International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 5 (2013) pp. 472-479
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Fuzzy-DDE: a fuzzy method for the extraction of document cluster descriptors

Tatiane M. Nogueira[1], **Heloisa A. Camargo**[2] **and Solange O. Rezende**[1]

[1]Institute of Mathematics and Computer Science,
University of São Paulo - São Carlos,
P.O. Box 668, 13560-970 São Carlos, SP, Brazil.
{tatiane,solange}@icmc.usp.br

[2]Department of Computer Science,
Federal University of São Carlos,
P.O. Box 676, 13565-905 São Carlos, SP, Brazil.
heloisa@dc.ufscar.br

Abstract: **Good cluster descriptors facilitate the efficient storage and retrieval of information. In particular, when the imprecision and uncertainty of textual information is considered, the extraction of cluster descriptors, which represent the compatibility of a document with a cluster in a more precise way, is a challenging problem. Therefore, in this paper we present the method named Fuzzy-DDE (Fuzzy method for document descriptors extraction), by which two issues are addressed: (1) how to consider the imprecision and uncertainty present in the document clustering and (2) how to extract cluster descriptor from this kind of information. The experimental evaluation shows that the insertion of the information about the compatibility of a document with a cluster improves the fuzzy cluster descriptor extraction. Furthermore, the proposed method demonstrate its usefulness and effectiveness not only as a descriptor extraction, but also as a feature selection method for document categorization.**

Keywords: fuzzy clustering, text mining, cluster descriptor, information retrieval.

I. Introduction

Clustering techniques have been widely used to solve text mining and information retrieval problems, such as the ever-increasing amount of textual documents available online [1]. Using this technique, clusters can be found directly from the data without relying upon background knowledge [2]. Therefore, a document collection organized into clusters is very useful for users of information retrieval systems.

However, most document clustering methods still suffer from challenges in dealing with the problems of high dimensionality, scalability, accuracy, and meaningful cluster descriptors [3]. Moreover, these issues are even more challenging when addressed the problem of imprecision and uncertainty of textual documents. The imprecision and uncertainty are present in all textual information, since writers or readers deal with texts from different perspectives and representation of the document content when organizing them. Some of the clustering algorithms are suitable to the management of imprecision and uncertainty in textual document organization by allowing the assignment of documents to more than one cluster.

To illustrate the usefulness of such a flexibility, consider a context in which news are to be organized according to their main topic, each topic being identified by one or more descriptors. Consider a news (textual document) with the title "*Experts affirm the adventure sport strengthens heart health*", which discusses complementary topics: *Sports* and *Health*. This news can be assigned to distinct clusters: the cluster whose descriptors represent the *Sports* topic or the cluster whose descriptors represent the *Health* topic. Nevertheless, the cited news deals with both topics simultaneously, what suggests that the assignment of this news to both clusters would be a better option than choosing one of them.

One way to provide the assignment of documents to more than one cluster simultaneously is by means of fuzzy clustering. Fuzzy clustering algorithms scatter a document collection so that each document may belong to different clusters with different membership degrees. The interpretation of these membership degrees can be used to quantify the compatibility of a document with a particular topic.

Therefore, in this paper we present an improved version of the method proposed in [4] that extracts cluster descriptors which represent the compatibility of a document with a cluster in a more precise way. Firstly, we provide an overview about fuzzy clustering and cluster descriptor extraction. Secondly, the new method for extracting fuzzy cluster descriptors is presented. The novelty in this proposal is that the new version of the method, named Fuzzy-DDE (Fuzzy method for document descriptors extraction), evaluates the relevance of a term to identify a cluster considering the degree of membership of the documents in the clusters. Third, some experiments are carried out on a document collection by comparing our proposal with a state-of-the-art method and our method presented in [4] in terms of the quality of the extracted descriptors for document categorization.

To evaluate the proposed method, experiments were performed using the Opinosis customer reviews document collection [5]. The Opinosis documents were organized in a flexible way by means of fuzzy clustering. The membership degree of a document to a cluster represent the compatibility degree between the document and the topic associated to that cluster. The topics are represented by the cluster descriptors extracted using the Fuzzy-DDE method.

The remainder of this paper is organized as follows. Section II describes the main issues related to fuzzy document clustering and fuzzy cluster descriptor extraction, citing some related work. Section III presents the method for extracting fuzzy cluster descriptors and explains the basic concepts on which our method relies, followed by an evaluation of the method proposed and discussion of the experimental results in Section IV. Finally, Section V concludes the paper and points future directions of this research.

II. Related work

Document clustering is a technique commonly used to organize a document collection, since by clustering is possible: i) to identify the similarity among documents, ii) to create a document hierarchy, vi) to produce a document classifier and v) to summarize a document collection [6]. Moreover, according to Bordogna and Pasi in [7], within the text mining area, document clustering is one of the most effective techniques to organize documents in an unsupervised manner.

According to Jayabharathy et. al in [8], a good document clustering method can assist computers in organizing the document collection automatically into meaningful clusters, since if documents are well clustered, searching within the cluster with relevant documents improves efficiency and reduces the time for search. Moreover, in applications where document clustering is used for information retrieval, good cluster representatives are as important as a good clustering [9]. Good cluster representatives facilitate the efficient storage and retrieval of information. However, the extraction of good representatives is a challenging problem, since documents are represented by a high dimensional feature space.

In cluster analysis, in general, the extraction of the cluster representatives occurs naturally because the representative candidates are probabilistic models or cluster prototypes. However, in textual document clustering, representatives such as the cluster prototype are not very useful to identify the topic or the subject addressed in the textual documents in each cluster.

The document clusters are better identified by descriptors, which are terms that are present in the documents and significant to the topic of the documents. Although we refer to terms representing the documents in the clusters as descriptors, usually the task of cluster descriptor extraction is also called cluster labeling, cluster naming, label identification, topic discovery, cluster description, and descriptive clustering [10, 8].

There is a document clustering research field that aims to develop methods that deal with multi-topic documents, which are usually addressed by clustering algorithms that are designed to produce overlapping clustering solutions. Cluster overlapping is achieved by documents that share terms or phrases with other documents, reason why a document may be assigned to more than one cluster [11]. Fuzzy clustering algorithms are examples of approach by which documents are assigned to multiple clusters simultaneously and relationships among the domains can be found [12, 13, 14, 15, 16, 17, 7, 18]. In this context, some methods have been proposed and shown to be effective in finding overlapping information. Some recent methods are brief cited as follows.

A hierarchical fuzzy clustering algorithm for dynamically supporting information filtering was proposed by Bordogna et. al in [19]. According to them, users can have either general or specific interests depending on their profile. Therefore, they must be provided with documents belonging to the categories of interest that can correspond to either a high level topic, such as sport news, or to a subtopic, such as football news, or even to a very specific topic such as football matches of their favorite team. The authors proposed an hierarchical structure in which clusters are automatically identified. Each level of the hierarchy corresponds to a distinct level of overlapping of the clusters. Therefore, in the upper levels of the hierarchy the value of overlapping increases, since the topics represented in these levels are more general, and then, fuzzier.

Deng et. al proposes in [20] an improved fuzzy clustering-text method based on the Fuzzy C-means (FCM) clustering algorithm and the edit distance algorithm. The authors used the feature evaluation to reduce the dimensionality of text vectors. Due to the boundary value attribution of the traditional FCM, the authors recommend the edit distance algorithm. The results obtained from this approach demonstrated that the improved algorithm can be applied to the text clustering, making the clustering results more stable and accurate than the traditional FCM clustering algorithm.

Chen et. al proposed in [3] an effective Fuzzy Frequent Itemset-Based Hierarchical Clustering (F^2IHC) approach, which uses fuzzy association rule mining algorithm to improve the clustering accuracy of Frequent Itemset-Based Hierarchical Clustering (FIHC) method. In the approach a fuzzy association rule mining algorithm for text is employed to discover a set of highly-related fuzzy frequent itemsets, which contain key terms to be regarded as the labels of the candidate clusters.

Song et. al proposes in [21] a weighted conceptual model for document presentation, since, according to the authors, document clustering techniques mostly rely on single term analysis which can not reveal the potential semantic relationship among terms. The proposed model divides the document concepts into centroid concepts and peripheral concepts due to their semantic relations to the subject. A fuzzy semantic clustering method was proposed based on the new semantic model to better capture the semantic subject of the documents.

According to Carmel et. al in [22], a lot of research has being done on fuzzy clustering algorithms and their applications in information retrieval and text mining. However, compared with clustering algorithms, little work has been done on fuzzy cluster descriptor extraction. This is due to the fact that the documents can belong to more than one cluster. Therefore, it is more difficult to obtain good descriptors for fuzzy clusters.

In general, there are two types of cluster description extraction: (1) descriptors are extracted from the collection of documents before the document clustering or (2) terms are extracted after document clustering. The first type is named DCF (Description Comes First), and the second one is named DCL (Description Comes Last) [10].

According to [10], cluster descriptors obtained before document clustering decreases the readability of clustering description. Therefore, in this paper we present a DCL method by which descriptors of fuzzy clusters are extracted considering the information about the compatibility of a document with a cluster obtained after the fuzzy document clustering.

III. The Fuzzy-DDE method

The selection of an appropriate number of clusters to organize a given set of documents is a difficult task because it is usually necessary for clustering documents with overlapping information to be represented by many terms.

Besides the selection of an appropriate number of clusters, an efficient organization of documents using clusters should ensure the relationship among the documents from different clusters. This issue is related to the selection of cluster descriptors, which are meant to identify the topic of each cluster. However, the extraction of cluster descriptors is very challenging in the document organization using fuzzy clustering, since the same descriptor can be representative for more than one cluster with different weights of representativity.

In order to overcome these problems, we have proposed in [4] a method to extract fuzzy cluster descriptors. Therefore, we aim to improve the previous method extracting cluster descriptors that better represent the compatibility of a document with a cluster.

As in the previous version of the method [4] the documents are preprocessed and clustered using the Fuzzy c-means algorithm. After that, since we are proposing a DCL (Description Comes Last) method, the most representative terms are selected as cluster descriptors.

All terms present in the documents are descriptors candidates. The representativity of each term with relation to each cluster is calculated by the classic measures of information retrieval [23] (precision, recall and f-measure). The default balanced f-measure, which equally weights precision and recall, was used. This balanced f-measure is commonly written as $f1$-measure [23]. These measures consider the contingency matrix showed in Table 1, for each descriptor candidate $t = 1, ..., T$, where T is the number of terms of the collection, and each cluster $c = 1, ..., C$, where C is the number of clusters.

Table 1: Contingency matrix for information retrieval measurement

	Documents of cluster c	Documents that are not in cluster c
Documents which have the descriptor candidate t	*hits*	*noises*
Documents which do not have the descriptor candidate t	*losses*	*rejects*

By means of a fuzzy clustering, documents can belong to more than one cluster. Therefore, considering a document collection $D = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_K\}$, where K is the number of documents of the collection, the membership degree of document $\mathbf{d}_k$, in the c-th cluster is $A_c(\mathbf{d}_k)$. A document $\mathbf{d}_k = [d_{k1}, d_{k2}, ..., d_{kT}]$ where $1 \leq k \leq K$, comprises the tf-idf of each term, *i.e.*, $\mathbf{d}_{kt} = tf\text{-}idf(t, \mathbf{d}_k)$ where $1 \leq t \leq T$. Where $tf - idf$ is the ratio between the frequency of a particular term in the collection and the inverse of the frequency of this term in the document (tf-idf Term Frequency-Inverse Document Frequency). By this measure, the importance of the terms in a document is weighted, so that terms which are present in a lot of documents have a smaller weight than the terms that occur more rarely in the collection.

In the fuzzy cluster descriptor extraction, a document $\mathbf{d}_k$ is considered belonging to the cluster c if it has a membership degree $A_c(\mathbf{d}_k) \geq s$, where $s = \frac{1}{C}$. The threshold s is considered for two reasons. Firstly, its use allows the selection of descriptors candidates from documents that belong to more than one cluster with different degrees, instead of considering only the cluster with the highest membership degree. Secondly, using this threshold it is possible to penalize the descriptor candidates that occur in documents that have low membership degree in a cluster.

By means of the Fuzzy-DDE, the efficiency of each descriptor candidate in identifying the documents in a cluster is evaluated including the membership degree of the documents in each cluster in the information retrieval measurement. This new form of evaluation was adopted based on the assumption that membership degree carries an additional information about the representativity of the terms that can contribute to a more precise evaluation of its relevance as descriptors of the cluster. Therefore, the extraction of the descriptors of a particular cluster begins with the calculation of the $f1$-measure of each descriptor candidate. A rank of terms weighted by their $f1$-measure is obtained for each cluster as follows.

i. Calculate the precision of a descriptor candidate t in a cluster c:

$$p(t,c) = \frac{(hits + losses) \cdot \sum_{k=1}^{K} \Theta(c, \mathbf{d}_k, t)}{hits + noises} \qquad (1)$$

in which $\Theta(\cdot)$ is defined in Equation (2).

$$\Theta(\alpha, \beta, \gamma) = \begin{cases} A_\alpha(\beta), & A_\alpha(\beta) \geq s \text{ and } \gamma \in \beta \\ 0, & A_\alpha(\beta) < s \end{cases} \qquad (2)$$

According to Equation (2), only the document β that has the term γ and belongs to the cluster α with membership degree, $A_\alpha(\beta)$, higher than or equal to the threshold s is considered in the measurement of the descriptor candidate weight.

ii. Calculate the recall of a descriptor candidate t in a cluster c:

$$r(t,c) = \sum_{k=1}^{K} \Theta(c, \mathbf{d}_k, t) \qquad (3)$$

in which $\Theta(\cdot)$ is defined in Equation (2).

iii. Calculate the $f1$-measure, the weighted harmonic mean of precision and recall of a descriptor candidate t in a cluster c:

$$f1(t,c) = \frac{2 \cdot p(t,c) \cdot r(t,c)}{p(t,c) + r(t,c)} \qquad (4)$$

The use of the membership degree in the evaluation of a descriptor candidate is useful. This measure warrants that the extracted descriptors are able to represent the information that a document can belong to more than one cluster with different compatibility degrees.

The defined number of descriptors is extracted for each cluster and a flexible organization of documents is obtained, in which the documents can belong to more than one cluster and each cluster is represented by descriptors obtained from the document collection.

Finally, the Fuzzy-DDE method obtains a clustering $G = \{(\sigma_1, g_1), ..., (\sigma_C, g_C)\}$, where each cluster g_i represents a collection topic ($g_i \subseteq D$) and σ_i represents its set of descriptors.

The next section presents a discussion of the experimental results to evaluate the proposed method.

IV. Evaluation and discussion of the experimental results

This section initially describes the knowledge domains and the experimental methodology adopted. The results from the experiments are then presented and analyzed.

The Fuzzy-DDE method to extract fuzzy cluster descriptors was evaluated using the Opinosis collection [24], which contains documents composed by customer reviews about characteristics of some products. The customer reviews in the collection were obtained from the websites: Tripadvisor.com, Amazon.com and Edmunds.com. Each one of these websites provides customer reviews about hotels, cars and electronics products, respectively.

Such documents are composed of sentences with high subjectivity, imprecision and uncertainty, since different documents with reviews about different characteristics of different products may have similar sentences. The goal of organizing these documents by means of fuzzy clustering is to find clusters of documents that present some similarity regarding topic reviews. For example, suppose there are two products: car and notebook. There are two reviews for these products in two different documents, respectively: “The speed limit of this car is very good” and “This notebook has a speed performance”. In this context, the topic of review “speed” is a common topic between these two products. In a flexible organization of documents as the one provided by the method proposed here, different products such as these can be assigned to the cluster represented by the review topic “speed”.

The documents in the Opinosis collection have their text already preprocessed. For example, the sentences separated by —” in the text presented in Table 2 does not have stopwords, which are words that are not relevant in the analysis of documents and usually consist of prepositions, pronouns, articles, interjections, among others. Moreover, the representative terms of these documents are composed by only one word.

Table 2: Preprocessed of an Opinosis review document

plug usb hub comput charg batteri charg cord design clever — page tru — page book — page — time chapter chapter — excit low batteri time — user replac batteri — bui extend warranti — year pai ship send devic amazon kindl replac — batteri chang — fact kindl sd card capabl batteri user — (...)

In our experiments the documents of the Opinosis collection were preprocessed again in order to compose terms by 2, 3 and 4 consecutive words. The Opinosis documents were preprocessed using the Pretext [25] tool, by which the representative terms of the documents were obtained. The terms were also stemmed, i. e., the terms were reduced to their root form in order to reduce the number of terms needed to represent the document collection. For example, the terms *battery* and *batteries* were reduced to *batteri*. Finally, the Opinosis collection was clustered by the execution of the Fuzzy C-Means [26] clustering algorithm, according to the approach proposed in [27].

The appropriate number of clusters to organize documents is often chosen by the repetitive execution of the clustering algorithm modifying the number of clusters in each execution and evaluating each execution according to some criterion of cluster evaluation [28].

In the experiments presented here seven clusters were chosen as the appropriate number of cluster to organize documents. This number was obtained evaluating the document clustering using an extension of a simplified version of the Average Silhouette Width Criterion [29], named Fuzzy Silhouette (FS) [30]. This method considers a balance between effectiveness and computational cost, besides using the degrees of membership and the data values in its calculation.

The descriptors of each fuzzy cluster were extracted by the Fuzzy-DDE method presented in Section III. In order to perform the comparative analysis, the fuzzy cluster descriptors were also extracted by the method based on centroids [31] and the method proposed by Nogueira et. al in [4].

Considering the fuzzy cluster descriptor extraction methods as methods of feature selection, the performance of the methods were measured applying some classification algorithms to categorize the documents from the attribute-value matrix obtained by the extracted descriptors. The document class was considered as the cluster in which it document has the higher membership degree, since there is not a previous classification of the documents.

For a vocabulary as large as the vocabulary of the Opinosis collection, a reasonable number of terms must be considered for the categorization of documents. Therefore, 100 descriptors were chosen to represent each cluster.

An example of the descriptors obtained by the method proposed by Nogueira et. al, and Fuzzy-DDE can be observed in Table 3.

The classification was carried out using a few well-known classification algorithms of machine learning implemented in the Weka tool [32]: Naive Bayes, Multinomial Naive Bayes, J48, SVM and KNN.

The SVM constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, measuring the complexity of hypotheses based on the margin with which they sep-

Table 3: Ten first descriptors of each cluster

Clusters	*Nogueira et. al*	*Fuzzy-DDE*
Cluster 1	long batteri life great, continu hold, solid feel, asu softwar, life continu	asu softwar, small screen, extra asu, keyboard respon, keyboard larg comfort type
Cluster 2	staff pleasant, room good, concierg servic, hotel perfect, nice furnish	room good, concierg servic, servic good, hotel perfect, room nice
Cluster 3	brighter screen video batteri, screen video batteri, gb version huge, small devic, version huge	brighter screen video batteri, screen video batteri, gb version huge, small devic, version huge
Cluster 4	easi read, friendli servic, servic good, locat excel, long time	easi read, free wine recept, servic good, locat excel, fun drive
Cluster 5	easi read, friendli servic, servic good, locat excel, long time	easi read, free wine recept, servic good, locat excel, fun drive
Cluster 6	easi read, friendli servic, servic good, locat excel, long time	easi read, free wine recept, servic good, locat excel, fun drive
Cluster 7	post speed, street speed, found map inaccur, downfal product, reason give star fact	post speed, street speed, navig downfal, downfal product, displai road

arate the data, not the number of features. Thus, the biggest advantage of SVM is its ability to learn independent of the dimensionality of the feature space. However, according to Shanahan and Roma in [33], the SVM, when applied to text classification, provides excellent precision, but poor recall.

The Naive Bayes classifier is based on the Bayes rule of conditional probability. It uses all the attributes contained in the data, and analyses them individually. According to Schneider in [34], this method is often used in text classification applications and experiments because of its simplicity and effectiveness.

The K-Nearest Neighbor (KNN) is an Instance-Based Learning (IBL) method. IBL approaches can construct a different approximation to the target function for each distinct query instance that must be classified. In fact, the KNN constructs a local approximation to the target function that applies in the neighborhood of the new query instance, and never constructs an approximation designed to perform well over the entire instance space. This has significant advantages when the target function is very complex. However, the disadvantage to KNN is that it typically considers all attributes of the instances when attempting to retrieve similar training examples from memory [35].

The C4.5 algorithm [36] is a predictive machine-learning model that decides the target value of a new sample based on various attribute values of the available data.

Experimental results obtained by Joachims in [35] show that SVMs consistently achieve good performance in text categorization tasks, outperforming the other compared methods. However, Gabrilovich and Markovitch in [37] demonstrate that in such datasets C4.5 significantly outperforms SVM and KNN, although the latter are usually considered substantially superior to text classifiers. According to the authors, when no feature selection is performed, C4.5 constructs small decision trees that capture the concept much better than either SVM or KNN. Furthermore, even when feature selection is optimized for each classifier, C4.5 formulates a powerful classification model, significantly superior to that of KNN and only marginally less capable than that of SVM.

In this context, as suggested in the literature, the chosen document categorization methods are good options to evaluate the ability of features selection methods to extract the information of a collection.

In this paper, the Naive Bayes, Multinomial Naive Bayes and J48 algorithms (the weka implementation of the C4.5 classification method) were executed using the default parameters of the Weka tool. However, the performance of the SVM was tuned up using the Normalized Polynomial Kernel and the complexity parameter c=2.0. The IBk (the weka implementation of the KNN classification method) was experimented ranging the number of neighbors from 1 to 7. The best result was obtained using 5 neighbors. The 5-fold cross validation method was used in all experiments. The performance rates (correct classification rate and standard deviation) obtained from each classifier are presented in Table 4. With these performance rates, the representativity of the descriptors obtained from the centroid-based, the Nogueira et. al, and the Fuzzy-DDE methods were checked.

Table 4: Correct classification rate

Algorithm	*Centroid-based*	*Nogueira et. al*	*Fuzzy-DDE*
SVM	54.60(14.17)	58.80(13.80)	**63.60**(14.25)
Naive	47.80(13.75)	47.40(12.26)	**53.20**(14.35)
M.Naive	60.40(12.28)	66.40(11.20)	**69.80**(14.36)
KNN-5	36.80(12.20)	**51.80**(13.95)	38.40(4.68)
J48	**52.00**(16.16)	50.80(13.83)	46.00(10.30)

From these results, it is possible to conclude that the Fuzzy-DDE is able to extract terms that are more representative for document categorization in the form of fuzzy cluster descriptors than the other ones. It is important to highlight that the high standard deviation was obtained because of the small number of documents used in the 5-fold cross-validation experiments.

Multiple comparisons among all methods were carried out to test whether there is a statistically significant difference between them. Our statistic test do not have the goal of test whether the newly proposed method is better than the existing ones, but to carry out a multiple comparison in which all possible pairwise comparisons need to be computed. Therefore, the Friedman test with Nemenyi's post-hoc was used [38, 39].

The average ranks obtained by applying the Friedman procedure are showed in Table 5. The Friedman statistic considering reduction performance (distributed according to chi-square with 2 degrees of freedom) was 1.6. The P-value computed by Friedman Test was 0.4493. The results achieved on post hoc comparisons for $\alpha = 0.05$ are showed in Table 6.

Table 5: Average Rankings of the methods

Method	Ranking
Fuzzy-DDE	1.6
Nogueira et. al	2
Centroid-based	2.4

The Nemenyi's post hoc procedure rejects those hypothe-

Table 6: P-values for $\alpha = 0.05$

Methods	$z = \frac{(R_0 - R_i)}{SE}$	p
Centroid-based vs. Fuzzy-DDE	1.264911	0.205903
Centroid-based vs. Nogueira et. al	0.632456	0.527089
Nogueira et. al vs. Fuzzy-DDE	0.632456	0.527089

ses that have a p-value ≤ 0.016667. Therefore, the null-hypothesis was not rejected with a 95% confidence level and the results demonstrated that it is not possible to detect statistically significant difference between the methods.

In addition to these results, the flexible organization of the Opinosis documents was obtained by the distribution of the documents in more than one cluster. Considering the membership degree of each document in each cluster as a degree of compatibility of a document with a topic represented by descriptors, the flexible organization of documents allows a document to belong to various topics simultaneously.

Furthermore, the performance of the Fuzzy-DDE method was measured considering it as a method of feature selection. Therefore, considering the results obtained, the proposed method demonstrates its usefulness and effectiveness also as a feature selection method for document categorization.

V. Conclusion

In this paper, we have proposed a method to extract fuzzy cluster descriptors based on the membership degrees obtained from the fuzzy clustering of documents. The experiments demonstrated that this is a promising method to deal with the problem of imprecision and uncertainty when organizing textual documents, since the information about the compatibility of a document with a cluster is considered in the cluster descriptor extraction.

Furthermore, the proposed method has the advantage for the fact that the organization allows the arrangement of the documents in topics in a totally unsupervised manner, *i.e.*, without expert domain or user's participation. Hence, this organization can be generalized and exploited by different users.

It is important to highlight that the analysis was conducted in 51 documents. Therefore, a different response about the representativity of the descriptors to the topics can be obtained when organizing a large document collection. However, it is highly unlikely that a lower representativity of descriptors for a large collection be obtained, since the number of clusters can also be increased proportionally to the number of documents.

Even though the proposed method has a good performance compared to the centroid-based and Nogueira et. al methods, it is necessary to consider the preprocessing of the documents because it can interfere on the results. Moreover, the time complexity about the use of 2, 3, and 4-gram to represent the document collection used in the experiments was not evaluated.

Therefore, it is very important to the process of analysis to select a reasonable number of terms to represent the document collection, making the set of terms more concise but no less representative in relation to the original set. Investigations in the future include experiments with different collections and document preprocessing.

Acknowledgment

Authors acknowledge the Brazilian research agencie CAPES for their support. This paper is also based upon work supported by FAPESP (Sao Paulo Research Foundation), Brazil, under the grant 2011/19850-9.

References

[1] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *In KDD Workshop on Text Mining*, 2000, pp. 1–20.

[2] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper, "Topic discovery based on text mining techniques," *Information Processing and Management*, vol. 43, pp. 752–768, 2007.

[3] C.-L. Chen, F. S. C. Tseng, and T. Liang, "Mining fuzzy frequent itemsets for hierarchical document clustering," *Information Processing and Management*, vol. 46, pp. 193–211, March 2010.

[4] T. Nogueira, S. Rezende, and H. Camargo, "Fuzzy cluster descriptor extraction for flexible organization of documents," in *11th International Conference on Hybrid Intelligent Systems (HIS)*, 2011, pp. 528–533.

[5] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China, 2010, pp. 340–348.

[6] J. Peltonen, J. Sinkhonen, and S. Kashi, "Discriminative clustering of text documents," *IEEE 9th International Conference on Neural Information Processing*, vol. 4, pp. 1956–1960, 2002.

[7] G. Bordogna and G. Pasi, "Soft clustering for information retrieval applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 138–146, 2011.

[8] J. Jayabharathy, S. Kanmani, and A. A. Parveen, "A survey of document clustering algorithms with topic discovery," *Journal of Computing*, vol. 3, no. 2, pp. 21–27, 2011.

[9] R. Feldman and J. Sanger, *The text mining handbook: Advanced approaches in analyzing unstructured data*, C. U. Press, Ed., 2007.

[10] C. Zhang, "Document clustering description based on combination strategy," in *2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC)*, 2009, pp. 1084–1088.

[11] A. Tagarelli and G. Karypis, "A segment-based approach to clustering multi-topic documents," in *Text Mining Workshop, SIAM Datamining Conference*, 2008, pp. 1–12.

[12] M. Mendes and L. Sacks, "Evaluating fuzzy clustering for relevance-based information access," in *The 12th IEEE International Conference on Fuzzy Systems*, vol. 1, 2003, pp. 648–653.

[13] K. Kummamuru, A. Dhawale, and R. Krishnapuram, "Fuzzy co-clustering of documents and keywords," in *The 12th IEEE International Conference on Fuzzy Systems*, vol. 2, 2003, pp. 772–777.

[14] M. E. S. M. Rodrigues and L. Sacks, "A scalable hierarchical fuzzy clustering algorithm for text mining," in *Proceedings of the 5th International Conference on Recent Advances in Soft Computin*, 2004.

[15] E. M. Rodrigues and L. Sacks, "Learning topic hierarchies from text documents using a scalable hierarchical fuzzy clustering method," in *International Conference on Recent Advances in Soft Computing*, 2005, pp. 269–274.

[16] K. Mizutani, R. Inokuchi, and S. Miyamoto, "Algorithms of nonlinear document clustering based on fuzzy multiset model," *International Journal of Intelligent Systems*, vol. 23, no. 2, pp. 176–198, 2008.

[17] R. Saracoglu, K. TuTuncu, and N. Allahverdi, "A new approach on search for similar documents with multiple categories using fuzzy clustering," *Expert Systems with Applications*, vol. 34, pp. 2545–2554, 2008.

[18] T. M. Nogueira, H. A. Camargo, and S. O. Rezende, "Fuzzy rules for document classification to improve information retrieval," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 3, p. 210?217, 2011.

[19] G. Bordogna, M. Pagani1, and G. Pasi, "A dynamic hierarchical fuzzy clustering algorithm for information filtering," in *Soft Computing in Web Information Retrieval*, ser. Studies in Fuzziness and Soft Computing, E. Herrera-Viedma, G. Pasi, and F. Crestani, Eds., 2006, vol. 197, pp. 3–23.

[20] J. Deng, J. Hu, H. Chi, and J. Wu, "An improved fuzzy clustering method for text mining," in *NSWCTC 2010 - The 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing*, vol. 1, 2010, pp. 65–69.

[21] S. Song, Z. Guo, and P. Chen, "Fuzzy document clustering using weighted conceptual model," *Information Technology*, vol. 10, pp. 1178–1185, 2011.

[22] D. Carmel, H. Roitman, and N. Zwerdling, "Enhancing cluster labeling using wikipedia," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '09. ACM, 2009, pp. 139–146.

[23] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[24] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: [http://archive.ics.uci.edu/ml]

[25] M. V. B. Soares, R. C. Prati, and M. C. Monard, "PRETEXT II: Description of restructuring tool preprocessing of texts," ICMC-USP, Tech. Rep. 333, 2008, (in portuguese).

[26] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[27] T. M. Nogueira, S. O. Rezende, and H. A. Camargo, "On the use of fuzzy rules to text document classification," *10th International Conference on Hybrid Intelligent Systems*, pp. 19–24, 2010.

[28] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. Wiley, 1999.

[29] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data–An Introduction to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics, 1990.

[30] R. Campello and E. Hruschka, "A fuzzy extension of the silhouette width criterion for cluster analysis," *Fuzzy Sets and Systems*, vol. 157, no. 21, pp. 2858 – 2875, 2006.

[31] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. Cambridge University Press, 2008.

[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[33] J. Shanahan and N. Roma, "Improving svm text classification performance through threshold adjustment," in *Machine Learning: ECML 2003*, ser. Lecture Notes in Computer Science, N. Lavrac, D. Gamberger, H. Blockeel, and L. Todorovski, Eds. Springer Berlin-Heidelberg, 2003, vol. 2837, pp. 361–372.

[34] K.-M. Schneider, "Techniques for improving the performance of naive bayes for text classification," in *In Proceedings of CICLing 2005*, 2005, pp. 682–693.

[35] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of ECML-98, 10th European Conference on Machine Learning*, C. Nédellec and C. Rouveirol, Eds. Heidelberg et al.: Springer, 1998, pp. 137–142.

[36] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[37] E. Gabrilovich and S. Markovitch, "Text categorization with many redundant features: using aggressive feature selection to make svms competitive with c4.5," in *Proceedings of the twenty-first international conference on Machine learning*. New York, NY, USA: ACM, 2004, pp. 41–49.

[38] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[39] J. Derrac, S. Garca, D. Molina, and F. Herrera, “A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms,” *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3 – 18, 2011.

Author Biographies

Tatiane Marques Nogueira was born in Feira de Santana, Bahia, Brazil, in 1982. She is graduated in Computer Science from State University of Southwest Bahia (UESB), master in Artificial Intelligence from the Federal University of São Carlos (UFSCar) and is currently a PhD student in Computational Intelligence at the University of São Paulo, ICMC-USP São Carlos. She has experience in Computer Science with emphasis on Logic and Semantics of Programs, acting on the following topics: Fuzzy Logic, Clustering and Text Mining.

Heloisa de Arruda Camargo was born in São Carlos-SP, Brasil, in 1956. She is graduated in Computer Science and master in Computer Science (1984) from ICMC-USP São Carlos, and PhD in Electrical Engineering (1993) from FEEC-UNICAMP, Campinas-SP. She is an associate professor at the Department of Computer Science, Federal University of So Carlos (UFSCar), where she works since 1980. In the period from 2001 to 2002, she did a postdoctoral work at the University of Alberta, Edmonton, AB, Canada. Her main research lines are: Genetic Fuzzy Systems, Semi-supervised Learning, Reasoning and Approximate Methods for Hybrid Fuzzy Modeling.

Solange Oliveira Rezende is graduated in Mathematics from Federal University of Uberlandia (1986), Master in Computer Science and Computational Mathematics from the University of São Paulo (1990) and Ph.D. in Mechanical Engineering from the University of São Paulo, São Carlos (1993). She is currently an associate professor at the University of So Paulo. She has experience in Computer Science with emphasis in Computer Methods and Techniques, and Artificial Intelligence, working mainly on issues related to data mining and text.