Keywords Extraction in Clusters of Related Documents

Leticia Arco, Damny Magdaleno, Rafael Bello,
Manuel Llanes and Libernys Valdés

Central University of Las Villas, Carretera a Camajuaní km 5 ½,
54830 Santa Clara, Villa Clara, Cuba
{leticiaa, dmg, rbellop, manuela, libernys}@uclv.edu.cu

Abstract. The aim of this work is to develop a model that allows the application of feature selection techniques for the extraction of relevant terms that characterize the clusters of related documents and discriminate among clusters. The main feature selection techniques are described as well as their applications to text mining, particularly the induction of decision trees in feature selection. We outline a flexible model that justifies the design and subsequent application of the stages that make up the proposed procedure, which are: the discretization of the features that describe the documents, the induction of the decision tree and the keywords extraction of textual homogeneous clusters. The feasibility of the developed model is demonstrated through its applications in three study cases using the CorpusMiner tool. The validation process comprised a linguistic expert's analysis of the obtained keywords and their relation with the topics corresponding to the textual clusters that they characterize.

Keywords: Feature Selection, Decision Trees, Text Mining, Relevance Term.

1 Introduction

A field where feature selection has a significant practical interest is the mining of information, especially, text mining, where the volume of features considered to describe the documents is extremely big and in many cases irrelevant and redundant. Several areas within text mining require a process of feature selection and many techniques of feature selection have been applied to these areas of textual processing; nevertheless, in the majority of cases the efforts in feature selection have been focused in the stage of reduction of dimensionality in textual representation. Nevertheless, there exist many other stages of the textual processing where feature selection becomes necessary. The aim of this work is to apply the feature selection techniques to a textual corpus previously classified to obtain the relevant features that are capable of characterizing the textual clusters and simultaneously manage to discern among clusters.

This paper is organized as follows. In section 2 there will appear a classification of the feature selection techniques and the principal algorithms applied in text mining will be mentioned briefly. The model proposed for the selection of relevant terms in homogeneous clusters of documents is presented in section 3. The evaluation of the

Advances in Computer Science and Engineering
Research in Computing Science 27, 2007, pp. 137-148

Received 23/02/07
Accepted 08/04/07
Final version 23/04/07

model appears in section 4. Finally, in section 5 the principal conclusions are outlined and possible applications are proposed.

2 Feature Selection in Text Mining

Feature selection that is done in the textual representation uses a filter approach. Thus, the final vocabulary is established by selecting all those features whose score is higher or lower than a predetermined threshold or selecting the best m features. To apply feature selection techniques in textual domains it is necessary preprocess the documents and represent them structurally. One of the representations most widely used in textual domains is the Vector Space Model (VSM) [10].

We will mention only some filter methods to select features used in dimensionality reduction in the textual representation task. A well-known linguistic approach is the stop word elimination [9][12]. Several numerical measurements are frequently used to evaluate the quality of the terms; e.g. eliminate all the terms whose frequencies are either higher or lower than a predefined threshold [9], consider the importance of the terms (term frequency / inverse document frequency (tfidf)) [11], consider the entropy of the probability distribution of the terms among the documents [12] and calculate measurements that are used to calculate the quality of the terms [2]. So far examples of forms of feature selection in the stage of textual representation have been given. Nevertheless, other stages exist in the textual processing that need to apply feature selection techniques, generally those that extract knowledge from texts. For example, if it is desirable to obtain an extract from every obtained cluster as result of a clustering process, it is not possible to consider all the words that were obtained in the process of dimensionality reduction of the VSM, but it becomes necessary to submit every cluster to a new dimensionality reduction process.

3 Model for the Selection of Relevant Terms in Textual Clusters

The objective of the model is to achieve the selection of keywords that characterize homogeneous clusters of related documents and simultaneously manage to discern between the clusters. The input to the model is the result of the documents clustering, where the classes to which every document corresponds are the result of the clustering process and the principal output are the keywords that characterize and discern between the homogeneous clusters of documents. Two secondary outputs, but also of big profit are: the decision tree and the rules of induction.
The relevant terms obtained can be used in later processes such as the extraction of summaries of the multiple documents that compose a homogeneous cluster and in the labeling of the clusters.

As part of the conceptual model a general procedure is developed that includes several specific procedures, structured in three stages that as a whole summarize the content of the model. The stages of the general procedure are (see Fig. 1): (1) discretization of the features that describe the documents, (2) induction of the decision tree, and (3) extraction of keywords from homogeneous textual clusters.

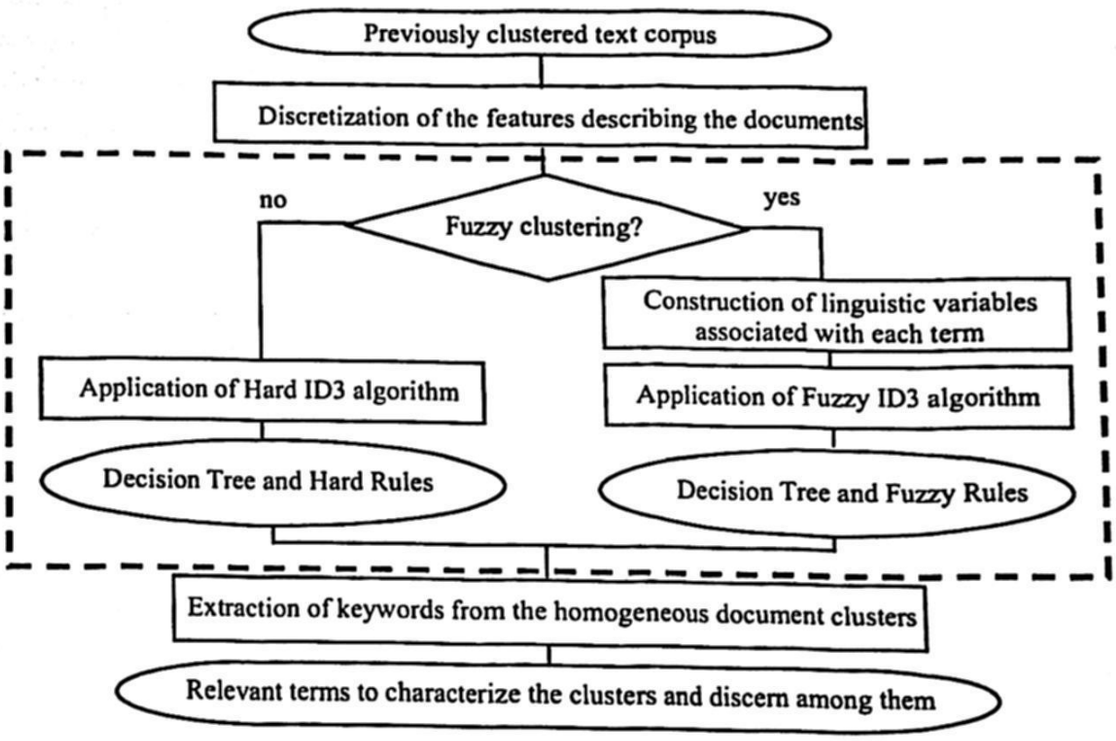

Fig. 1. Model proposed for the selection of relevant terms that characterize textual related clusters and can discern among them.

3.1 Input to the General Procedure

The input to the model is the result of the document clustering, where the classes to which each document corresponds are the resulting clusters of this process. They are considered to be outputs of clustering methods that include any of the three following techniques: hard and deterministic, fuzzy, or hard and overlap [4]. This general procedure is incorporated into the CorpusMiner system [1], that starts from a VSM representation of the document collection, whether modified or not by the application of some technique of normalization, weighting of the matrix elements, reduction of dimensionality or any combination of these, and it clusters the documents following some of these algorithms: Simultaneous Keyword Identification and Clustering of Text Documents (SKWIC) [2], Simultaneous Keyword Identification and Fuzzy Clustering of Text Documents (Fuzzy SKWIC) [2], and Extended Star algorithm [3], or the concatenated variants Extended Star – SKWIC and Extended Star – Fuzzy SKWIC [1].

We represent both input possibilities in Table 1. Clusters for hard ID3 specifies for each document to what cluster it belongs whereas cluster for fuzzy ID3 shows for each document the membership degree to each cluster obtained.

Table 1. Matrix of input to the hard and fuzzy ID3 algorithms.

	T_1	T_2	...	T_m	... for ID3	... for Fuzzy ID3
D_1	$tf_{d_1}(t_1)$	$tf_{d_1}(t_2)$		$tf_{d_1}(t_m)$	C_{k1}	$(\delta_{Cluster1}(D_1), \ldots, \delta_{Clusterk}(D_1))$
D_2	$tf_{d_2}(t_1)$	$tf_{d_2}(t_2)$		$tf_{d_2}(t_m)$	C_{k2}	$(\delta_{Cluster1}(D_2), \ldots, \delta_{Clusterk}(D_2))$
...			...		...	...
D_n	$tf_{d_n}(t_1)$	$tf_{d_n}(t_2)$		$tf_{d_n}(t_m)$	C_{kn}	$(\delta_{Cluster1}(D_n), \ldots, \delta_{Clusterk}(D_n))$

Where D_i is the i-th document of the corpus, with $i \in \{1, n\}$ and T_j is the j-th term that describes the documents, with $j \in \{1, m\}$; the values tf are the normalized and weighted frequency of each term for each document.

3.2 Stage 1: Discretization of the Features that Describe the Documents

The general procedure proposed considers in the third stage the application of the ID3 algorithm, whether in its hard or fuzzy variant. In the first case it is required that the features describing the problem are already discretized, in the second one it becomes necessary to construct the linguistic variables associated with the terms that describe the documents. Thus, stage 1 is so important, whether the processing is begun from the result of a hard or a fuzzy clustering. The proposed procedure considers in this first stage the discretization using the equal width method, that is to say, it partitions the set of possible values of the term in intervals of equal size [5].

3.3 Stage 2: Induction of the Decision Tree

The classic algorithm in the induction of decision trees is the algorithm ID3 for discreet values [6][7] and its extension C4.5 for continuous values [8]. To obtain the induction of the decision tree from the collection of documents (documents that originally are not labeled), it is necessary to consider the result of a clustering process as the set of classes of the document collection. The classification of every document consists of the clusters to which it belongs after the clustering process, or, in the fuzzy variant, its membership degree to each cluster obtained. Thus, the methods ID3 [7], C5.4 [8] and fuzzy ID3 [13] are those which will be used in the selection of the terms that characterize the homogeneous clusters of related documents. The first one is used if the input to the general procedure is the result of an algorithm that applies a hard and deterministic technique and the second one will be used if the technique is fuzzy. When the technique is fuzzy it becomes necessary to obtain the linguistic variables associated with every term that describes the collection.

The fuzzy ID3 algorithm [13] offers the possibility of considering weights associated with the objects of the decision system. In this paper two ways are considered of obtaining information to weight the documents: the highest membership degree of the document to a cluster and the variance of the membership degrees of the document to all the clusters. The highest membership degree of every document to the clusters is a way that is proposed in this paper to weight the documents, because the

higher the membership degree, the more typical or representative the document is of this cluster, therefore it must have a higher responsibility in the induction of the fuzzy decision tree. Those documents that have a similar membership degree for all clusters are documents that are not typical or typical of no cluster in particular, therefore, these documents must have a lesser influence in the selection of the algorithms to induce the fuzzy tree. Hence the documents that have a higher variance of the membership degrees to the clusters must have a major leading role in the induction. Thus, one of the criteria proposed in this paper to weight the documents is to consider the variance of the membership degrees of each document to all the clusters.

In the hard decision trees, it is simple to select which examples belong to the node that is being constructed associated with a value of a certain attribute. On the other hand, fuzzy logic states that all the elements belong to all the sets but with a given membership degree. Therefore, in the FDT after an attribute branches for a certain linguistic term, all the examples have a membership degree to this linguistic term. Let's suppose that there is a node corresponding to the linguistic variable j and that this variable is composed of k linguistic terms. Let's also suppose that the new node to form is that corresponding to the linguistic term a_t^j (linguistic term t of the linguistic variable j). What examples to consider in the node to branch associated to a_t^j ? To solve this problem in this paper two variants are proposed. The first is to apply the Principle of Maximum Membership, so that those examples which have the highest membership degree to that linguistic term t of the linguistic variable j will be included in the corresponding node of a_t^j. The second variant considers the specification of a threshold α. To include an example in the corresponding node of a_t^j, we apply an α- cut and select those examples for which it is fulfilled that its membership degree to the linguistic term t is higher than the α- cut.

Another important aspect to define is when to finish the ramification of the tree. In this stage three stopping criteria have been included. The most general is to stop the ramification when there are no any more attributes or features for the classifications. The second included criterion is to consider a node leaf when all the examples belong to the same class. The third and last criterion considered in the induction compares the value of the measurement of the information of the attribute with a stated threshold, if the studied value is minor that a threshold given by the user or calculated by the proper algorithm, stops the ramification. This stopping criterion avoids choosing attributes with very low information gain. The automated variant calculates the threshold as the average of the information gain of the attributes that initially describe the set of examples.

It is necessary to specify how it is identified to what class or set of classes an example belongs, to determine the classes associated with a node leaf. In the hard variant of the ID3 it is trivial. Nevertheless, in the fuzzy variant of the ID3 all the examples have a membership degree to each of the classes. We propose the use of the following two ways of choosing the classes corresponding to an example or node leaf:

- To apply the Principle of Maximum Membership, in such a way that the only class associated with the example will be that class for which the example has the highest membership degree. To identify the class associated with a node leaf this

principle is applied to every example of the node and the class that has the highest number of associated examples is selected.
- To define a threshold α and to apply an α- cut, thus all those classes are included in the classification of an example for which the example belonged with a degree above the threshold α. To identify the classes associated with a node leaf, an α- cut is applied to every example of the node and the classes associated with this node will be all those obtained from the examples that belong to the node according to this criterion. This is one of the advantages of the FDT, because this method allows a node leaf to have more than one value of the decision attribute.

Another element to bear in mind to obtain a node leaf is the definition of its certainty, an important aspect when generating and applying the rules from the FDT. We have considered two ways of calculating the certainty of a node leaf (certainty of the rule that generates from the root up to the given node leaf). (i) To calculate the certainty of the node leaf as the average of the membership degrees of the examples those are in the node to the classes selected for this node. (ii) To consider the weighted sum of the membership degrees of the existing examples in the node leaf to the classes selected for this node. The weighting is based on the weight associated with every example. In both variants a decision tree is generated and the rules that describe every document cluster are obtained from the generated tree. The antecedents of these rules are the resultant intervals of the discretization process (hard variant) or the terms associated to the linguistic variables (fuzzy variant). The extraction of keywords is performed based on the analysis of the obtained rules. This process will be described in stage 3.

Fuzzy ID3 variant needs the construction of linguistic variables associated with every term. Several researches have been carried out with the purpose of automatically building membership functions. We used the method propose in [12] for the construction of triangular and the Beta bell functions.

Generation of Rules that Describe a Textual Corpus from the Hard and Fuzzy Variants of the Algorithm ID3. After the tree has been constructed the rules that describe the text corpus can be generated, bearing in mind that every path in the decision tree of the root to the leaves is a rule, where the precedent is a conjunction of all the internal nodes of the tree that belong to the path (with its respective discrete associate values or linguistic terms for ID3, hard or fuzzy variant, respectively) and the consequent is the node leaf (i.e., associate classes and certainty of the rule). The rules that are obtained from the tree induced for fuzzy ID3 are Sugeno grade 0.

3.4 Stage 3: Extraction of Keywords of Textual Homogeneous Clusters

In this stage of the model we propose three variants of feature selection to extract the keywords of the homogeneous clusters of related documents obtained from the clustering results. This may be useful, for example, for a possible later stage of automatic generation of the summary extract of every cluster or labeling of textual clusters. Thus we may identify those terms that characterize every cluster, through the selection of the words of higher relevance of clustering methods, the selection of the

terms with higher values of quality in the cluster, and the selection of the terms from the rules generated by the algorithm ID3 in any of its variants.

The selection of the words of higher relevance resulting from the clustering method, and the selection of the terms with higher quality values in the cluster, are forms of selection that coincide when the clustering was performed applying either a hard or a fuzzy technique. The only difference is that when the technique is fuzzy it becomes necessary to apply the principle of maximum membership or to define one α- cut to determine what documents belong to each cluster. Nevertheless, the election of the terms from the rules generated by the algorithm ID3 depends on whether the variant was hard or fuzzy.

Selection of the Words of Higher Resultant Relevancy of the Clustering Methods. There exist clustering algorithms that along with the collection of document clusters return the relevance of the terms for each cluster; such is the case of the algorithms SKWIC and Fuzzy SKWIC. Considering the relevance of the terms for clusters it is possible to select the relevant terms for every obtained cluster in two ways: selection of the words whose relevance is higher than a certain threshold and selection of n words with better relevance value. The first one goes through all the terms in the cluster and chooses those whose relevance value is higher than a stated threshold. The second one sorts all the terms in decreasing order according to their relevance and selects the first n terms of the list. Notice that this way of keyword selection is only applicable to results of the clustering with SKWIC and Fuzzy SKWIC.

Extraction of Keywords as the Quality of Terms. As mentioned in section 2 of this paper, there exist functions that determine the quality of a term in a document collection. Using these quality measurements it is possible to reduce the dimensionality of the VSM representation of a text corpus by eliminating the words of lesser quality value. It is possible to extrapolate this form of selection of words to each cluster of documents obtained by a method of clustering. To achieve this, it is necessary to create a VSM representation for each cluster of documents. This representation has the same terms as the matrix from which the collection of clusters of documents was obtained, but for each cluster the representation has only those documents belonging to it. Then, for each of these representations the quality of all the terms is calculated and those are chosen which have higher or lower value than a threshold (lower in the case of entropy) or the n better quality terms.

Selection of the Terms from the Rules Generated by the Algorithm ID3. From the obtained rules it is possible to generate the words that discern among clusters. This process of selection differs according to whether the rules are hard or fuzzy.

The selection of terms from hard rules is carried out as follows. With every cluster there are associated those terms that are a part of the antecedents of the rules of which that they are consequents and whose value (i.e., interval associated in the process of discretization) is one of the n better values that this term can reach, where n is a value of input to the algorithm. Let's suppose, for example, that one of the terms that describes a textual corpus is the word SOFTWARE, and in the process of

discretization the frequencies of appearance in the corpus for this term were divided into as VERY LOW, LOW, NORMAL, HIGH and VERY HIGH, and it has been specified except that only it is desirable to consider in a cluster those terms that should describe it with a frequency of appearance HIGH or VERY HIGH, then only SOFTWARE will be considered to be a relevant term in those clusters where this term is a part of the antecedents of the rules and that these rules have the above mentioned cluster as a consequent.

The selection of terms from fuzzy rules is carried out as follows. With each cluster are associated those terms that are a part of the antecedents of the rules of which they are consequents and that its value (i.e., linguistic term associated with the linguistic variable corresponding to the attribute that describes the node) is one of n better values that this term can have, where n is a value of input to the algorithm. Therefore, this variant is similar to the processing of hard rules, with the only difference that is processed through linguistic terms and not through discrete intervals.

Combined Forms of Selection. Notice that the ways of selection showed above are used to extract relevant terms for clusters, but these do not necessarily manage to discern among clusters. On the other hand, the use of the rules obtained by the induced decision trees can generate terms that do discern among clusters, but that do not necessarily have a high frequency of appearance in the cluster, and therefore, are terms that can hardly be used in later processing as extraction of summaries or labeling of clusters. It is for these reasons, that combined variants offer the best solutions. If the algorithm that generated the collection of clusters of documents also generated the relevance of every term for cluster, the lists of keywords that are obtained by the ID3 (hard or fuzzy) can be intercepted with the lists of words that are obtained when terms are chosen based on their relevance above a certain threshold. Another possible combination is to intercept the results of the ID3 with the high quality words obtained for each cluster.

4 Evaluation of the Model

Evaluating is an arduous work in tasks of text mining. To evaluate the model study cases were designed and experts' opinion was considered to perform the semantic analysis of the words extracted in the context of the textual corpus used.

The general procedure was implemented in the software CorpusMiner [1]. This software also implements the initial processing and clustering of the textual corpus. In the stage of textual representation, the corpus was transformed by conversion of all the letters to capital letters, replacement of the contractions with their expansions and of abbreviations with their full forms, the elimination of numbers and symbols, the establishment of orthographic homogeneity and lemmatization. Then the VSM representation was carried out on the transformed corpus, with a weighting based on a variation of the formula TF-IDF [2], allowing that the weight of the terms should reflect the relative importance of a term in a document with regard to other terms in the document. The reduction of the dimensionality was performed by the elimination of the grammatical words and the selection of those 600 better terms, that is to say,

terms that have a higher quality than a certain threshold for the applied measurement of term quality (Term Quality II) [1][2]. The concatenated methods applied for the clustering were Extend Star – SKWIC and Extend Star – Fuzzy SKWIC [1][2][3].

4.1 Definition of the Study Cases for the Application of the General Procedure

The first study case included a textual corpus that was created from the Reuters Agency news collection published by David D. Lewis. The second study case is an artificial corpus created by expert linguists for this validation. Finally, the third study case is a collection of Bioinformatics' scientific papers published in the BioMed Central's open access full-text corpus for data mining research.

The first study case includes the textual corpus from Reuters Agency news collection[1]. The created corpus has a size of 353 KB. It possesses 113 pieces of news, previously labeled. These documents tackle 6 topics; 12 news about cocoa, 23 news about acq, 12 news about money-supply, 17 news about trade, 24 news about crude, and 25 news about earn.

The second study considers a textual corpus that was constructed intentionally by expert linguists to validate clustering for meaning association. The construction of this corpus starts from a collection of documents, from which the lexical high frequency words were selected and the sentences containing them were assigned to a document. Thus for each word selected there is a pseudo-text that contains a variable number of sentences containing that word. The built corpus is composed by 35 documents (corresponding to the 35 most frequent words of the original corpus) and it occupies 2.78 MB.

The third study case includes the textual Corpus from BioMed Central's open access full-text corpus[2]. The created corpus has a size of 3.08 MB. It possesses 123 scientific papers, previously labeled. These documents tackle 7 topics; 16 papers about Cystic fibrosis, 12 papers about genic therapy, 6 papers about diabetes mellitus (therapy and diet), 32 papers about diabetes mellitus (research, molecular biology), 31 papers about AID, 16 papers about lung cancer, and 10 papers about microarrays.

4.2 Validation of the Results

To validate the results of the first and second study cases we considered the opinion of experts in English to determine the appropriateness of the keywords selected for the clusters. In their opinion the words obtained manage to describe the clusters and to discern among them.

We apply a fuzzy clustering technique for these study cases; thus, we had to apply the induction of the fuzzy decision trees using fuzzy ID3 algorithm. Tables 2 and 3 reflect fragments of the results of the keyword selection process from the clusters for the first and second study cases. The terms shown are the result of the selection

---

[1] Reuters-21578 Text Categorization Collection, 135 topics.
http://www.daviddlewis.com/resources/testcollections/reuters21578

[2] BioMed Central has so far published 22003 articles.
http://www.biomedcentral.com/info/about/datamining/

method that considers the interception of the terms obtained by fuzzy ID3 and the relevance of the terms calculated by the Extended Star – Fuzzy SKWIC (i.e, we used a fuzzy clustering technique). In the discretization process for the automatic construction of the membership function three frequency intervals were considered, LOW, NORMAL and HIGH. In the selection we only considered those terms that were part of the antecedents associated to linguistic terms with an HIGH frequency and a relevance based on the results of the Extended Star – Fuzzy SKWIC algorithm above a given threshold (which was different for the two study cases). In the induction of the fuzzy decision tree each document was weighted with the variance of the membership degrees of the document to each of the clusters. We applied the principle of maximum membership to determine what examples to include in the branching of the node. The branching of each node was pruned when the attribute gain was smaller than the means of the gains of the attributes that initially describe the set of all examples. The certainty of the rules was calculated as the weighted sum of the examples' membership degrees to the classes selected for this node leaf (classes whose examples' membership degree is above a given α-cut).

Table 2. An excerpt of keywords obtained using Fuzzy ID3 with the first study case.

Cluster 4		Cluster 6		Cluster 5		Cluster 1	
<TRADE> (0.76)		<EARN> (0.72)		<CRUDE> (0.86)		<COCOA> (0.91)	
System	important	dollar	Exchange	military	risk	Cocoa	bean
Foreign	export	represent	Pay	platform	involve	Strong	cost
Tender	trade	cause	Earn	Washington	predict	Carry	rise
Sight	spokesman	financial	Asset	target		Land	dollar
Gulf	congress	partly	intangible	Iran		Pound	condition
		asset				Offer	good

The headings for each cluster in Tables 2 and 3 are the topics most widespread in the cluster and their degrees of importance. Observe in Table 3 that the chosen terms have a relationship with the main topics dealt with in these homogeneous document clusters, so the semantics of the selection is correct and adequately describes the clusters under study. These words also discern among the clusters.

See in Table 3 the results of the keywords obtained for a fragment of the clusters of the collection of the second study case. Notice that the chosen terms are in correspondence with the content of the clusters.

Table 3. An excerpt of keywords obtained using Fuzzy ID3 with the second study case.

Cluster 1	Cluster 2	Cluster 4	Cluster 5	Cluster 6
<virtual> (0.89)	<wireless> (0.77) <users> (0.74)	<virus> (0.84)	<woman > (0.83)	<warming> (0.90) <weather> (0.83)
virtual	wireless	virus	woman	weather
system	user	system	female	warm
future	computer	seriously	men	troposphere

To validate the results of the third study case, we considered the opinion of experts in Bioinformatics to determine the appropriateness of the keywords selected for the clusters of this kind of scientific papers. In their opinion the words obtained manage to describe the clusters and to discern among them.

We decided to use a hard and deterministic clustering technique; thus we needed to apply Hard ID3 to induce the hard decision tree. Table 4 reflects the complete results of the keyword selection process from the homogeneous clusters of Bioinformatics' papers. The keywords shown are the result of the selection method that considers the interception of the terms obtained by ID3 and the relevance of the terms calculated by the Extended Star – SKWIC (i.e, we used a hard and deterministic clustering technique). In the discretization process three frequency intervals were considered, LOW, NORMAL and HIGH. In the selection we only considered those terms that were part of the antecedents associated to intervals with a HIGH frequency and a relevance based on the results of the Extended Star –SKWIC algorithm above a given threshold. The threshold is given considering the characteristics of this study case.

Table 4. The keywords obtained using ID3 with the third study case.

Cystic Fibrosis	Genetic Therapy	Diabetes (Diet and therapy)	Diabetes (Research)	AID	Lung Cancer	Microarray
cystic	Virus	people	experiment	HIV	response	clone
fibrosis	Transfer	prevention	cell	program	tumor	microarray
surface	DNA	primary	clone	human	Lung	experiment
line	Joint	model	pain	prevention	clone	
	AAV	status	mouse	transmission	DNA	
	Tumor		human	information		
				health		

See in Table 4 that the chosen terms have a relationship with the main topics dealt with in these homogeneous document clusters, so the semantics of the selection is correct and adequately describes the clusters under study. These words also discern among the clusters.

5 Conclusions and Future Work

A model has been presented that allows the selection of keywords that characterize homogeneous clusters of documents and can simultaneously discern among the clusters. The characteristics present in the general procedure of the developed model provide advantages with regard to the consideration in the input of new ways of document clustering and the inclusion of other variants construction of membership functions associated to the linguistic variables for every feature. The stage of keyword extraction allows the combination of the relevancy of the features obtained by the clustering processes, with the words selected from the process of induction of the hard or fuzzy decision trees. This element is fundamental to obtain relevancy for clusters and differentiation among them.

As a future work, it is possible use the relevant terms obtained in the extract summarization of textual homogeneous clusters, as well as in the labeling clustering process.

Acknowledgments. This work was supported in part by VLIR (Vlaamse Inter Universitaire Raad, Flemish Interuniversity Council, Belgium) under the IUC Program VLIR-UCLV and by Informatics for Enterprises Project between Cuba and Germany. Thanks also to Prof. Rudolf Kruse, Prof. Christian Borgelt and Prof. Andreas Nuernberger for their suggestions.

References

1. Arco, L., Bello, R., Mederos, J.M., Pérez, Y.: Agrupamiento de documentos textuales mediante métodos concatenados. Revista Iberoamericana de Inteligencia Artificial, 10(30) (2006) 43-53
2. Berry, M.: Survey of text mining. Clustering, classification, and retrieval. Springer-Verlag. (2004)
3. Gil-García, R., Badía-Contelles, J.M., Pons-Porrata, A.: Extended star clustering algorithm. Proceedings of CIARP. Lecture Notes in Computer Science, 2905, Springer-Verlag (2003) 480-487
4. Höppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy cluster analysis. Methods for classification. Data Analysis and Image Recognition. John Wiley & Sons Ltd. (1999)
5. Liu, H., Setiono, R.: Feature selection via discretization. IEEE Transactions on Knowledge and Data Engineering, 9(4) (1997) 642-645
6. Mitchell, T.: Machine learning. McGraw-Hill Science (1997)
7. Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann Series in Machine Learning (1993)
8. Quinlan, J.R. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4 (1996) 77–90
9. Rijsbergen, C.J.: Information Retrieval. London, Butterworths (1979)
10. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic text retrieval. Communications of the ACM, 18(11) (1975) 613-620
11. Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. Information Processing and Management 24(5) (1988) 513-523
12. Salton, G., McGill, M.: Introduction to modern information retrieval. New York: McGraw-Hill (1983)
13. Wang, X., Borgelt, C.: Information measures in fuzzy decision trees. IEEE International Conference on Fuzzy Systems (2003)