W(d, t) = B(d, t) \times IDF(t) \quad (4)$$

where  $B(d, t) = \{0, 1\}$ .

- **TF-IDF**. In [20] Salton proposes combining TF and IDF to weight terms. The combination weight of a feature  $t$  in a document  $d$  is given by:

$$W(d, t) = TF(d, t) \times IDF(t) \quad (5)$$

- **WIDF**. This is an extension of IDF to incorporate the feature frequency over the collection of documents [26]. The WIDF weight is given by:

$$WIDF(d, t) = TF(d, t) \sum_{i \in D} TF(i, t) \quad (6)$$

In addition to these five representations we use two functions which combine several criteria. Both consider as main features of a web page a subset of the set of words it contains. The criteria which are combined are: word frequency in the text, the appearance of a word in the title of the web page, the position of a word throughout the text, and whether or not the word appears in emphasized tags. For the last criterion a

set of HTML elements are selected like emphasized tags, and the “title” tag, because we think they can capture the author’s intention.

These two representation are the Analytic Combination of Criteria (ACC), and the Fuzzy Combination of Criteria (FCC). The difference between them lies in how they evaluate and combine the criteria. The first [3] uses a linear combination of criteria, whereas the second [19] combines the criteria by using a fuzzy system.

#### A. Analytic Combination of Criteria

The ACC is a linear combination of the criteria. Once they are fixed, the functions corresponding to each of them and their combination have to be described. We use the following functions:

- The **frequency function of a word on a web page**:

$$f_f(d, w) = \frac{n_f(d, w)}{N_{tot}(d)} \quad (7)$$

Where  $n_f(d, w)$  is the number of occurrences of a word  $w$  in the page  $d$ , and  $N_{tot}(d)$  is the total number of words in the web page  $d$ . This definition allows the function to be normalized using  $\sum_1^k f_f(d, w) = 1$ , where  $k$  is the number of different words in the document.

- The **frequency function of a word in the title**:

$$f_t(d, w) = \frac{n_t(d, w)}{N_{tit}(d)} \quad (8)$$

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

- The **emphasized function of a word**:

$$f_e(d, w) = \frac{n_e(d, w)}{N_{emph}(d)} \quad (9)$$

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

- The **position function**. To compute the position criteria the web page is split into four equal parts according to the number of features in the <body> tag. Considering  $n_f(d, w) = n_{1,4}(d, w) + n_{2,3}(d, w)$ , where  $n_{1,4}(d, w)$  and  $n_{2,3}(d, w)$  are the number of times that the word  $w$  appears in the introduction or in the conclusion (first and fourth parts of  $d$ ) and in the document body (second and third parts) respectively, the function is:

$$f_p(d, w) = \frac{2n_{1,4}(d, w) + n_f(d, w)}{2n_f(d, w) + N_{tot}(d)} \quad (10)$$

This expression is a particularization of a more general expression [3], where we impose a weight ratio of 3/4 and 1/4 for preferential and standard quarters respectively.

Finally, the combination of these criteria for each feature  $i$  in ACC weight function is given by the following relevance function:

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

In section 5 the values of  $C_1$ ,  $C_2$ ,  $C_3$ , and  $C_4$  will be commented.

#### B. Fuzzy Combination of Criteria

The FCC weight function uses the same criteria to evaluate feature relevance on a web page. It is a fuzzy system for the assignation of feature weights and their combination. Thus, the linguistic variables of the fuzzy system are:

- *Text-frequency* with associated labels “Low”, “Medium” and “High”;
- *Title-frequency* with associated labels “Low” and “High”;
- *Emphasis* with associated labels “Low”, “Medium” and “High”;
- *Global-position* with associated labels “Standard” and “Preferential”.

To make the fuzzy rules independent of document size, *text-frequency*, *title-frequency* and *emphasis* variables are normalized to the greatest of the frequency in each criterion. So, in equations 7, 8, and 9 the denominator is the maximum number, instead of the total number.

In the position criterion, the variable *global-position* is calculated from an auxiliary fuzzy system. This auxiliary system has as input the *line-position* variable, and as output the *global-position*, with associated labels “Standard” and “Preferential”. The output value is calculated from the different positions where a feature appears.

As output of the fuzzy module we have only one variable for each feature on a web page: *relevance*, with associated labels “NonRelevant”, “LowRelevant”, “MediumRelevant”, “High-Relevant”, and “VeryRelevant”.

The inference engine is based on a Center of Mass Algorithm (COM) that weights the output of each rule against the truth degree of each antecedent. The set of rules and a detailed explanation of the fuzzy module can be found in [19].

The foundation of the rules is based on the following points:

- A web page may have no emphasized words.
- A word appearing in the title may not always be relevant (for instance, the title can be generated automatically by a HTML editor).
- In general, position is a criterion that gives more weight on longer pages than on shorter ones.
- A word with a high frequency on a page could mean that the word is an all-purpose word and, consequently, it does not discriminate.

## IV. WEB PAGES COLLECTION

We have used the BankSearch as the web document collections to evaluate the different representations. The BankSearch [24] is a dataset of 11,000 web documents pre-classified into 11 equally-sized categories, each containing 1,000 web documents and grouped in hierarchical levels. It

was generated by Mark Sinka and David Corne, from Reading University in the U.K., with the main aim of proposing a general dataset for web document clustering and similar experiments.

In this work, we only considered the main 10 categories because the eleventh one is an extra category that is a superset of one of the 10 main categories. The first three belong to the general "Banking & Finance" theme: Commercial Banks (A), Building Societies (B) and Insurance Agencies (C). The next three belong to the more general theme "Programming Languages"; Java (D), C/C++ (E) and Visual Basic (F). Finally, the other two are Soccer (I) and Motor Sport (J), which belong to the theme "Sport".

We use the vector space model, so that each web page is represented by using a vector where each component is the weight of a feature in this page. In this work, a feature is a character stream between two space characters. We fix the maximum length of the feature to be 30 characters. In order to calculate the values of the vector components for each web page we follow these steps:

- 1) We eliminate all the punctuation marks except some special marks that are used in URLs, e-mail addresses, and multiword terms.
- 2) The words on a stoplist used in the Cross Language Evaluation Forum (CLEF) are eliminated from the web pages.
- 3) We considered only features that occur more than one time in the collection.
- 4) We obtain the stem of each feature by using Porter's stemming algorithm [18].
- 5) We count the number of times each feature appears on each web page, and the number of web pages where each feature appears.
- 6) To calculate the ACC and FCC representations, we memorize the position of each feature throughout the web page, and whether or not the feature appears in emphasized HTML tags.

## V. EXPERIMENTS

We tried the seven functions of the collection by means of a clustering process. The coefficients for the linear combination of the ACC representation were estimated as a result of previous research [3] on the influence of each criterion in the representation. That research was carried out with a web page collection different from the one involved in the current experiments. The research indicated that the position and the frequency of a word on a web page are better criteria separately than emphasis and frequency of a word in the title. Nevertheless, none of them is as good as the linear combination of all of them. As result of this research, the best coefficients of the linear combination were estimated as:

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

The selection of these coefficients should be independent of the evaluation data and so, in these experiments, we use the same values that were found in [3]. The ACC representation will be called  $ACC(0.3, 0.15, 0.25, 0.3)$ .

Next we describe the main characteristics of the experiments carried out.

### A. Features Reduction

One of the main problems in representation and later clustering is the high number of features that have to be taken into account when documents are dealt with. In these experiments we tried two types of feature reduction:

- 1) Only the features that appear more than  $FFmin$  times in more than  $DFmin$  web pages, and less than  $FFmax$  times in less than  $DFmax$  web pages are selected. This is a variant of [23]. This type of reduction is applied to the seven term weighting functions. We called this type of reduction "Min-Max".
- 2) The proper weight functions are used as reduction method. So, the  $N$  most relevant features (those of higher function values) on each web page are selected. In this case, this type of reduction has been able to be applied to every term weighting function, except to the Binary function. We have tried seven values for  $N$ , from 1 to 70.

We tried several values for the variables  $FFmin$ ,  $DFmin$ ,  $FFmax$ ,  $DFmax$ , and  $N$  in order to obtain different magnitude reductions. We fix the maximum number of features to be one magnitude order less than the initial number of features.

It is important to emphasize that the Binary, TF, ACC and FCC representations are independent of the collection information; in other words, they only need information from the web page itself to represent it. However, the Binary-IDF, TF-IDF and WIDF representations need the whole collection information to represent each web page.

### B. Clustering Algorithm

Clustering involves dividing a set of  $n$  objects into a specified number of clusters  $k$ , so that objects are similar to other objects in the same cluster, and different from objects in other clusters. In our case, the objects are web documents.

In this work, we evaluated which term weighting functions improve the results of the web page clustering. So, a good term weighting function is one which leads to a good clustering solution. Since we work with a known number of classes we use a partition clustering algorithm. We select the Direct algorithm from the well-known CLUTO library. In the Direct method, the clustering solution is computed by simultaneously finding two clusters. If the number of clusters is small, this method is appropriate for finding a good partition [10].

### C. Evaluation Measures

We test the performance of the clustering algorithm with the different term weighting functions. We carry out an external evaluation to determine the quality of the clustering results by means of F-measure.

## Proper function reduction - CLUTO (k=10)

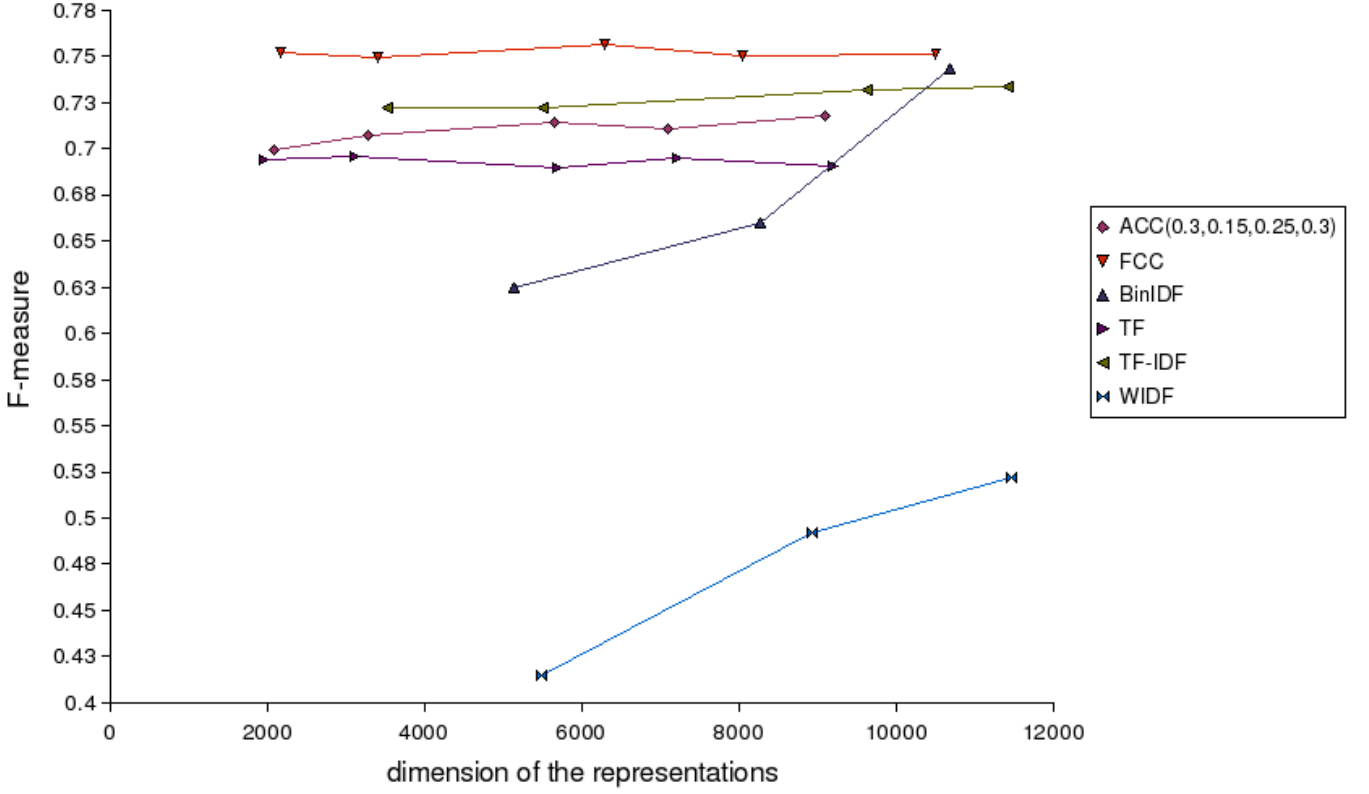


Fig. 1. Clustering results with proper function reduction

We call the groups given by the Benchmark Dataset “classes”, and the groups obtained by the algorithm “clusters”.

The F-measure [25] combines the precision and recall measures. The F-measure of cluster  $j$  and class  $i$  is given by the following equation:

$$F(i, j) = \frac{2 \times \text{Recall}(i, j) \times \text{Precision}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)} \quad (11)$$

with

$$\text{Recall}(i, j) = \frac{n_{ij}}{n_i} \quad (12)$$

$$\text{Precision}(i, j) = \frac{n_{ij}}{n_j} \quad (13)$$

where  $n_{ij}$  is the number of members of class  $i$  in cluster  $j$ ,  $n_j$  is the number of members of cluster  $j$  and  $n_i$  is the number of members of class  $i$ . For all the clusters:

$$F = \sum_i \frac{n_i}{n} \max\{F(i, j)\} \quad (14)$$

where  $n$  is the number of web documents. The higher the F-measure values the better the quality of the clustering.

### D. Results

The results of the experiments can be seen in Figures 1 and 2. The first one corresponds to the results of the clustering using the proper function reduction and Figure 2 fits to “Min-Max” reduction.

The results show that when the proper function is used for reducing the number of features, the fuzzy combination function FCC shows the best behavior. In addition this behavior is enough stable in all the dimensions. TF-IDF function is the second better followed by the Analytic Combination of Criteria ACC. WIDF and B-IDF show a less stable performance. It is necessary to emphasize that with this feature reduction for the same value of  $N$ , the B-IDF and WIDF dimensions are larger than the dimensions for the rest of the representations.

When the “Min-Max” feature reduction is applied to FCC and TF-IDF, in this order, show the best performance. Nevertheless, TF is the third better function followed by ACC. It is remarkable that with the two reductions applied, the worst results are obtained with the WIDF function.

If we examine the results obtained, the Fuzzy Combination of Criteria FCC, which takes account of the author’s intention by means of emphasized tags, improves the results of the representations that do not take account of this criterion. FCC also improves the results of the ACC function which provides

## MinMax reduction - CLUTO (k=10)

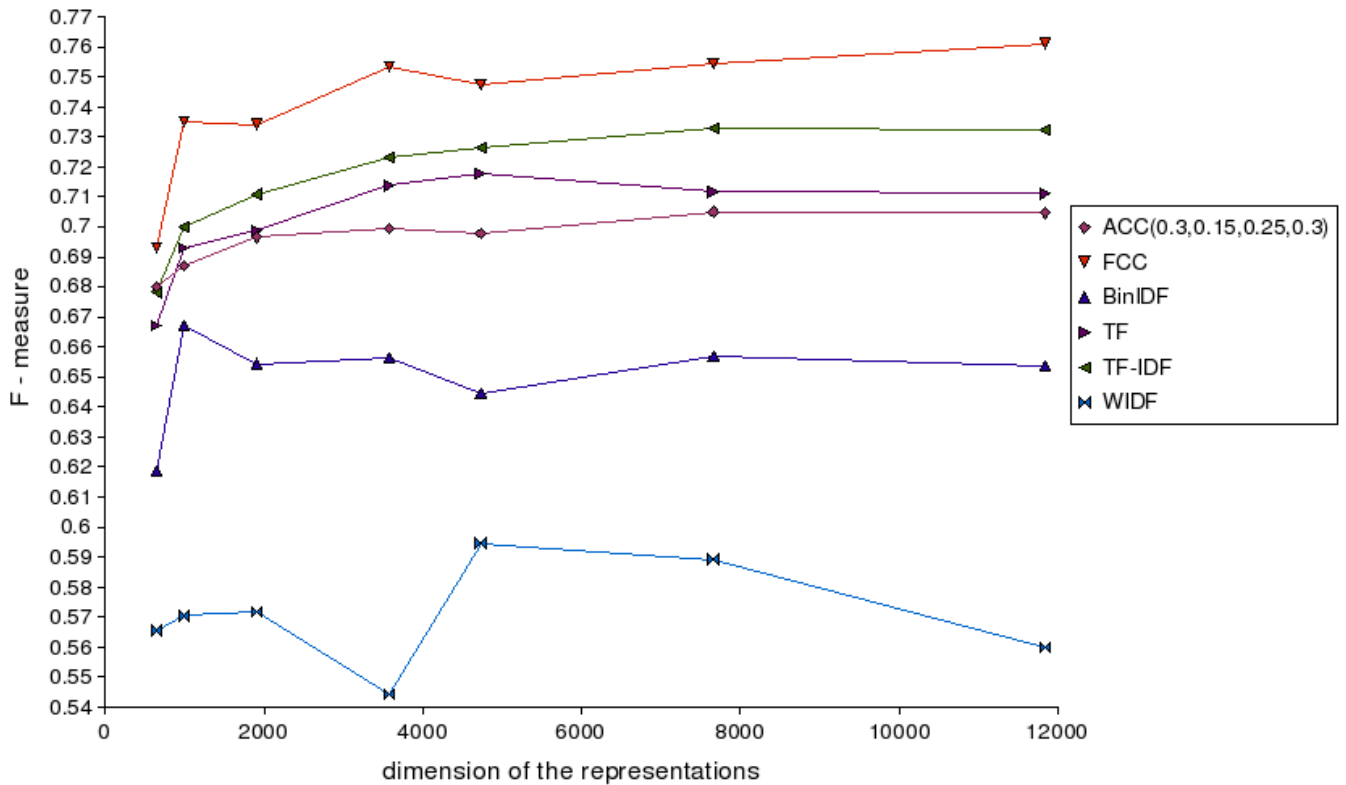


Fig. 2. Clustering results with "Min-Max" reduction

the author's intention by means of an analytical combination.

An additional advantage of the FCC function is that it does not use collection information, only uses information of the proper page, which is quite desirable in representing web pages.

## VI. CONCLUSIONS

We tried seven term weighting functions and two types of features reductions in order to study which of them lead to a better clustering results. We used a partitioning clustering algorithm from the CLUTO library. The quality of the clustering solution is determined by means of an external evaluation, the F-measure, by comparing the algorithm results with those of the Benchmark collection. The experiments are carried out with a collection of web documents created for general use in web document clustering research. It is composed of 10 main categories.

The best clustering results were obtained by using the proposed term weighting function which carries out a fuzzy combination of criteria, FCC. This function allows to obtain representations with mark-up HTML enrichment. These criteria are: (1) the frequency of a feature in a web page, (2) the frequency of a feature in the title, (3) whether a feature is emphasized and, finally, (4) the position of a feature in the web

page. The (2) and (3) criteria are obtained from the HTML mark-up of the web pages. The function which combine the same criteria in an analytic way, ACC, obtained worse results. Clearly, the fuzzy module makes the most of the knowledge provided by the different criteria. Besides, they show quite homogeneous behavior, even with a low number of features. This is a very desirable virtue when large collections of web page have to be dealt with.

The TF-IDF function presents the second better performance. However, the FCC function has an additional advantage: it only uses information of the proper web page, not information of the collection. In a context in which a collection can have a high activity rate we think this is an valuable characteristic.

The results suggest that using information from the HTML mark-up combined with textual information leads to good results in web page clustering. The main contribution of this work has been to prove that by using term weighting functions based on heuristic combination of criteria, the web page clustering can improve against the use of functions based solely on frequencies, in the web page or in the collection.

## ACKNOWLEDGMENT

This work has been partially supported by the CICYT project TIN2005-08943-C02-02.

## REFERENCES

- [1] A. Casillas, V. Fresno, M. González de Lena and R. Martínez. *Evaluation of Web Page Representations by Content through Clustering*. String Processing and Information Retrieval. LNCS series of Springer-Verlag, 129-130, 2004.
- [2] S. Chakrabarti, M. Joshi and V. Tawde. *Enhanced topic distillation using text, markup tags, and hyperlinks*. SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, 208-216, 2001.
- [3] V. Fresno and A. Ribeiro. *An Analytical Approach to Concept Extraction in HTML Environments*. Journal of Intelligent Information Systems - JIIS. Kluwer Academic Publishers, 215-235, 2004.
- [4] J. Guo. *Integrating Automatic Document Clustering into Web Log Association Rule Mining*. Thesis Faculty of Computer Science Dalhousie University, Canada, 2004.
- [5] M. Halkidi, B. Nguyen, I. Varlamis and M. Vazirgiannis. *THESUS: Organizing Web Document Collections Based on Link Semantics*. In VLDB Journal, special issue on Semantic Web, 2003.
- [6] E. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher and J. Moore. *WebACE: A Web Agent for Document Categorization and Exploration*. Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98), 1998.
- [7] V. Harmadas, M. Sanderson and M. D. Dunlop. *Images retrieval by hypertext links*. Proceeding of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval, 296-303, 1997.
- [8] F. Iavernaro. *Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming*. <http://www.knowledgeboard.com/cgi-bin/item.cgi?id=129238&d=pnd>, 2004.
- [9] S. Jones. *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, Vol. 28, N. 1, 11-21, 1972.
- [10] G. Karypis. *CLUTO: A Clustering Toolkit*. Technical Report: 02-017. University of Minnesota, Department of Computer Science, Minneapolis, MN 55455.
- [11] A. Leuski and J. Allan. *Improving interactive retrieval by combining ranked lists and clustering*. Proceedings of RIAO2000, 665-681, 2000.
- [12] H. P. Luhn. *A statistical approach to mechanized encoding and searching of literary information*. IBM Journal of Research and Development, Vol. 1, N. 4, 307-319, 1957.
- [13] D. Merkl. *Text data mining. A handbook of Natural Languages Processing Techniques and Applications for the Processing of Languages as Text*. R. Dale, H. Moisl and H. Sommer (Eds). New York: Marcel Dekker, 1998.
- [14] A. Molinari and G. Passi. *A Fuzzy representation of HTML documents for Information Retrieval Systems*. Proceedings of the IEEE International Conference on Fuzzy Systems, New Orleans. Vol. 1, 107-112, 1996.
- [15] A. Molinari, G. Passi and R. A. Marques Pereira. *An indexing model of HTML documents*. SAC '03: Proceedings of the 2003 ACM symposium on Applied computing, Melbourne, Florida, 834-840, 2003.
- [16] J. Moore, E. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar and B. Mobasher. *Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering*. Workshop on Information Technologies and Systems, 1997.
- [17] J. M. Pierre. *On the Automated Classification of Web Sites*. Linking Electronic Articles in Computer and Information Science. Vol. 6, 2001.
- [18] M.F. Porter. *An algorithm for suffix stripping*. Reprinted in Sparck Jones, Karen, and Peter Willet, Readings in Information Retrieval, San Francisco: Morgan Kaufmann, 1997.
- [19] A. Ribeiro, V. Fresno, M. García-Alegre and D. Guinea. *A Fuzzy System for the Web Page Representation*. Intelligent Exploration of the Web, Springer-Verlag Group, 19-38, 2002.
- [20] G. Salton, C. S. Yang. *On the specification of term values in automatic indexing*. Journal of Documentation, Vol. 29, N. 4, 351-372, 1973.
- [21] G. Salton and M. McGill. *Introduction to Modern information Retrieval*. McGraw Hill, New York, 1983.
- [22] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1988.
- [23] F. Sebastiani. *Machine Learning in Automated Text Categorization*. ACM Computing Surveys, Vol. 34, N. 1, 1-47, 2002.
- [24] M. P. Sinka and D. W. Corne. *A Large Benchmark Dataset for Web Document Clustering*. Soft Computing Systems: Design, Management and Applications, Frontiers in Artificial Intelligence and Applications, Vol. 87, 881-890, 2002.
- [25] C. J. van Rijsbergen. *Foundations of evaluation*. Journal of Documentation, Vol. 30, 365-373, 1974.
- [26] T. Tokunaga and M. Iwayama. *Text categorization based on Weighted Inverse Document Frequency*. Technical Report 94 TR0001, Department of Computer Science, Tokyo Institute of Technology, 1994.
- [27] Y. Yang, S. Slattery and R. Ghani. *A Study of Approaches to Hypertext Categorization*. Journal of Intelligent Information Systems - JIIS. Kluwer Academic Publishers, Vol 18, 1-25, 2002.
- [28] L. Yi, B. Liu. *Web Page Cleaning for Web Mining through Feature Weighting*. [www.cs.uic.edu/liub/publications/ijcai03-webClean.pdf](http://www.cs.uic.edu/liub/publications/ijcai03-webClean.pdf), 2003.
- [29] O. Zamir and O. Etzioni. *Grouper: A dynamic clustering interface to web search results*. Proceedings of the WWW8 Conference, 1999.
- [30] Y. Zhongmei and B. Choi. *Bidirectional Hierarchical Clustering for Web Mining*. IEEE/WIC International Conference on Web Intelligence (WI'03), 2003.