Using Fuzzy Logic to Leverage HTML Markup for Web Page Representation

Alberto P. García-Plaza[1], Víctor Fresno[1], Raquel Martínez[1] and Arkaitz Zubiaga[2]
[1] NLP&IR Group at UNED, Spain
[2] University of Warwick, UK
{alpgarcia,vfresno,raquel}@lsi.uned.es, arkaitz@zubiaga.org

***Abstract*—The selection of a suitable document representation approach plays a crucial role in the performance of a document clustering task. Being able to pick out representative words within a document can lead to substantial improvements in document clustering. In the case of web documents, the HTML markup that defines the layout of the content provides additional structural information that can be further exploited to identify representative words. In this paper we introduce a fuzzy term weighing approach that makes the most of the HTML structure for document clustering. We set forth and build on the hypothesis that a good representation can take advantage of how humans skim through documents to extract the most representative words. The authors of web pages make use of HTML tags to convey the most important message of a web page through page elements that attract the readers' attention, such as page titles or emphasized elements. We define a set of criteria to exploit the information provided by these page elements, and introduce a fuzzy combination of these criteria that we evaluate within the context of a web page clustering task. Our proposed approach, called Abstract Fuzzy Combination of Criteria (AFCC), can adapt to datasets whose features are distributed differently, achieving good results compared to other similar fuzzy logic based approaches and TF-IDF across different datasets.**

***Index Terms*—Document Representation, Fuzzy Systems, Term Weighting Function, Web Page Clustering.**

I. Introduction

ACCESS to and retrieval of web documents in large collections can be substantially eased when the documents are properly clustered into topics. The organization of documents into clusters then facilitates focusing search on the topic or topics of interest, shrinking down the large collection to smaller sets of topically related resources. While a body of research has studied clustering of web documents, little attention has been paid to the improvement of document representation techniques and the definition of robust term weighting functions.

We are interested in the study of document representation techniques based on fuzzy logic that can generalize across datasets when the purpose is to group documents by topic in the absence of category information. This can be particularly useful in cases where new categories can emerge, so that the system should be able to accommodate its clustering process to be able to find these new categories with the information extracted from the documents.

Prior to the clustering process, document representation plays a very important role in web page clustering, and constitutes the central point of research of this work. In the document representation phase we choose the characteristics of the document that we consider useful, and assess how this information could be exploited.

The textual content is often used for the representation of web pages, given that it is readily available and is easy to process; however, an unweighted bag-of-words representation of the content does not always lead to optimal results. Interestingly, the content of an HTML document is structured in tags, providing additional clues on how different parts of the content differ from one another, and ultimately affecting its visual presentation [1]. The HTML structure of a web document can be further exploited to identify the most representative words within its content. We pay special attention to document contents, introducing a representation that makes the most of information inherent to the document. Hence, we set out to delve into the study of approaches that garner the additional information that HTML tags provide for improved representation of web documents. Moreover, we also look into the use of additional context information, using anchor texts pointing to web pages, as well as statistics inferred from the whole collection. We assess the suitability of using these additional characteristics for web document representation in a clustering task.

We make use of a fuzzy system, as a flexible solution that enables to handle the importance of the different characteristics of web pages. For instance, the titles of web pages can often be deemed rhetorical, where some words are very representative of its content, but other words are solely used to embellish the language. When considering frequency in titles within a linear combination of criteria in order to identify the most important words within a document, these words would get a high importance value, which would not correspond with their real importance to describe the content of the page, since they are only embellishing the language. In these linear combinations, when a word is important with respect to a single criterion, the corresponding component will have a value which will always be added to the importance of the word in the document, regardless of the importance corresponding of the rest of the components. On the contrary, by using fuzzy logic it is possible to define related conditions, e.g., a word should

Manuscript received October 29, 2015; revised April 20, 2016; accepted June 6, 2016.
This work has been part-funded by the Spanish Ministry of Science and Innovation (MED-RECORD Project, TIN2013-46616-C2-2-R) and the PHEME FP7 project (grant No. 611233).

appear in the title and emphasized or within specific parts of the document to be considered important. In the same way, if a word appears in the title but not in other criteria, then we could consider that word less important. Here we delve into the use of fuzzy logic for the purposes of exploiting these characteristics of web pages.

Building on the state-of-the-art unsupervised fuzzy logic approach for HTML document representation [2], known as Fuzzy Combinations of Criteria (FCC), we propose three alternative approaches, namely EFCC, AddFCC, and AFCC. We perform the evaluation of these and additional baseline approaches over three benchmark web page collections through a clustering task using the well-known Cluto library [3]. Our proposed approach AFCC, which more suitably adapts to datasets with different characteristics, consistently outperforms the other approaches on the three datasets under study. AFCC provides a flexible, straightforwardly applicable approach that makes the most of the structure and content of HTML documents for web mining purposes.

In what follows, we provide background on the task of web page representation, followed by a summary of previous work in the literature as well as their relevance to our work in Section III. We move on then to the experimentation, describing first the experimental settings in Section V, introducing and evaluating two new approaches, AddFCC and EFCC, to improve the existing FCC approach in Section VI, and further studying the analysis and tuning of membership functions through the so-called AFCC in Section VII. We outline the contributions of our work and conclude the paper in Section VIII.

II. Background

The document representation process can be split into three stages: (1) selection of feature sources, (2) weighing of those features, and (3) dimensionality reduction. Throughout this paper we delve into these three stages, paying close attention at how to model a term weighing function.

Within the **selection of feature sources**, the information that needs to be represented within each document is picked, e.g., plain textual content, titles, or hyperlinks. There are mainly three different approaches. First, *content based*, which make use of the textual content of documents. This kind of approaches were initially developed for document retrieval in static collections, but with the popularity of the Internet, they have also been adapted to the Web. Further exploiting the characteristics of the Web, the textual content of the documents has also been enhanced with the information provided by HTML tags about document formatting, page structure, visual aspects, etc. Second, *link based*, which take advantage of the link structure among the pages in the collections. It considers hyperlinks as citations between pages. When two documents have many incoming links in common, or both documents have outgoing links to a similar set of documents, then the documents are likely related. Third and last, the *hybrid approach*, which combines features from the textual content of the document and from the context of the page. Here context can include not only hyperlinks or anchor texts, but also other information sources, such as information inferred from the entire collection, or definitions extracted from external resources such as Wikipedia.

In the subsequent step of **weighing features**, each feature is assigned a weight in each document, the weight being representative of the feature's importance in the document. There are different elements that can determine the importance of a word within the document. One can then define a set of criteria to make the most of the different elements when it comes to improving the document representation. The initial hypothesis of the present work lies in that a good representation should take advantage of how humans skim through web documents to pick out salient words. For example, some words are explicitly highlighted with specific HTML tags. Then, if one wants to determine the importance of a word in a document, in addition to the rather straightforward frequency of the word in the document, one can also take advantage of these highlighted words as a signal that conveys the remarkable importance of the word.

In the final **dimensionality reduction step**, useless features are removed by keeping the document's most representative features, which makes it more efficient to be handled computationally.

III. Related Work

There have been multiple attempts at exploiting the structure of web pages to maximize understanding of their contents for different purposes. Kwon and Lee [4] aimed to classify web sites by using not only their home pages, but also the content of pages that linked to the home page of each site. Their weighing scheme to establish term importance takes into account different HTML tags such as titles, headlines, and boldfaced texts, to identify the most representative words in a web page. They show that the use of the extended set of pages boosts the performance with respect to the ordinary classifier using only the home pages. Golub and Ard [5] studied how setting the importance of different parts of a web page could have an impact on the outcome of a web page classification task. They classified a set of $1,003$ web pages based on titles, headings, metadata, and text. As a single feature, they found the titles to be the most useful; however, since not all web pages have titles, they found that combining all features leads to the best overall performance. In an earlier work, making use of the link structure among documents, Fisher and Everson [6] analyzed the usefulness of links for web page classification tasks. They conclude that links may be useful, but it depends on link density and quality.

Besides links and anchor texts, other kinds of information have also been exploited over the years. For instance, Kovacevic et al. [7], Shih and Karger [8], Bohunsky and Gatterbauer [9] and Bartik [10], or more recently Herzog et al. [11], have used the visual appearance of a web page, after rendering its content in a browser, for the purposes of representing web documents. Another work along these lines is that performed by Gasparetti et al. [12], which describes an approach based on the implicit signal that can be captured through web browsing interactions, defining a DOM-based representation of visited

pages. While these approaches might be handy for systems that exploit the visual appearance of web pages, our objective instead is to avoid reliance on the visual rendering by solely exploiting the HTML structure.

Information from external knowledge bases such as Wikipedia has also been exploited by others such as Hu et al. [13] and Li et al. [14]. The use of these knowledge bases can help enrich the content inherent to the web documents. In these cases, the classification structure of articles within Wikipedia's taxonomy is leveraged to associate web documents with Wikipedia concepts and categories; this process of linking concepts in documents to Wikipedia articles is also known as wikification [15]. Then, Wikipedia entries or their n-grams are matched with documents to expand the content of each document with related content.

While recent years have seen a growing body of research in the use of fuzzy logic to make the most of the document representation for clustering purposes [16], [17], [18], the exploitation of the characteristics of HTML documents, which are rich in structure, remains relatively unexplored. One of the most recent approaches making use of fuzzy logic representation for semi-structured documents is that introduced by Ensan and Biletskiy [19]. The caveat of this approach is the need of a human in the loop for generating templates, which boosts the system's performance by extracting additional information within a supervised approach. The authors did not however explore an alternative solution for fully automating the process. Our work intends to fill this gap, performing a comprehensive study on the use of the HTML structure and content with fuzzy logic for web document clustering in an unsupervised approach.

The works which are closest to ours are by Molinari and Pasi [20], focused on an Information Retrieval task, and by Fresno and Ribeiro [21], who presented an Analytical Combination of Criteria (ACC) to represent web pages in web page classification and clustering tasks. It is based on a linear combination of different heuristic criteria within the Vector Space Model. These criteria were selected taking into account how a human reader skims through a document to identify the most representative words. The criteria used by ACC are `title`, `emphasis`, `position`, and `frequency`. Based on the same criteria, Fresno [2] proposed an approach called Fuzzy Combination of Criteria (FCC), an alternative way of combining them in a non-linear way. In this case, a fuzzy logic based system is employed to define the expert knowledge about how to combine these criteria. The output is also a single vector within the Vector Space Model, representing the estimated importance of each term in a given document. One of the main advantages of FCC is its flexibility, which can be easily utilized for different purposes within different tasks. In fact, recent works have adapted FCC for different purposes, including Nassem et al. [22] for the detection of near duplicate web pages, and Bartik [10] for web page classification. The use of fuzzy logic for feature selection and web representation is still an active topic of interest, and is used as can be seen in recent research [23], [24]. To the best of our knowledge, however, no alternatives to FCC have been proposed, and therefore FCC represents, at the time of this writing, the state of the art in the fully automated, unsupervised fuzzy model for web page representation based on web page structure.

In the present work, we rely on the FCC fuzzy representation as a starting point for our research in order to study the fuzzy combination model in different ways, from analyzing its original definition, to proposing new ways of exploiting the system to perform the combination, as well as to explore the possibility of adapting the system to the input we want to represent. In what follows we further describe the FCC approach, which our work builds on.

IV. FCC: Fuzzy Combinations of Criteria

The fuzzy system in FCC is built over the concept of linguistic variable and its fuzzy sets. Each variable describes the membership degree of an object to a particular class and it is defined by human experts. This membership degree is defined by a membership function. For each heuristic criterion (`frequency`, `title`, `emphasis`, and `position`), an associated linguistic variable is defined, as well as for the system output (importance):

1) **Text Frequency**: term frequency in the document. Its input is calculated by normalizing this frequency to the maximum number of occurrences of any term in that document. It is defined in three fuzzy sets: *low*, *medium*, and *high* (see Fig. 1a).
2) **Title**: term frequency within the `<title>` tag. Its input is calculated by normalizing this frequency to the maximum number of occurrences of any term in the title of that document. It is defined in two fuzzy sets: *low* and *high* (see Fig. 1b).
3) **Emphasis**: term frequency in emphasized parts of the text[1]. Its input is calculated by normalizing this frequency to the maximum number of occurrences of any term in emphasized text segments in that document. It is composed of three fuzzy sets: *low*, *medium* and *high* (see Fig. 1c).
4) **Position**: the global position of a term in the document, defined in two fuzzy sets: *standard* and *preferential* (see Fig. 1d). It is obtained by means of an Auxiliary Fuzzy System that takes as input all the positions of a term within a document (captured by the other linguistic variable **term position**) and returns the global position value in terms of two fuzzy sets, *standard* and *preferential*.
5) **Importance**: it is the output of the fuzzy system and equates to the estimated importance of a term in the document content. It has five homogeneously distributed fuzzy sets: *no*, *low*, *medium*, *high* and *very high*.

These membership functions have a trapezoidal shape. All the variables except `emphasis` are defined by sets of equal size symmetrically distributed along the possible input values. These sets were defined without restricting to specific datasets. However, `emphasis` is considered separately because when the maximum frequency value for emphasized words in a document is small, the normalization could have high impact

[1]We use a manually created list of HTML tags that add emphasis: `<em>`, `<b>`, `<u>`, `<strong>`, `<big>`, `<h*>`, `<cite>`, `<dfn>`, `<i>`, `<blockquote>`

on the importance of other emphasized terms. For example, using symmetrical sets and having a maximum of 4 would lead to consider the importance of terms emphasized once as *low*, when we may want to increase the importance of these terms. For this reason, the sets for `emphasis` were asymmetrically defined. This way, frequencies that would be strictly *low* can also be considered as *medium*, since we can expect small maximum values in `emphasis`.

The other part of the knowledge base is a set of IF-THEN rules. The aim of the rules is to combine one or more input fuzzy sets (antecedents or premises) and to associate them with an output fuzzy set (consequent). Once the consequents of each rule have been calculated, and after an aggregation stage, the final set is obtained representing the word based on its importance within the document content. The complete set of 31 rules defined in the FCC approach can be found in [2, p. 130]. Example 1 shows an example of an IF-THEN rule.

Example 1:
IF Title IS High AND Frequency IS Low AND Emphasis IS Low AND Position is Standard THEN Importance IS Low

The rule set is complete, so that every possible input has to trigger at least one rule. The inference engine evaluates all the triggered rules on the basis of the *Center Of Mass* (COM) algorithm, which weighs the output of every triggered rule, taking into account the truth degree of their antecedents. It takes the balance point or centroid of all the scaled membership functions taken together for that variable [25]. The output for each term input to the system is calculated by scaling the membership functions by product and combining them by summation.

The rule base presented in [2] relies on the following three considerations:

1) If a word appears in the title or the word is emphasized, it should also appear in one of the other criteria in order to be considered important. This aims to alleviate the problem of rhetoric titles or non-informative highlighting;
2) Words occurring in the beginning or in the end of a document are more likely to be important than the rest of the words, as some documents contain overviews and summaries in order to attract the interest of the reader. When the words in a preferential position do not occur also in the title or emphasized, then we could assume that the document does not adhere to the mentioned structure and we could reduce the importance value of that word;
3) It might be the case that there are no emphasized words in a document, the document has no title, or the title has no important words. In these cases we have to take care of the penalization it could cause to the combination. If the previous criteria did not pick important words, the word frequencies in the whole document are used. Different from the others, the `frequency` criterion is always available.

V. Experimental Framework

In this section, we describe the experimental settings that we use in our research.

A. Datasets

To make results comparable to those by Fresno [2], we also use the same two datasets, Banksearch [26] and WebKB [27]. Additionally, we use the Social-ODP-2k9 Dataset [28], which provides the features we need for the extended analysis looking at anchor texts.

1) **Banksearch [26].** A benchmark dataset designed for evaluation of web page clustering. We use the 10 main categories –*A* to *J*–,(Commercial Banks, Building Societies, Insurance Agencies, Java, C/C++, Visual Basic, Astronomy, Biology, Soccer, Motor Sports). We removed the other category (*K*, Sport) for being of a different granularity level and hence not comparable to the rest. This results in $9,897$ documents evenly distributed across categories.
2) **WebKB [27].** A dataset that includes web pages from computer science departments of various universities. We use $4,518$ web pages that are categorized into 6 imbalanced categories (*Student*, *Faculty*, *Staff*, *Department*, *Course*, *Project*), after removing the *Other* miscellanea category that is not comparable to the rest. This dataset is more heterogenous than the others, as web pages on a common subject can be found in different categories, such as *Java programming* categorized into *Student*, *Course* or *Department*.
3) **Social ODP 2k9 [28].** A dataset that consists of HTML documents retrieved from links bookmarked by users on Delicious.com. The classification of these documents was inferred from the taxonomy of the Open Directory Project[2]. From this dataset, we used $12,148$ documents that passed a valid HTML test. The documents are classified into 17 categories. This dataset is also imbalanced, where the most prominent category accounts for 26% of the documents. In addition to the documents themselves, we collected up to 300 anchor texts per document in the collection. The anchor texts were retrieved by querying Google for links pointing to collection pages.

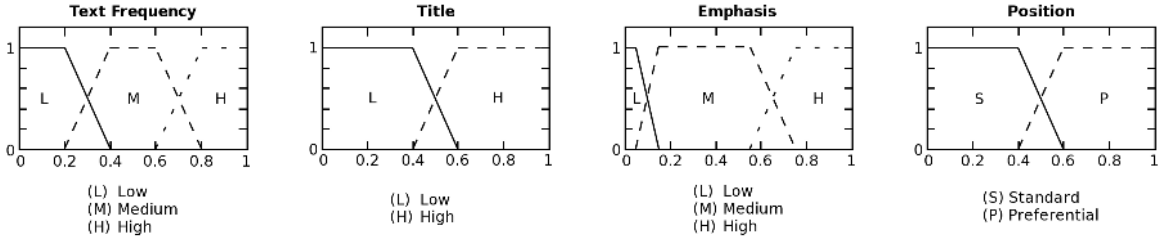
B. Baseline

As a baseline, we compute the weight of each word occurring in a document by using the well-known TF-IDF term weighing function, where the term frequency (TF) in a document is combined with the Inverse Document Frequency (IDF) of that term in the whole collection:

$$\text{TF-IDF}(t_i, d_j, D) = \text{TF}(t_i, d_j) \times \log \frac{|D|}{|\{d_j \in D : t_i \in d_j\}|} \quad (1)$$

where t_i is a term, $_j$ a document, D the whole corpus, $|D|$ is the total number of documents in the corpus and $|\{d_j \in D : t_i \in d_j\}|$ is the number of documents where the term t_i appears.

[2]http://www.dmoz.org/



(a) Frequency sets (b) Title sets (c) Emphasis sets (d) Position sets

Fig. 1: Data base for FCC. Input linguistic variables.

C. Dimensionality Reduction

Dimensionality reduction aims to reduce the number of vector components, consequently attempting to reduce the computational cost while the performance loss is as little as possible. Many different dimensionality reduction approaches have been introduced in the literature, aiming to address the limitations of traditional techniques such as Principal Component Analysis and classical scaling. These approaches range from simpler techniques relying solely on term frequencies, to more complex methods derived from approaches originally defined for text classification. Van der Maaten et al. [29] present a review and comparison of nonlinear dimensionality reduction techniques, which they group into two types: (1) convex techniques (full spectral or sparse spectral), optimizing an objective function that does not contain any local optima; and (2) non-convex techniques (weighted euclidean distances, alignment of local linear models, or neural networks) that optimize objective functions that do contain local optimal.

On the other hand, from the perspective of availability and use of labeled data for training, feature selection can be categorized as supervised, semisupervised or unsupervised. When it comes to supervised approaches, He et. al [30] introduce a feature selection algorithm called Laplacian Score, and Kala et al. [31] use Fuzzy C Means clustering to find clusters in the given training dataset. Others like [32] and [33] introduced approaches within a semi-supervised learning scenario. For an unsupervised scenario, in the abscense of class information, there are feature selection and dimensionality reduction methods which preserve the local geometrical structure such as Multi-Cluster Feature Selection [34] and L1 Graph Based on Sparse Coding for Feature Selection [35].

We introduce a unsupervised reduction method called Most Frequent Terms (MFT), which is based on term importance estimated by a term weighing function. The MFT method works as follows. First, the terms in each document are ranked based on the values of the weighing function. Then, the terms in the first position of the ranked list of each document are sorted according to the number of times they occur in the rankings. In case of a tie, we order them according to the maximum weight between them. We then do the same for terms in the second position of ranking, in the third position, and so on. The process stops when the desired number of terms is reached. Even though the resulting list may be larger than the size sought, the ordered list enables us to get the exact number of terms from the top.

As an alternative dimensionality reduction method, we also compare Latent Semantic Indexing (LSI) [36]. LSI projects the initial space of documents and their words into a reduced vector space, where the mapping is performed in such a way that the independence is kept for terms that do not co-occur.

D. Clustering Algorithm

We chose Cluto rbr (k-way repeated bisections globally optimized) as the clustering algorithm [3] for our experiments. The number of clusters k is set to the number of categories in each dataset to make the evaluation process more intuitive. Having k set to the actual number of clusters enables to explore differences between representation approaches, leaving aside the effect of the selection of the number of clusters. The rest of the parameters are set to their default values.

E. Evaluation Measure

We use the F_1 score [37] as the evaluation measure (see Equation 2).

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

where precision and recall are defined as follows:

$$precision = \frac{|\{relevant\ docs\} \cap \{retrieved\ docs\}|}{|\{retrieved\ docs\}|} \tag{3}$$

$$recall = \frac{|\{relevant\ docs\} \cap \{retrieved\ docs\}|}{|\{relevant\ docs\}|} \tag{4}$$

From these, the F_1 score for each category can be computed. The overall F_1 score is computed as the weighted average of the F_1 scores for each category.

VI. Improving the Combination of Criteria

In this section, we evaluate the performance of the state-of-the-art fuzzy logic approach FCC. We also propose and evaluate two novel alternative approaches, EFCC and AddFCC.

A. Study of FCC and Individual Criteria

As an initial comparative study over the existing FCC approach, we propose and analyze four variations of this term weighing function, one for each criterion, in such a way that the output of the system will depend only on one criterion at a time. Table I shows an example rule base for a system that solely relies on the `emphasis` criterion to determine the output.

We used the MFT reduction given that it selects the highest weighted features without transforming them. This enables us to perform a fairer comparison of different term weighing approaches.

Table II and Fig. 2 show the results of each individual criterion compared to FCC, where each column shows the performance for different numbers of features ranging from 100 to 5,000, as well as the average.

For Banksearch, FCC outperforms all individual alternatives, showing the importance of the combination of criteria. Among the individual criteria, `frequency` performs best, while `position` is the worst.

The results for WebKB are quite different. On one hand, `frequency` is not always the best among individual criteria and, on the other hand, FCC does not always outperform individual criteria, specifically `title` and `emphasis` obtain equal or higher F1-measure values in some cases when the vector dimensions are reduced to $2,000$ and $5,000$, respectively.

In this collection, the frequency distribution of emphasized terms shows a more restricted use of emphasis. It could be due to the limited number of web domains and the similarity among web page contents that only come from Universities. These factors could limit the number of different writing styles, fact that would be reflected in a less scattered distribution of emphasized term frequencies. The same consideration about the restrictions on the creation of WebKB can explain the good results achieved by the `title` criterion. We can expect that authors use titles in a similar way to emphasis, as both resources are used to highlight important words. In the cases where `title` and `emphasis` lead to a better clustering, their combination with `frequency` and `position` harms the results. In particular, WebKB documents within categories can be much more heterogeneous than in Banksearch, fact that negatively affects the `frequency` criterion; the combination should help correct this problem, but it does not. Thus, it suggests that `frequency` and `position` are hindering the combination.

B. AddFCC and EFCC: Modifying the Knowledge Base

The first step to try to improve the fuzzy combination is to understand the bad performance of FCC in WebKB. In the rules of FCC [2], when `frequency` is *low*, the output can be *very high* (the maximum) depending on `position`, if `title` and `emphasis` are *high*. As we saw before, `frequency` contributes to a good clustering much more than `position`, so the output should reflect that fact. But, in this case, `frequency` is totally ignored. This occurs again when `title` is *low* and `frequency` *medium*. Both criteria are important for a good grouping, but the output is *very high* based on `position`, the same as the previous case. In these cases we are underestimating the discrimination power of `frequency` and `title`. The same happens when `frequency` is *medium*, being `title` and `emphasis` *low*: `position` decides again that `importance` can be the minimum or not, but `frequency` should count more than `position`.

On the other hand, the whole set of 31 rules in FCC makes the possible combinations more difficult to understand and evaluate. As the fuzzy system is able to combine the conclusions of the rules, an alternative that we propose is the use of a set of single-input rules for each criterion. Thus, the alternative system calculates the output by combining the different outputs of the fired rules. We refer to this approach AddFCC, whose rule base is shown in Table III, which reduces the number of cases that are set to the minimum to keep the rule set complete.

Since the reduced expressiveness of AddFCC system may give rise to mistakes due to a bad specification of the heuristic knowledge, we introduce another intermediate approach, Extended Fuzzy Combination of Criteria (EFCC). Its rule base combines some criteria explicitly and for others lets the combination to the fuzzy engine (see Table IV). It has two sets of rules: one for `frequency` and one for the rest of the criteria. This guarantees having at least one rule of each set fired by the system. This avoids underestimation of `frequency` while also reducing the discriminative power of `position`.

Table V and Fig. 3 show the clustering results for AddFCC and EFCC, which are compared to FCC. We observe that EFCC improves FCC clustering results in WebKB in all cases while AddFCC does not, while AddFCC outperforms the other approaches for Banksearch in all cases. Nevertheless, EFCC also achieves good results in Banksearch, particularly with small feature sets. AddFCC has the problem of considering all criteria equally important, and hence overestimating `position` in the combination, as we observed with FCC too.

At this point, we opted for EFCC as an alternative to FCC for our subsequent experiments. We also apply LSI and compare the results of EFCC with TF-IDF and FCC (see Table VI and Fig. 4).

Globally, EFCC MFT achieves the most stable results among collections, and is generally the best approach, with a few exceptions in Banksearch. If one is thinking of applying the representation to a new collection, EFCC MFT would be the best option. It requires fewer terms to achieve its optimal performance for balanced, homogeneous collections. This posits EFCC MFT as a suitable approach to be applied to new, unseen collections. Furthermore, the additive properties of the fuzzy system make it possible to reduce the number of rules needed to specify the knowledge base of EFCC and therefore, the system is easier to understand.

On the other hand, the good behavior of MFT depends on the term weighing function applied before. Because of this, we believe that the use of light dimension reduction techniques is a good alternative, at the price of selecting a proper term

IF	Title	AND	Frequency	AND	Emphasis	AND	Position	THEN	Importance
					High			⇒	Very High
					Medium			⇒	Medium
					Low			⇒	No

TABLE I: Rule base for the system based on `emphasis` criterion.

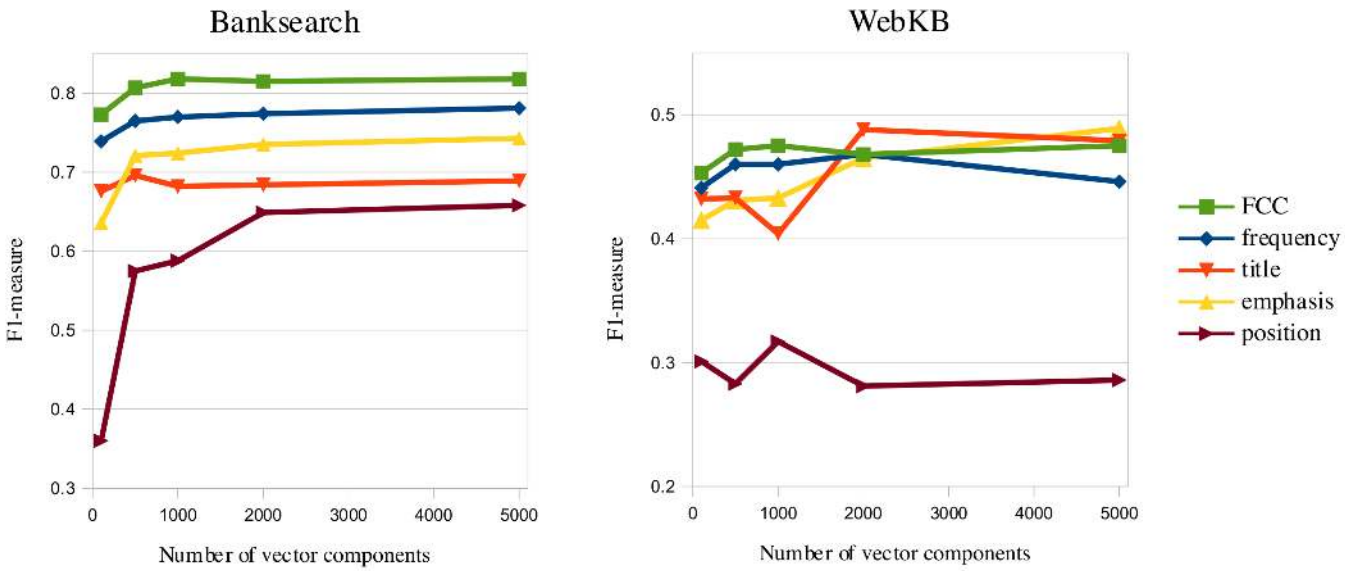


Fig. 2: Graphical representation of data in Table II.

Rep.	100	500	1,000	2,000	5,000	Avg.
Banksearch						
FCC	**0.723**	**0.757**	**0.768**	**0.765**	**0.768**	**0.756**
title	0.626	0.646	0.632	0.634	0.639	0.635
emphasis	0.586	0.671	0.674	0.685	0.693	0.662
frequency	0.689	0.715	0.720	0.724	0.731	0.716
position	0.310	0.525	0.538	0.599	0.608	0.516
WebKB						
FCC	**0.453**	**0.472**	**0.475**	0.468	0.475	**0.469**
title	0.432	0.433	0.404	**0.488**	0.479	0.447
emphasis	0.415	0.431	0.433	0.465	**0.489**	0.447
frequency	0.441	0.460	0.460	0.468	0.446	0.455
position	0.301	0.283	0.317	0.281	0.286	0.294

TABLE II: F1 results for criteria analysis experiments (all with MFT reduction).

weighing function, for the clustering problem to solve.

C. Incorporating Context: Criteria Beyond the Document Itself

Moving away from the sole use of the document's content itself, now we explore the application of two techniques to improve EFCC with contextual information: (1) Inverse Document Frequency, and (2) anchor texts.

1) Inverse Document Frequency (IDF): With IDF we incorporate information from the whole collection to the representation, which we do by using the product of both:

$$\text{EFCC-IDF}(t_i, d_j, D) = \text{EFCC}(t_i, d_j) \times \text{IDF}(t_i, D) \quad (5)$$

where t_i is a term, d_j a document, and D the whole corpus.

Looking at the Table VII and Fig. 5, EFCC-IDF works really well with over 500 features in Banksearch, but much worse with 100. WebKB EFCC IDF results are much worse in all cases. This is due to the penalization that IDF applies to common terms. In a clustering task, instead, we look for terms that are common across documents of the same group. Hence, this suggests that the combination of EFCC and IDF is not suitable for the purposes of a clustering task.

2) Anchor Texts: There are a number of ways of adding anchor texts to document representation methods. We are interested in elucidating whether anchor texts could help improve web page representation in clustering or not, but at the same time, we want to investigate different alternatives for

IF	Title	AND	Frequency	AND	Emphasis	AND	Position	THEN	Importance
	High							⇒	Very High
	Low							⇒	No
			High					⇒	Very High
			Medium					⇒	Medium
			Low					⇒	No
					High			⇒	Very High
					Medium			⇒	Medium
					Low			⇒	No
							Preferential	⇒	Very High
							Standard	⇒	No

TABLE III: Rule base for AddFCC. Inputs are related to normalized term frequencies.

IF	Title	AND	Frequency	AND	Emphasis	AND	Position	THEN	Importance
	High				High			⇒	Very High
	High				Medium		Preferential	⇒	High
	High				Medium		Standard	⇒	Medium
	High				Low		Preferential	⇒	Medium
	High				Low		Standard	⇒	Low
	Low				High		Preferential	⇒	High
	Low				High		Standard	⇒	Medium
	Low				Medium		Preferential	⇒	Medium
	Low				Medium		Standard	⇒	Low
	Low				Low		Preferential	⇒	Low
	Low				Low		Standard	⇒	No
			High					⇒	Very High
			Medium					⇒	Medium
			Low					⇒	No

TABLE IV: Rule base for EFCC. Inputs are related to normalized term frequencies.

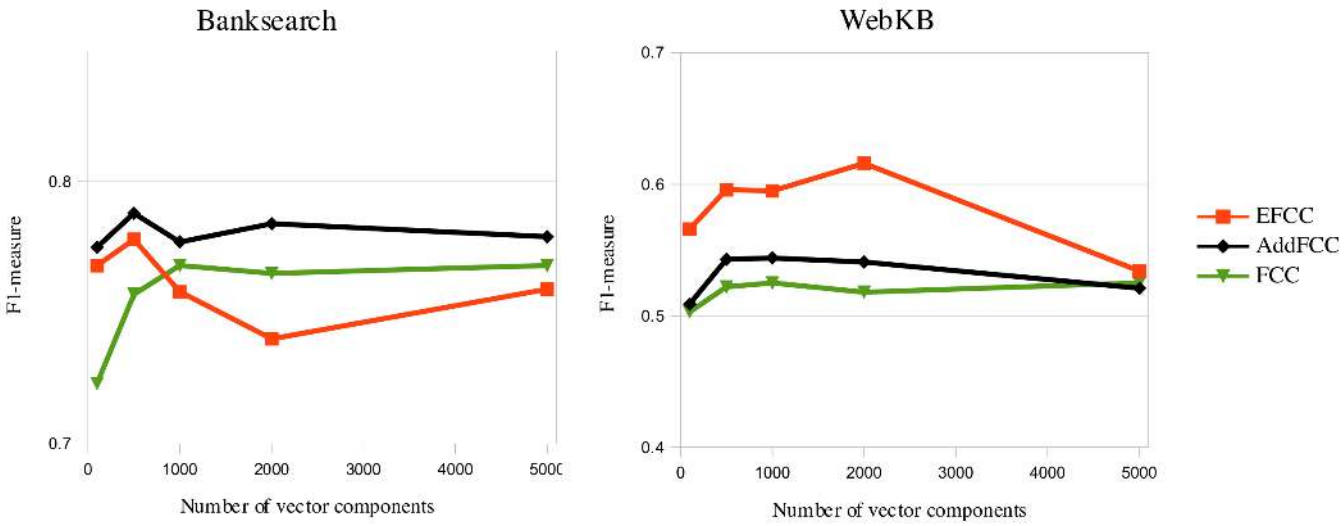


Fig. 3: Graphical representation of data in Table V.

Rep.	100	500	1,000	2,000	5,000	Avg.
Banksearch						
FCC	0.723	0.757	0.768	0.765	0.768	0.756
EFCC	0.768	0.778	0.758	0.740	0.759	0.760
AddFCC	**0.775**	**0.788**	**0.777**	**0.784**	**0.779**	**0.781**
WebKB						
FCC	0.453	0.472	0.475	0.468	0.475	0.469
EFCC	**0.516**	**0.546**	**0.545**	**0.566**	**0.484**	**0.532**
AddFCC	0.459	0.493	0.494	0.491	0.471	0.482

TABLE V: Fuzzy logic-based alternatives in terms of F1 (all with MFT reduction).

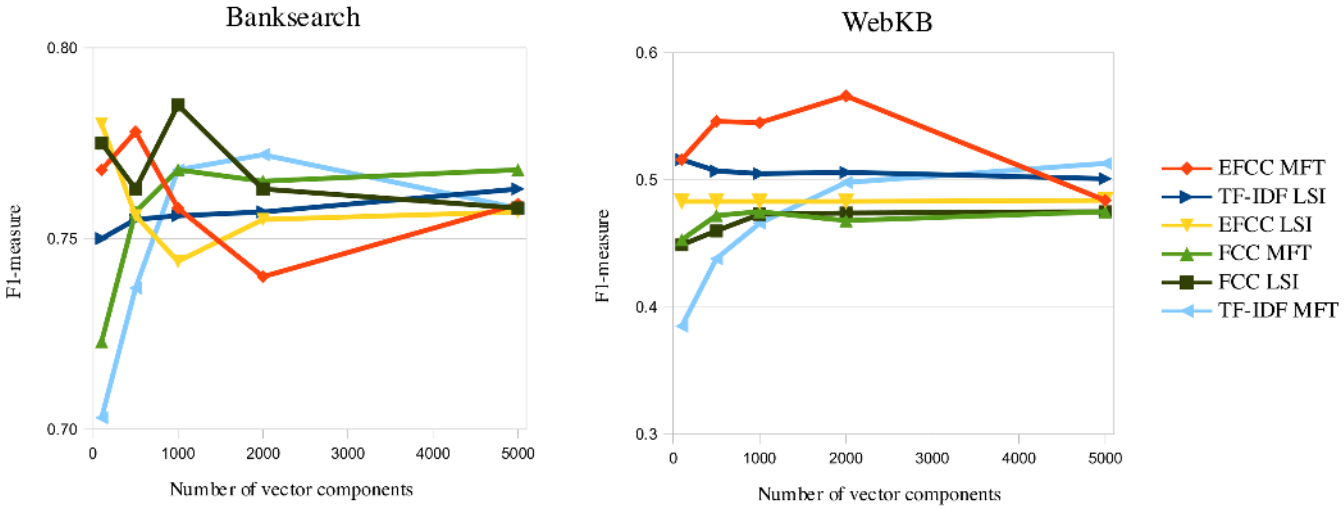


Fig. 4: Graphical representation of data in Table VI.

Rep.\Dim.	100	500	1,000	2,000	5,000	Avg.
Banksearch						
TF-IDF LSI	0.750	0.755	0.756	0.757	0.763	0.756
TF-IDF MFT	0.703	0.737	0.768	**0.772**	0.758	0.748
FCC LSI	0.775	0.763	**0.785**	0.763	0.758	**0.769**
FCC MFT	0.723	0.757	0.768	0.765	**0.768**	0.756
EFCC LSI	**0.780**	0.756	0.744	0.755	0.757	0.758
EFCC MFT	0.768	**0.778**	0.758	0.740	0.759	0.760
WebKB						
TF-IDF LSI	**0.516**	0.507	0.505	0.506	0.501	0.507
TF-IDF MFT	0.385	0.438	0.466	0.498	**0.513**	0.460
FCC LSI	0.449	0.460	0.473	0.474	0.475	0.466
FCC MFT	0.453	0.472	0.475	0.468	0.475	0.469
EFCC MFT	**0.516**	**0.546**	**0.545**	**0.566**	0.484	**0.532**
EFCC LSI	0.483	0.483	0.483	0.483	0.484	0.483

TABLE VI: F1 performance values for different dimensionality reduction methods with EFCC and other previous alternatives.

the combination within a term weighing function.

To analyze whether and how anchor texts can contribute to the document representation, we explore two different ways of incorporating them using EFCC:

a) Appended to the document's content itself, and hence contributing to the `frequency` criterion; and
b) Appended to the document's title, and therefore contributing at the same level as the title itself. These approaches considering anchor texts are in line with those described by Wang and Kitsuregawa [38] and Huang et al. [39].

We did the experiments in three different settings in each case:

1) Adding anchor texts;
2) Adding anchor texts and removing textual content from outlinks; and
3) Removing words that are frequently used across anchor texts, such as 'click', 'link' or 'homepage'.

We use the SODP dataset in these experiments, as it is the only dataset that includes anchor texts. As it is a new dataset not explored in previous sections, we also compare results with FCC and AddFCC.

Table VIII and Fig. 6 show the results of different alternatives using anchor texts. Each approach has a letter ('a' or 'b') and a number appended ('1', '2' or '3'), referring to the way in which anchor texts are exploited, as described above. The first three rows of the table show that EFCC outperforms FCC and AddFCC in all cases. This corroborates the limitations of FCC, and reinforces our motivation looking into an alternative approach where not all the criteria contribute equally to the combination. When it comes to the contribution of anchor texts, no approach improves EFCC clearly in all the cases, as the slight differences suggest when looking at the averages. Anchor texts do have a positive impact when we use vectors of small size, particularly when the terms in the anchor texts are considered as page titles (b alternative). However, as we increase the size of the vectors, anchor texts are not useful any more, leading to worse performance. Regarding the use of anchor texts as titles, the best option is to just add anchor texts as title terms (named b-1). The slight improvement achieved with anchor texts might not always pay off, given that the collection of anchor texts is a time consuming process.

Different reasons might explain the unsatisfactory results using anchor texts. The collection may have a link structure

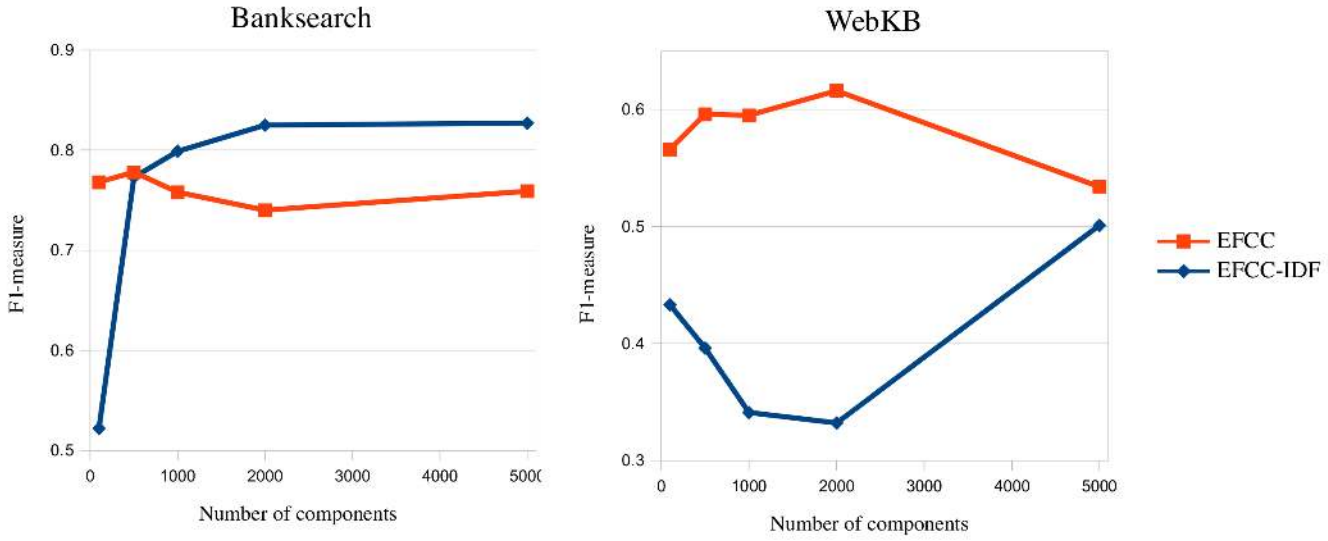


Fig. 5: Graphical representation of data in Table VII.

Rep.	100	500	1,000	2,000	5,000	Avg.
Banksearch						
EFCC	**0.768**	**0.778**	0.758	0.740	0.759	**0.760**
EFCC-IDF	0.522	0.773	**0.799**	**0.825**	**0.827**	0.749
WebKB						
EFCC	**0.516**	**0.546**	**0.545**	**0.566**	**0.484**	**0.532**
EFCC-IDF	0.383	0.346	0.291	0.282	0.451	0.350

TABLE VII: F1 results for EFCC IDF experiments (all with the MFT reduction method).

that is not sufficiently dense, or anchor texts might not be descriptive enough, hence not enabling to capture the topic of documents. This finding is in line with Eiron and McCurley [40] and Noll and Meinel [41], where authors posited that anchor text terms rather resemble terms used in search queries.

VII. AFCC: Analyzing and Tuning the Membership Functions

We now set forth a proposal to tune the membership functions, which leads to the definition of a revised and novel approach called Abstract Fuzzy Combination of Criteria (AFCC). We first perform a qualitative analysis of the membership functions that we are utilizing, and then we test AFCC, evaluating and analyzing its performance in comparison with the techniques studied previously.

A. Analysis of the Membership Functions

It is worthwhile considering that different datasets will have different frequency distributions for each criterion. Few terms in a collection tend to be in many documents, while many terms are used seldom. The effect of normalizing frequencies with respect to the most frequent term is that low values are compressed, and hence under-represented. This compression effect would exacerbate if the total maximum of the collection was used for the normalization process.

The fuzzy sets for FCC and EFCC were symmetrically defined, except for emphasis. Thus, some of the fuzzy sets defined for FCC and EFCC would match the initial state of most of the tuning processes of fuzzy rule-based systems.

In fact, what we call *high* or *low* are not absolute, but relative values. Therefore, a term is considered important because its normalized frequency is higher than most of the rest, and a certain value being *high*, *medium* or *low* depends on the frequency distribution of the dataset. In an ideal case, all term frequencies would be uniformly distributed between 0 and 1 (see Fig. 1a), configuring the basic parameters of the fuzzy set using the original heuristic information. However, the fact that texts tend to follow Zipf's law, suggests that the uniform distribution is not always the case and more sophisticated approaches are needed. Hence, we believe that each particular dataset should have its own features and tuning of membership functions.

B. Tuning of the Membership Functions

Given the limitations of FCC, EFCC and AddFCC to deal with varying term distributions across different datasets, we now delve into alternative considerations that further exploit these characteristics, which ultimately leads to the definition of AFCC. In order to automatically adjust the basic parameters of the membership functions, we assume the two base cases that both the words in the documents as well as the

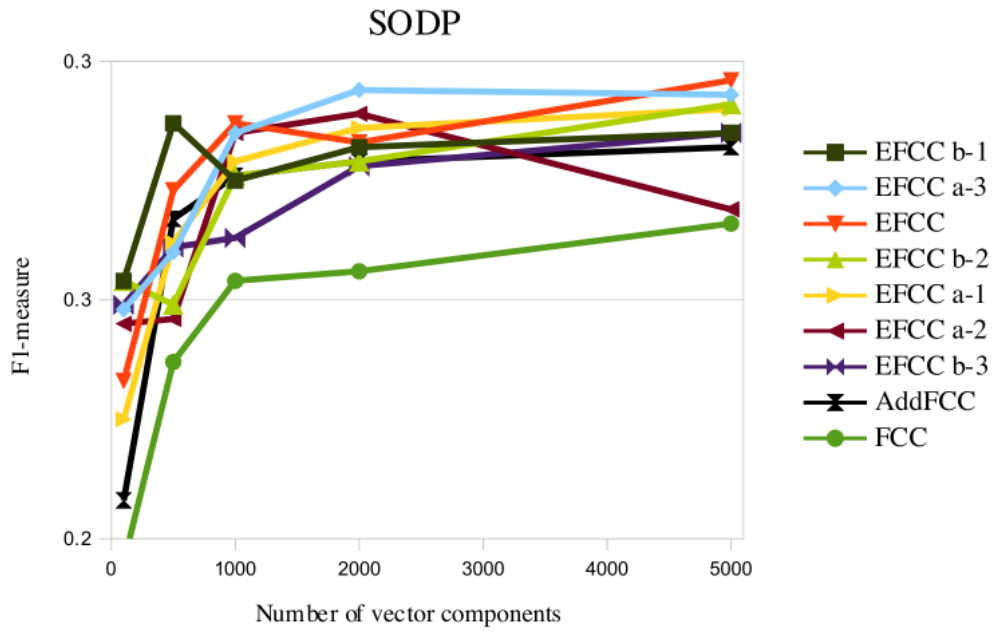


Fig. 6: Graphical representation of data in Table VIII.

Rep.	100	500	1,000	2,000	5,000	Avg.
SODP						
FCC	0.195	0.237	0.254	0.256	0.266	0.242
AddFCC	0.208	0.267	0.276	0.279	0.282	0.262
EFCC	0.233	0.273	**0.287**	0.283	**0.296**	0.275
EFCC a-1	0.225	0.262	0.279	0.286	0.290	0.268
EFCC a-2	0.245	0.246	0.285	0.289	0.269	0.267
EFCC a-3	0.248	0.260	0.285	**0.294**	0.293	0.276
EFCC b-1	**0.254**	**0.287**	0.275	0.282	0.285	**0.277**
EFCC b-2	**0.254**	0.249	0.276	0.279	0.291	0.270
EFCC b-3	0.249	0.261	0.263	0.278	0.285	0.267

TABLE VIII: F1 results for anchor text experiments (all with MFT reduction).

emphasized terms will approximate a Zipfian distribution, as defined by Zipf's law [42]. For the first base case based on the `frequency` criterion, we consider we have a distribution tending to a power law when the majority of terms, i.e., more than a half of them (55%) have normalized frequencies below 0.2. Depending on whether this condition is fulfilled or not, we set the membership functions with one of the following two alternatives:

1) When the precondition is fulfilled, we assume a distribution tending to a power law. As we need 5 intervals to build three sets (*low*, *medium* and *high* and two intersection areas between them, see Fig. 1a), our worst case would be to have only one possible value for each interval, that is, a maximum frequency of 5. Thus, to guarantee at least one possible value for the low set in that case, we chose the first interval from 0 to $1/5$. The rest of the intervals are selected using equidistant percentiles for term frequencies from $1/5$ to 1, because this is suitable for the normalized frequencies that we found in our test data;
2) When our precondition is not fulfilled, then we assume that the distribution tends to be closer to uniform, so that we can establish the fuzzy sets with the original heuristic, that is, all of them will have the same size. We use the corresponding percentiles to fit the distribution slightly better than using exact values (0.2, 0.4, 0.6, 0.8, see Fig. 1a). Notice that in case of a uniform distribution, the adjustment for the first case—distribution tending to Zipf's law—would lead to these exact values too, because as the distribution moves towards a uniform distribution, the percentile 0.2 will approximate to $1/5$ and the rest of the parameters belong to equidistant percentiles relative to this initial value in both cases. In those cases, the fuzzy sets would be symmetrical, that is, not only the case of the original sets in FCC and EFCC, but also the initial case used by most of the tuning methods of fuzzy rule-based systems.

With regard to the `emphasis` criterion, we follow the same precondition as with `frequency` to determine whether or not the distribution tends to a power law, but modifying the fitting

rules due to the different meaning of emphasis. Again, we have two alternatives for `emphasis`:

1) When the distribution tends to a power law, we set the first interval as in the `frequency` case, and the rest with decreasing percentiles, each being a half of the previous. The reason is that in the original heuristic-based fuzzy sets, the medium set is the biggest one, and we want to preserve the original heuristic knowledge, but always taking into account the relative difference between the number of elements in each set instead of absolute exact values;
2) If the collection does not fulfill our precondition, we assume that the distribution tends to be more uniform, so that we can establish the basic parameters of the membership functions by using the original heuristic rules but, as in the case of `frequency`, we use the percentiles instead of the exact values to fit slightly better the distribution (in this case the values were 0.05, 0.15, 0.55, 0.75, see Fig. 1c).

In the case of titles, we use the lowest value of the distribution to set the first interval, dividing the rest of the space in equidistant percentiles. Finally, it must be noted that we do not adjust the auxiliary system because the positions of words in a page do not depend on anything else than the number of words in the document.

We refer to this new approach we came up with after the analysis and tuning of the membership functions as Abstract Fuzzy Combination of Criteria (AFCC), which we test and evaluate next.

C. Empirical Analysis of AFCC

As AFCC represents a modification over the fuzzy logic based approaches, we use FCC and EFCC as baselines, as well as TF-IDF. We apply the MFT reduction in all cases to compare the weighing functions in the same conditions.

Table IX and Fig. 7 show F1 scores for these representations. On the one hand, looking at the results, among the fuzzy logic based representations, AFCC outperforms the rest in WebKB in all cases, while in Banksearch got better results than the others with 2 out of 5 vector sizes, having also a higher average F1 score. This varying performance across collections could be due to the fact that frequency distributions in Banksearch rather approximate a power law. In those cases, the least frequent terms are assigned to the low fuzzy set, with few terms remaining for the medium and high sets. This explains the small difference between the EFCC and FCC fixed sets. The same occurs with SODP, where EFCC and AFCC get similar results, although FCC performs worse, probably due to its underestimation of `frequency`. However, with a rather uniform term frequency distribution, as in WebKB, adjusting the fuzzy sets has a much bigger effect in results, where more terms are assigned to the medium and high fuzzy sets, and small variations of the basic parameters of the membership functions will have a much bigger effect. It is indeed important to adapt to this kind of distributions, as the terms are differently used and structured.

On the other hand, TF-IDF obtained surprisingly good results in SODP compared to the results of the same function with Banksearch and WebKB datasets. In general, results with all the representations tend to be worse in SODP, due to the special difficulties of this collection. We believe that the use of IDF could help improve the results of TF-IDF because it would alleviate the effect of the bigger categories, whose terms would be penalized giving more representativeness to those belonging to smaller categories. This fact would reduce slightly the bias introduced by the bigger categories, allowing to cluster the smaller ones slightly better. This improvement in the clustering of smaller categories could lead to an improvement in the overall clustering results of TF-IDF.

In general, adjusting the membership functions to the dataset seems to be useful not only to add more automatism to the document representation process, but also because this automation allows the system to adapt better to datasets with specific characteristics. The proposed method is able to achieve similar results to EFCC when dealing with exponential distributions. Moreover, when the shape of the distribution changes, the adjustment helps improve clustering results, as is the case of WebKB. Fig. 8 shows a summary of the resulting membership functions and the distributions of input values for each dataset and criterion.

D. Statistical significance

We analyze in depth the difference between using membership function tuning and the original representation with fixed fuzzy sets. Besides, we also include FCC in the comparison. We are interested in seeing the global improvements of the new proposal, AFCC, with respect to the original baseline. Each dataset was divided in 100 different sub-datasets 50% smaller than the original, where the size of each category is in proportion to the original ones. We performed 100 experiments per vector size corresponding to each sub-dataset, resulting in a total of 4,500 different clustering experiments. We calculated the statistical significance between F1 scores of each pair of representations (AFCC-FCC, EFCC-FCC, AFCC-EFCC) with a paired two-tailed t-test for each vector size.

In Table X and Fig. 9, for each vector size and representation we show the average F1 scores of the 100 clustering experiments (one per sub-dataset), and in Table X we also show the difference between the corresponding averages, and the *p*-value resulting from applying the statistical t-test between each pair of representations.

In most of the cases AFCC outperforms EFCC, and consequently also FCC. Therefore, the difference between term frequency distributions of the datasets, in combination with all of these results allow us to conclude that membership function tuning helps determine each criterion in a better way, ultimately improving clustering results.

Adjusting the membership functions to a dataset leads to results as good as or better than FCC in 91.6% of the cases, and as good as or better than EFCC in 86.5% of the cases. EFCC and AFCC outperform FCC in most of the cases, and between them, AFCC allows to improve the results of EFCC in 28.5% of the cases.

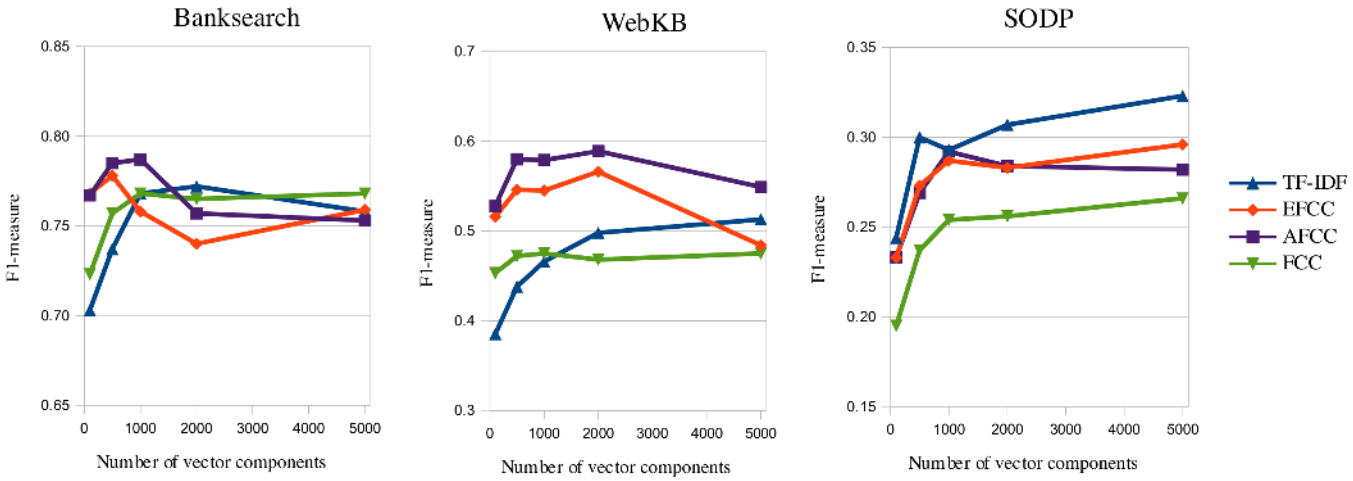


Fig. 7: Graphical representation of data in Table IX.

Rep.	100	500	1000	2000	5000	Avg.
Banksearch						
TF-IDF	0.703	0.737	0.768	**0.772**	0.758	0.748
FCC	0.723	0.757	0.768	0.765	**0.768**	0.756
EFCC	**0.768**	0.778	0.758	0.740	0.759	0.760
AFCC	0.767	**0.785**	**0.787**	0.757	0.753	**0.770**
WebKB						
TF-IDF	0.385	0.438	0.466	0.498	0.513	0.460
FCC	0.453	0.472	0.475	0.468	0.475	0.469
EFCC	0.516	0.546	0.545	0.566	0.484	0.532
AFCC	**0.528**	**0.580**	**0.579**	**0.589**	**0.549**	**0.565**
SODP						
TF-IDF	**0.244**	**0.300**	**0.293**	**0.307**	**0.323**	**0.293**
FCC	0.195	0.237	0.254	0.256	0.266	0.242
EFCC	0.233	0.273	0.287	0.283	0.296	0.275
AFCC	0.233	0.269	0.292	0.284	0.282	0.272

TABLE IX: F1 results for membership functions experiments (all with MFT reduction).

VIII. Discussion

We have studied the application of fuzzy logic for the representation of web documents in a way that imitates humans skimming through the documents. The use of fuzzy logic enables us to separate the knowledge declaration from the calculation procedure, which also enables us to specify the knowledge by means of a set of rules close to natural language applying non linear combinations of criteria. Building on a state-of-the-art unsupervised document representation, Fuzzy Combinations of Criteria (FCC), we have introduced, evaluated, and analyzed three alternatives that make the most of the HTML structure and content of web documents, namely EFCC, AddFCC, and AFCC. We evaluate and compare the representation approaches in a web page clustering task, using three datasets with very different characteristics.

We first defined a set of rules fixed on the basis of expert knowledge. Although there are other options to automatically generate these rules (the rule base could be adjusted by using machine learning techniques that adapt sets of rules to sets of sample data, or by using bio-inspired approaches), both approximations could cause a loss of generality in the learned/generated model in the attempt to fit the system to specific sample data. This could lead to illogical rules. On the other hand, in this automated scenario, we would need to deal with the coherence of the rules, which would require to establish a methodology to measure this coherence among rules. Last but not least, our system evaluates each term within each document using a fuzzy approach, which implies a high computational cost. Therefore, the use of machine learning or bio-inspired algorithms would add a considerable cost to the system. Of course, the manual definition of the rules employed in our work could lead to mistakes in the knowledge definition. However, in the same way, the application of machine learning or bio-inspired techniques would always require an initial knowledge to start the process.

Considering these aspects, we analyzed three challenges concerning web page representation for clustering: (1) the selection of feature sources to extract essential information from; (2) the term weighing functions to estimate the weight of each feature; and (3) the dimensionality reduction techniques to select the most representative features and to reduce the computational cost of the clustering.

For feature selection, we explored the application of new, mostly unstudied criteria to improve the representation with

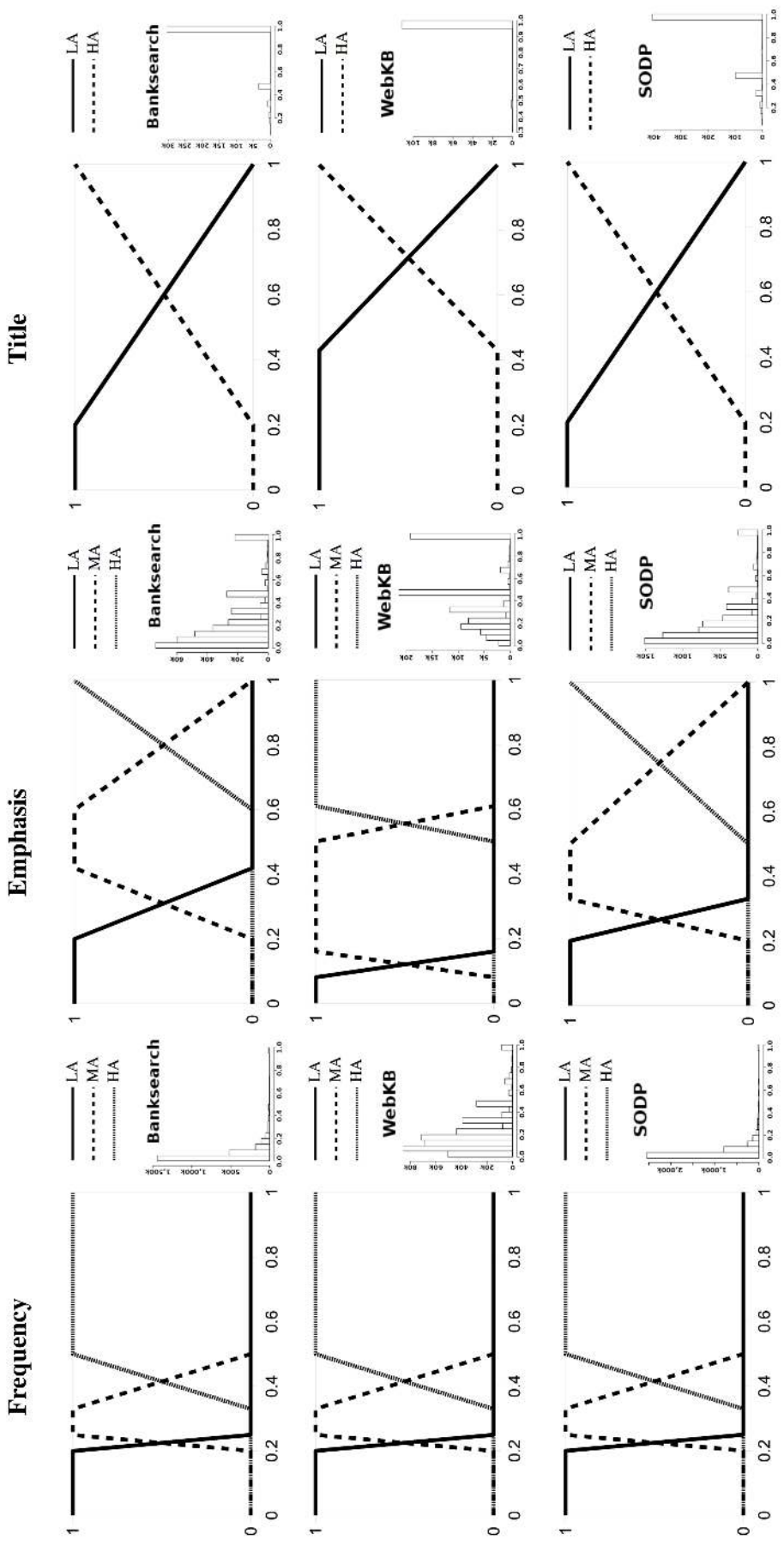


Fig. 8: Membership functions and distributions of input values for each criterion and dataset. Columns represent criteria, while rows represent datasets. Each pair of dataset and criterion includes two charts. The left-hand side, bigger chart, shows the final fuzzy sets for the membership functions after the automatic adjustment. The right-hand side, smaller chart, shows the distribution of term frequencies normalized by document in the dataset for each criterion, where X axis refers to the input value for the criterion, and Y axis refers to the number of terms that belong to that bin.

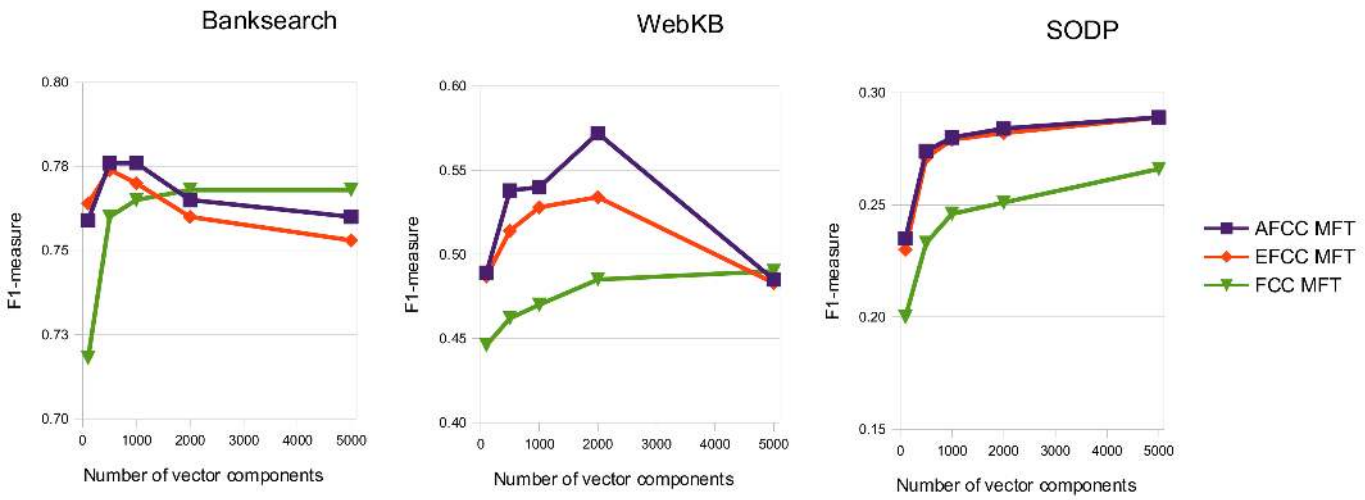


Fig. 9: Graphical representation of data in Table X.

		100	500	1000	2000	5000
		Banksearch				
F1-measure	AFCC MFT	0.759	**0.776**	**0.776**	0.765	0.760
	EFCC MFT	**0.764**	0.774	0.770	0.760	0.753
	FCC MFT	0.718	0.760	0.765	**0.768**	**0.768**
Difference	AFCC-FCC	0.041**	0.016**	0.011**	-0.003	-0.007**
	EFCC-FCC	0.047**	0.014**	0.006*	-0.008**	-0.015**
	AFCC-EFCC	-0.005**	0.003	0.005*	0.005**	0.008*
		WebKB				
F1-measure	AFCC MFT	**0.489**	**0.538**	**0.540**	**0.572**	0.485
	EFCC MFT	0.487	0.514	0.528	0.534	0.483
	FCC MFT	0.446	0.462	0.470	0.485	**0.490**
Difference	AFCC-FCC	0.043**	0.076**	0.070**	0.087**	-0.004
	EFCC-FCC	0.041**	0.051**	0.059**	0.049**	-0.007
	AFCC-EFCC	0.002	0.025**	0.011**	0.038**	0.002*
		SODP				
F1-measure	AFCC MFT	**0.235**	**0.274**	**0.280**	**0.284**	**0.289**
	EFCC MFT	0.230	0.271	0.279	0.282	**0.289**
	FCC MFT	0.200	0.233	0.246	0.251	0.266
Difference	AFCC-FCC	0.035**	0.040**	0.033**	0.033**	0.023**
	EFCC-FCC	0.030**	0.037**	0.033**	0.031**	0.023**
	AFCC-EFCC	0.005**	0.003*	0.000	0.001	0.000

TABLE X: Results for AFCC/EFCC/FCC t-test experiments. * indicates a statistically significant difference at $p < 0.01$, and ** indicates a statistically significant difference at $p < 0.001$.

information from the whole collection as well as from anchor texts. For term weighing we explored the fuzzy combination of criteria performed by FCC [2] aiming to get the most of the fuzzy system and the heuristics in which it is based. We use TF-IDF as the baseline, since it is a standard weighting method employed to represent documents. We presented an improved representation called EFCC, which outperformed the baselines, and another alternative called AddFCC, which did not work as well as expected. Both alternatives attempt to exploit the fuzzy system in a different manner to FCC, taking advantage of its additive properties. For dimensionality reduction, we introduced MFT, a lightweight dimensionality reduction technique, based on the term weighing function, which is able to improve the results of more complex techniques such as LSI when used together with EFCC in our test collections.

We also studied whether EFCC could be tuned to fit the specific characteristics of different collections. The aim of this adjustment is not only to improve clustering results in those collections, but also to adapt the representation to different datasets that could have different features. We found the case of the WebKB dataset, which has very different characteristics, particularly when looking at terms that are emphasized within the document contents. This led us to further study the tuning of the fuzzy system in an unsupervised way, for which we proposed AFCC. AFCC adjusts the basic parameters of the membership function on the basis of the term distributions of the collections. We showed that AFCC leveled or even improved the good results of EFCC and FCC in all kinds of datasets, outperforming the results of other approaches. Our results show that AFCC is a competitive approach that outperforms the rest of the techniques, with good performance

across datasets of very different characteristics.

Future work includes the study of the effect of non-linear scaling factors as a complementary tool to our proposal to adjust the representation to specific datasets, and to study new ways of considering the position criterion. A complementary analysis would include the exploitation of the position of words in documents through visual rendering of web pages. Finally, it would be interesting to study and assess the inclusion of additional criteria in the combination.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

REFERENCES

[1] X. Qi and B. D. Davison, "Web page classification: Features and algorithms," *ACM Computing Surveys*, vol. 41, no. 2, pp. 1–31, 2009.

[2] V. Fresno, "Representacion autocontenida de documentos html: una propuesta basada en combinaciones heuristicas de criterios [self-contained representation of html documents: an approach based on heuristic combinations of criteria]," Ph.D. dissertation, Universidad Rey Juan Carlos, 2006.

[3] G. Karypis, "CLUTO - a clustering toolkit," Tech. Rep. #02-017, Nov. 2003.

[4] O.-W. Kwon and J.-H. Lee, "Text categorization based on k-nearest neighbor approach for web site classification," *Information Processing and Management*, vol. 39, pp. 25–44, January 2003.

[5] K. Golub and A. Ardö, "Importance of html structural elements and metadata in automated subject classification," in *ECDL*. Springer, 2005, pp. 368–378.

[6] M. Fisher and R. Everson, "When are links useful? experiments in text classification," in *Advances in Information Retrieval*, 2003, vol. 2633, pp. 547–547.

[7] M. Kovacevic, M. Diligenti, M. Gori, and V. Milutinovic, "Visual adjacency multigraphs - a novel approach for a web page classification," in *Proceedings of the Workshop on Statistical Approaches to Web Mining*, 2004, pp. 38–49.

[8] L. K. Shih and D. R. Karger, "Using urls and table layout for web classification tasks," in *Proceedings of the 13th international conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 2004, pp. 193–202.

[9] P. Bohunsky and W. Gatterbauer, "Visual structure-based web page clustering and retrieval," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 1067–1068. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772807

[10] V. Bartík, "Text-based web page classification with use of visual information," in *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2010, Odense, Denmark, August 9-11, 2010*, 2010, pp. 416–420. [Online]. Available: http://dx.doi.org/10.1109/ASONAM.2010.34

[11] C. Herzog, I. Kordomatis, W. Holzinger, R. R. Fayzrakhmanov, and B. Krüpl-Sypien, "Feature-based object identification for web automation," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ser. SAC '13. New York, NY, USA: ACM, 2013, pp. 742–749. [Online]. Available: http://doi.acm.org/10.1145/2480362.2480504

[12] F. Gasparetti, A. Micarelli, and G. Sansonetti, "Mining navigation histories for user need recognition," in *HCI International 2014 - Posters' Extended Abstracts - International Conference, HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014. Proceedings, Part I*, 2014, pp. 169–173.

[13] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting wikipedia as external knowledge for document clustering," in *KDD*, 2009, pp. 389–396.

[14] H. Li, G. Sun, B. Xu, L. Li, J. Huang, K. Tanno, W. Wu, and C. Xu, "An information classification approach based on knowledge network," in *IEEE 8th International Symposium on Embedded Multicore/Manycore SoCs, MCSoC 2014, Aizu-Wakamatsu, Japan, September 23-25, 2014*, 2014, pp. 3–8. [Online]. Available: http://dx.doi.org/10.1109/MCSoC.2014.10

[15] T. Cassidy, H. Ji, L.-A. Ratinov, A. Zubiaga, and H. Huang, "Analysis and enhancement of wikification for microblogs with context expansion." in *COLING*, vol. 12, 2012, pp. 441–456.

[16] C. Lin and T. Hong, "A survey of fuzzy web mining," *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, vol. 3, no. 3, pp. 190–199, 2013. [Online]. Available: http://dx.doi.org/10.1002/widm.1091

[17] X. Wang, D. Luo, and H. He, "An improved feature weighted fuzzy clustering algorithm with its application in short-term prediction of wind power," in *Pattern Recognition*. Springer, 2014, pp. 575–584.

[18] Y. Zhou, H.-F. Zuo, and J. Feng, "A clustering algorithm based on feature weighting fuzzy compactness and separation," *Algorithms*, vol. 8, no. 2, pp. 128–143, 2015.

[19] A. Ensan and Y. Biletskiy, "Matching semi-structured documents using similarity of regions through fuzzy rule-based system," in *Advances in Data Mining. Applications and Theoretical Aspects - 13th Industrial Conference, ICDM 2013, New York, NY, USA, July 16-21, 2013. Proceedings*, 2013, pp. 205–217.

[20] A. Molinari and G. Pasi, "A fuzzy representation of html documents for information retrieval systems," *Fuzzy Systems*, vol. 1, pp. 107–112, 1996.

[21] V. Fresno and A. Ribeiro, "An analytical approach to concept extraction in html environments," *J. Intell. Inf. Syst.*, vol. 22, no. 3, pp. 215–235, 2004.

[22] R. Naseem, S. Anees, K. Muneer, and K. S. Farook, "Near duplicate web page detection with analytic feature weighting," in *ICACC*. IEEE, 2013, pp. 324–327.

[23] D. H. Kraft, E. Colvin, G. Bordogna, and G. Pasi, "Fuzzy information retrieval systems: A historical perspective," in *Fifty Years of Fuzzy Logic and its Applications*. Springer, 2015, pp. 267–296.

[24] A. Kolonin, "Automatic text classification and property extraction: Applications in medicine," in *SIBIRCON*, 2015.

[25] A. A. Hopgood, *Intelligent Systems for Engineers and Scientists*. Taylor & Francis, 2011.

[26] M. P. Sinka and D. W. Corne, "The banksearch web document dataset: investigating unsupervised clustering and category similarity," *J. Netw. Comput. Appl.*, vol. 28, pp. 129–146, April 2005.

[27] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to construct knowledge bases from the world wide web," *Artif. Intell.*, vol. 118, pp. 69–113, April 2000.

[28] A. Zubiaga, R. Martínez, and V. Fresno, "Getting the most out of social annotations for web page classification," in *ACM DocEng*, 2009, pp. 74–83.

[29] L. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," 2008.

[30] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2005, pp. 507–514.

[31] R. Kala, A. ShuLkla, and R. Tiwari, "Automatic text classification and property extraction: Applications in medicine," in *2009 WEE International Advaniee Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009*, 2009, pp. 541–545.

[32] J. Xu, H. He, and H. Man, "Dcpe co-training for classification," *Neurocomput.*, vol. 86, pp. 75–85, Jun. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2012.01.006

[33] F. Wanga, R. Li, Z. Lei, X. S. Ni, X. Huod, and M. Chena, "Kernel fusion-refinement for semi-supervised nonlinear dimension reduction," *Pattern Recognition Letters*, vol. 63, no. 1, pp. 16–22, October 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865515001671

[34] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 333–342.

[35] J. Xu, G. Yang, H. Man, and H. He, *Advances in Neural Networks – ISNN 2013: 10th International Symposium on Neural Networks, Dalian, China, July 4-6, 2013, Proceedings, Part I*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. L1 Graph Based on Sparse Coding for Feature Selection, pp. 594–601.

[36] T. K. Landauer, P. W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, no. 25, pp. 259–284, 1998.

[37] C. J. Van Rijsbergen, "Foundations of evaluation," *Journal of Documentation*, vol. 30, pp. 365–373, 1974.

[38] Y. Wang and M. Kitsuregawa, "Evaluating contents-link coupled web page clustering for web search results," in *CIKM*, 2002, pp. 499–506.

[39] S. Huang, Z. Chen, Y. Yu, and W. Ma, "Multitype features coselection for web document clustering," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 4, pp. 448 – 459, 2006.

[40] N. Eiron and K. S. McCurley, "Analysis of anchor text for web search," in *SIGIR*, 2003, pp. 459–460.
[41] M. G. Noll and C. Meinel, "The metadata triumvirate: Social annotations, anchor texts and search queries," in *Proceedings of the WI-IAT*, vol. 1, 2008, pp. 640–647.
[42] G. K. Zipf, "Human behavior and the principle of least effort." 1949.

Alberto P. García-Plaza was born in Madrid, Spain. He received the B.S. and M.S. degrees in Computer Engineering from Universidad Rey Juan Carlos (URJC), Madrid, Spain, in 2003 and 2006 respectively; he got the M.S. degree in Telematics and Computer Systems from URJC in 2007, and the PhD degree in computer science from the National University of Distance Education (UNED), Madrid, Spain, in 2012. From 2007 to 2012 he was a Teaching Assistant with the UNED Department of Computer Systems and Languages, being also a member of the UNED group in Natural Language Processing and Information Retrieval. From 2013 he has been working as a Research & Development Engineer for 4IQ, developing cyber-intelligence solutions based on state-of-the art IR approaches for their customers.

His main research interests are document representation for clustering related tasks, fuzzy logic, information retrieval and social media mining.

Víctor Fresno was born in Madrid, Spain. He received the B.S. and M.S. degrees in theoretical physics from the Autonomous University of Madrid (UAM), Madrid, Spain, in 1999; he got the M.S. degree in telecommunication engineering from the Polytechnic University of Madrid (UPM), Madrid, Spain, in 2004, and the Ph.D. degree in computer science from the King Juan Carlos University (URJC), Madrid, Spain, in 2006. From 2000 to 2001, he was a Research Assistant at the Spanish National Research Council (CSIC). Afterwards, from 2001 to 2007, he was a Teaching Assistant and Lecturer at URJC. Since 2007 he is an Associate Professor at the Department of Lenguajes y Sistemas Informáticos (LSI) at the National University of Distance Education (UNED) in Madrid, Spain.

He is the author of more than 60 articles, and his research interests include document representation models for classification/clustering and information retrieval, and fuzzy logic and NLP tools and techniques for text mining and social media information analysis.

Raquel Martínez Unanue was born in Portugalete, Basque Country, Spain. She received the B.S and M.S. degrees in computer science from Deusto University, Bilbao, in 1985; she got the Ph.D. degree in computer science also from Deusto University in 2000. She has a wide experience in teaching and researching in several Spanish universities, Cádiz University (UCA) in Cádiz, University Complutense of Madrid (UCM), King Juan Carlos University (URJC) in Madrid, and since 2005 is Associate Professor at the Department of Lenguajes y Sistemas Informáticos (LSI) at National Distance Learning University (UNED), in Madrid, Spain.

She has been project manager of several competitive research projects. She is author of more than 70 articles in different conferences and journals. Her research lines revolve around text mining, specially multilingual document clustering, as much in document representation as in clustering algorithms, application of NLP techniques, and modelling disambiguation problems as clustering problems.

Arkaitz Zubiaga was born in Arrasate, Basque Country, Spain. He received the B.S. and M.S. degrees in Computer Engineering from Mondragon University in 2006, the MSc in Language Technologies on the Web from National University of Distance Education (UNED) in 2008, and he got the PhD degree in Computer Science from National University of Distance Education (UNED) in 2011. He is a postdoctoral research fellow at the University of Warwick.

His research interests revolve around social media mining, social computing, computational journalism, computational social science and human-computer interaction. He is interested in researching the spread of news and events through social media, and especially in the role of citizen journalists in news reporting. He has conducted research at different institutions in 5 countries including the UK and the US, being involved in the organisation of workshops and conferences.