TERM WEIGHTING: NOVEL FUZZY LOGIC BASED METHOD VS. CLASSICAL TF-IDF METHOD FOR WEB INFORMATION EXTRACTION

Jorge Ropero, Ariel Gómez, Carlos León, Alejandro Carrasco
Department of Electronic Technology,University of Seville, Seville, Spain

Keywords: Term weighting, TF-IDF, Fuzzy logic, Information extraction, Information retrieval, Vector space model, Intelligent agent.

Abstract: Solving Term Weighting problem is one of the most important tasks for Information Retrieval and Information Extraction. Tipically, the TF-IDF method have been widely used for determining the weight of a term. In this paper, we propose a novel alternative fuzzy logic based method. The main advantage for the proposed method is the obtention of better results, especially in terms of extracting not only the most suitable information but also related information. This method will be used for the design of a Web Intelligent Agent which will soon start to work for the University of Seville web page.

1 INTRODUCTION

The great amount of available information caused by the rising of Information Technology constitutes an enormous advantage when it comes to search for needed information. However, at the same time, it is a great problem to distinguish the necessary information among the huge quantity of unneccessary data.

For this reason, the concepts of Information Retrieval (IR) and Information Extraction (IE) came up. IR is a field in which there have been great advances in the last decades (Kwok, 1989), especially in what concerns to the search of documents. Nevertheless, IR does not only come down to document searching. IR tools may be used for the objects in any collection of accumulated knowledge such as the objects stored in a shop or the photographies in an album. The generalization of this method is possible thanks to the substitution of every object for its representation in Natural Language (NL). IE involves a transformation of a collection of documents, generally helped by an IR system. This collection of documents is transformed into easier to assimilate and analyze information. IE tries to extract relevant facts from documents, whereas IR selects relevant documents. Therefore, it might be said that IE works with a higher level of granularity than IR. (Kosala, 2002). In our case, we are applying IE techniques to a web portal. A web portal consists of a collection of web pages, so the method is completely applicable.

IR has been widely used for text classification (Aronson et al., 1994; Liu et al., 2001) introducing approaches such as Vector Space Model (VSM), K nearest neighbour method (KNN), Bayesian classification model, neural networks and Support Vector Machine (SVM) (Lu et al., 2002). VSM is the most frequently used model. In VSM, a document is conceptually represented by a vector of keywords extracted from the document, with associated weights representing the importance of these keywords in the document. Typically, the so-called TF-IDF method is used for determining the weight of a term (Lee et al., 1997). Term Frequency (TF) is the frequency of occurrence of a term in a document and Inverse Document Frequency (IDF) varies inversely with the number of documents to which the term is assigned (Salton & Buckley, 1988).

Although TF-IDF method for Term Weighting (TW) has worked reasonably well for IR and has been a starting point for more recent algorithms, (Lee et al., 1997; Salton & Buckley, 1988; Liu et al., 2001; Zhao & Karypis, 2002; Lertnattee & Theeramunkong, 2002; Xu et al., 2003), it was never taken into account that some other aspects of keywords may be important for determining term

Ropero J., Gómez A., León C. and Carrasco A. (2009).
TERM WEIGHTING: NOVEL FUZZY LOGIC BASED METHOD VS. CLASSICAL TF-IDF METHOD FOR WEB INFORMATION EXTRACTION.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 130-137
DOI: 10.5220/0001982901300137

weights apart from TF and IDF: first of all, we should consider the degree of identification of an object if only the considered keyword is used. This parameter has a strong influence on the final value of a term weight if the degree of identification is high. The more a keyword identifies an object, the higher value for the corresponding term weight; secondly, we should also consider the existance of join terms.

In this paper, we introduce a fuzzy logic (FL) based term weighting scheme. This scheme bears in mind all these features for calculating the weight of a term, taking advantage of fuzzy logic flexibility. Fuzzy logic makes it possible to have a non-rigid term weighting.

2 METHODOLOGY FOR IE

As said above, we are applying IE to a web portal. Particularly, we have worked with the University of Seville web portal. To carry out Information Extraction, it is necessary to identify web page and object, that is to say, every web page in a portal is considered an object. These objects are gathered in a hierarchical structure. An object is classified under a unique criterion - or group of criteria -.

In our case, we have taken advantage of the hierarchical structure of a web page to divide the portal in three levels: Topic, Section and Object. The size of every level is variable. Every object is represented by means of a set of questions formulated in NL. We have called these questions standard questions. The number of standard questions associated to every web page is variable, depending on the amount of information contained in every page, its importance and the number of synonymous of index terms. Logically, System Administrator's knowledge about the jargon of the related field is pretty important. The higher his knowledge, the higher reliability of the proposed standard questions, as they shall be more similar to possible user consultations. After all, users are the ones who try to extract the information. Our study was based both on the study of the web pages themselves and on previous consultations - University of Seville bank of questions - .

Once standard questions are defined, index terms are extracted from them. We have defined these index terms as words, though they also may be compound terms. Index terms are the ones that better represent a standard question. Every index term is associated with its correspondent term weight. This weight has a value between 0 and 1 and depends on the importance of the term in every hierarchic level. The higher importance in a level, the higher is the term weight. In addition, term weight is not constant for every level, as the importance of a word to distinguish a topic from the others may be very different from its importance to distinguish between two objects.

An example of the followed methodology is shown in Table 1.

Table 1: Example of the followed methodology.

STEP	EXAMPLE
Step 1: Web page identified by standard/s question/s	- Web page: www.us.es/univirtual/internet - Standard question : Which services can I access as a virtual user at the University of Seville?
Step 2: Locate standard/s question/s in the hierarchical structure.	Topic 12: Virtual University Section 6: Virtual User Object 2.
Step 3: Extract index terms	Index terms: 'services', 'virtual', 'user'
Step 4: Term weighting	See section 4

When a user consultation is made, these term weights are the inputs to a fuzzy logic system, which must detect the object to which the correspondent user consultation refers. System operation is described in (Ropero et al., 2007).

3 TERM WEIGHTING

As said in previous sections of this paper, there are a few weights associated with every index term. The values of the weights must be related somehow to the importance of an index term in its corresponding set of knowledge - in our case, Topic, Section or Object -. We may consider two options to define these weights:

An expert in the matter should evaluate intuitively the importance of the index terms. This method is simple, but it has the disadvantage of depending exclusively on the knowledge engineer. It is very subjective and it is not possible to automate the method.

The generation of automated weights by means of a set of rules. The most widely used method for TW is the TF-IDF method, but we propose a novel Fuzzy Logic based method, which achieves better results in IE.

3.1 The TF-IDF Method

The idea of automatic text retrieval systems based on the identification of text content and associated identifiers is dated in the 50s, but it was Gerard Salton in the late 70s and the 80s who laid the foundations of the existing relation between these identifiers and the texts they represent (Salton & Buckley, 1988). Salton suggested that every document D could be represented by term vectors tk and a set of weights wdk, which represent the weight of the term tk in document D, that is to say, its importance in the document.

A TW system should improve efficiency in terms of two main factors, recall and precision. Recall bears in mind the fact that the most relevant objects for the user must be retrieved. Precision takes into account that strange objects must be rejected. (Ruiz & Srinivasan, 1998). Recall may be defined as the number of retrieved relevant objects divided by the total number of objects. On the other hand, precision is the number of retrieved relevant objects divided by the total number of retrieved objects. Recall improves if high-frequency terms are used, as such terms will make it possible to retrieve many objects, including the relevant ones. Precision improves if low-frequency terms are used, as specific terms will isolate the relevant objects from the non-relevant ones. In practice, compromise solutions are used, using terms which are frequent enough to reach a reasonable level of recall without producing a too low precision.

Therefore, terms that are mentioned often in individual objects, seem to be useful to improve recall. This suggests the utilization of a factor named Term Frequency (TF). Term Frequency (TF) is the frequency of occurrence of a term. On the other side, another factor should favor the terms concentrated in a few documents of the collection. The inverse frequency of document (IDF) varies inversely with the number of objects (n) to which the term is assigned in an N-object collection. A typical IDF factor is log (N/n). (Salton & Buckley, 1988). A usual formula to describe the weight of a term j in document i is:

$$w_{ij} = tf_{ij} \times idf_j . \quad (1)$$

This formula has been modified and improved by many authors to achieve better results in IR and IE (Lee et al., 1997; Liu et al., 2001; Zhao & Karypis, 2002; Lertnattee & Theeramunkong, 2002; Xu et al., 2003).

3.2 The FL based Method

The TF-IDF method works reasonably well, but it has the disadvantage of not considering two key aspects for us:

The degree of identification of the object if only the considered index term is used. This parameter has a strong influence on the final value of a term weight if the degree of identification is high. The more a keyword identifies an object, the higher value for the corresponding term weight. Nevertheless, this parameter creates two disadvantages in terms of practical aspects when it comes to carrying out a term weight automated and systematic assignment. On the one hand, the degree of identification is not deductible from any characteristic of a keyword, so it must be specified by the System Administrator. On the second hand, the same keyword may have a different relationship with every object.

The second parameter is related to join terms. In the index term 'term weighting', this expression would constitute a join term. Every single term in a join term has a lower value than it would have if it did not belong to it. However, if we combine all the single terms in a join term, term weight must be higher. A join term may really determine an object whereas the appearance of only one of its single terms may refer to another object.

The consideration of these two parameters together with classical TF and IDF determines the weight of an index term for every subset in every level. The FL based method gives a solution to all the problems and also gives two main advantages. The solution to both problems is to create a table with all the keywords and their corresponding weights for every object. This table will be created in the phase of keyword extraction from standard questions. Imprecision practically does not affect the working method due to the fact that both term weighting and information extraction are based on fuzzy logic, what minimizes possible variations of the assigned weights. The way of extracting information also helps to successfully overcome this imprecision. In addition, the FL based method also gives important advantages: on the one hand, term weighting is automated; on the other hand, the level of required expertise for an operator is lower. This operator would not need to know anything about the FL engine functioning, but only how many times does a term appear in any subset and the answer to these questions: a) Does a keyword undoubtedly

define an object by itself? b) Is a keyword tied to another one?

In our case, the application of this method to a web portal, the web portal developer himself may define simultaneously the standard questions and index terms associated with the object - a web page - and the response to the questions mentioned above.

4 METHOD IMPLEMENTATION

This section shows how the TF-IDF method and the FL based method were implemented in practise, in order to compare both methods applying them to the University of Seville web portal.

4.1 TF-IDF Method Implementation

As mentioned in previous sections, a reasonable measure of the importance of a term may be obtained by means of the TF-IDF product. However, this formula has been modified and improved by many authors to achieve better results in IR and IE. Eventually, the chosen formula for our tests was the one proposed by Liu et al. (Liu et al., 2001).

$$W_{ik} = \frac{tf_{ik} \times \log(N / n_k + 0.01)}{\sqrt{\sum_{k=1}^{m} tf_{ik} \times \log(N / n_k + 0.01))^2}} \quad (2)$$

Where tfik is the ith term frequency of occurrence in the kth subset - Topic / Section / Object -. nk is the number of subsets to which the termTk is assigned in a collection of N objects. Consequently, it is taken into account that a term might be present in other sets of the collection.

As an example, we are using the term ‘virtual’, above used in the example in Section 2.

At Topic level:

- ‘Virtual’ appears 8 times in Topic 12 (tf_{ik} = 8, K=12).
- ‘Virtual’ appears twice in other Topics (n_k = 3)
- There are 12 Topics in total (N=12) - for normalizing, it is only necessary to know the other tf_{ik} and n_k for the Topic -.
- Substituting, W_{ik} = 0.20.

At Section level:

- ‘Virtual’ appears 3 times in Section 12.6 (tf_{ik} = 3, K= 6)
- ‘Virtual’ appears 5 times in other Sections in Topic 12 (n_k = 6)
- There are 6 Sections in Topic 12 (N=6).
- Substituting, W_{ik} = 0.17.

At Object level:

- ‘Virtual’ appears once in Object 12.6.2 (tf_{ik} = 1, K = 2). – Logically a term can only appear once in an Object -.
- ‘Virtual’ appears twice in other Topics (n_k = 3)
- There are 3 Objects in Section 12.6 (N=3).
- Substituting, W_{ik} = 0.01. In fact, ‘virtual’ appears in all the Objects in Section 12.6, so it is irrelevant to distinguish the Object.

Consequently, ‘virtual’ will be relevant to find out that the Object is in Topic 12, Section 6, but irrelevant to find out the definite Object, which should be found according to other terms in a user consultation.

4.2 FL based Method Implementation

As said in section 3.2, TF-IDF has the disadvantage of not considering the degree of identification of the object if only the considered index term is used and the existance of tied keywords. Like TF-IDF method, it is neccesary to know TF and IDF, and also the answer to the questions mentioned in section 3.2. FL based Term Weighting method is defined below. Four questions must be answered to determine the Term Weight of an Index Term:

- Question 1 (Q1): How often does an index term appear in other subsets? - Related to IDF -.
- Question 2 (Q2): How often does an index term appear in its own subset? - Related to TF -.
- Question 3 (Q3): Does an index term undoubtedly define an object by itself?
- Question 4 (Q4): Is an index term tied to another one?

Question 1

Term weight is partly associated to the question ‘How often does an index term appear in other subsets?’. It is given by a value between 0 – if it appears many times – and 1 - if it does not appear in any other subset -. To define weights, we are considering the times that the most used terms in the whole set of knowledge appear.

Provided that there are 1114 index terms defined in our case, we have assumed that 1 % of these words must mark the border for the value 0 (11 words). As the eleventh most used word appears 12 times, whenever an index term appears more than 12 times in other subsets, we will give it the value of 0. Values for every Topic are defined in Table 2.

Table 2: Term weight values for every Topic for Q1.

Times appearing	0	1	2	3	4	5	6	7	8	9	10	11	12	>12
Value	1	0.9	0.8	0.7	0.64	0.59	0.53	0.47	0.41	0.36	0.3	0.2	0.1	0

Table 3: Term weight values for every Section for Q1.

Times appearing	0	1	2	3	4	5	> 5
Value	1	0.7	0.6	0.5	0.4	0.3	0

Table 4: Term weight values for every Object for Q1.

Times appearing	0	1	2	>2
Value	1	0.7	0.3	0

Table 5: Term weight values for every Topic and Section for Q2.

Times appearing	1	2	3	4	5	> 5
Value	0	0.3	0.45	0.6	0.7	1

Table 6: Term weight values for Q3.

Answer to Q3: Does a term define undoubtedly a standard question?	Yes	Rather	No
Value	1	0.5	0

Between 0 and 3 times appearing - approximately a third of the possible values - , we consider that an index term belongs to the so called HIGH set. Therefore, it is defined in its correspondant fuzzy set with uniformly distributed values between 0.7 and 1, as may be seen in Figure 1. Analogously, we may distribute all values uniformly according to different fuzzy sets. Fuzzy sets are triangular, on one hand for simplicity and on the other hand because we tested other more complex types of sets (Gauss, Pi type, etc) and the results did not improve at all.

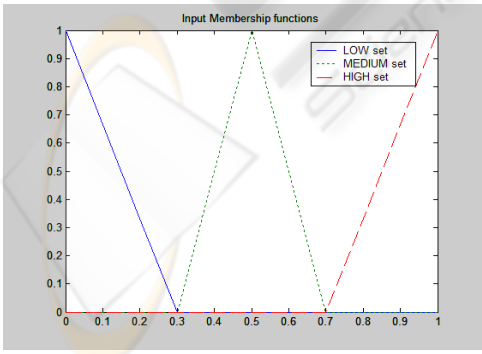

Figure 1: Input fuzzy sets.

Provided that different weights are defined in every hierarchic level, we should consider other scales to calculate them. As for the Topic Level we were considering the immediately top level – the whole set of knowledge - , for the Section level we should consider the times that an index term appears in a certain Topic. We again consider that 1 % of these words must mark the border for the value 0 - 11 words - . The eleventh most used index term in a unique Topic appears 5 times, so whenever a term appears more than 5 times in other subsets, its weight takes the value 0 at the Section of level.

Possible term weights for the level of Section are shown in Table 3. The method is analogous and considers the definition of the fuzzy sets. At the level of Object, term weights are shown in Table 4.

Question 2

To find out the term weight associated to question 2 - Q2, How often does an index term appear in its own subset? -, the reasoning is analogous. However, we have to bear in mind that it is necessary to consider the frequency inside a unique set of knowledge, thus the number of appearances of index terms decreases considerably. The list of the most used index terms in a Topic must be considered again. It also must be born in mind that the more an index term appears in a Topic or Section, the higher value for an index term. Q2 is senseless at the level of Object. The proposed values are given in Table 5.

Table 7: Term weight values for Q4.

Number of index terms tied to another index term	0	1	2	> 2
Value	1	0.7	0.3	0

Table 8: Rule definition for Topic and Section levels.

Rule number	**Rule definition**	**Output**
R1	IF Q1 = HIGH and Q2 ≠ LOW	At least MEDIUM-HIGH
R2	IF Q1 = MEDIUM and Q2 = HIGH	At least MEDIUM-HIGH
R3	IF Q1 = HIGH and Q2 = LOW	Depends on other Questions
R4	IF Q1 = HIGH and Q2 = LOW	Depends on other Questions
R5	IF Q3 = HIGH	At least MEDIUM-HIGH
R6	IF Q4 = LOW	Descends a level
R7	IF Q4 = MEDIUM	If the Output is MEDIUM-LOW, it descends to LOW
R8	IF (R1 and R2) or (R1 and R5) or (R2 and R5)	HIGH
R9	In any other case	MEDIUM-LOW

Question 3

In the case of question 3 – Q3, Does a term define undoubtedly a standard question? - , the answer is completely subjective and we propose the answers 'Yes', 'Rather' and 'No'. Term weight values for this question are shown in Table 6.

Question 4

Finally, question 4 – Q4, Is an index term tied to another one? – deals with the number of index terms tied to another one. We propose term weight values for this question in Table 7. Again, the values 0.7 and 0.3 are a consequence of considering the border between fuzzy sets – see Figure 1-.

After considering all these factors, fuzzy rules for Topic and Section levels are defined in Table 8. This rules cover all the 81 possible combinations. Note that, apart from the three input sets mentioned in previous sections, four output sets have been defined - HIGH, MEDIUM-HIGH, MEDIUM-LOW and LOW-, as may be seen in Figure 2. At the level of Object, we must discard question 2 and rules change.

The only aspect which has not been defined yet is about multiple appearances in a Topic or Section. I.e., it is possible that the answer to question 3 is 'Rather' in one case 'No' in another one. In this case, a weighted average of the corresponding term weights is calculated.

An example of all the process is shown below

Example

Object 12.6.2 is defined by the following standard question :

Which services can I access as a virtual user at the University of Seville?

If we consider the term 'virtual':

- At Topic level:

- 'Virtual' appears twice in other Topics in the whole set of knowledge, so that the value associated to Q1 is 0.80.

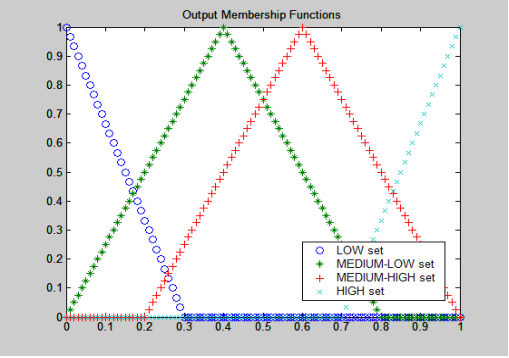

Figure 2: Output fuzzy sets.

- 'Virtual' appears 8 times in Topic 12, so that the value associated to Q2 is 1.

- The response to Q3 is 'Rather' in 5 of the 8 times and 'No' in the other three, so that the value associated to Q3 is a weighted average: (5*0.5 + 3*0)/8 = 0.375.

- Term 'virtual' is tied to one term 7 times and it is tied to two terms once. Therefore, the average is

Table 9: Test comparison between both methods.

	Cat1	Cat2	Cat3	Cat4	Cat5	Total
TF-IDF Method	466 (50.98%)	223 (24.40%)	53 (5.80%)	79 (8.64%)	93 (10.18%)	914
FL Method	710 (77.68%)	108 (11.82%)	27 (2.95%)	28 (3.06%)	41 (4.49%)	914

1.14 terms. A linear extrapolation leads to a value associated to Q4 of 0.65.

- With all the values as inputs for the fuzzy logic engine, we obtain a term weight of 0.53.

- At Section level

- 'Virtual' appears 5 times in other Sections corresponding to Topic 12, so that the value associated to Q1 is 0.30.

- 'Virtual' appears 3 times in Topic 12, so that the value associated to Q2 is 0.45.

- The response to Q3 is 'Rather' in all cases, so that the value associated to Q3 is 0.5.

- Term 'virtual' is tied to term 'user' so that the value associated to Q4 is 0.7.

- With all the values as inputs for the fuzzy logic engine, we obtain a term weight of 0.45.

- At Object level:

- 'Virtual' appears twice in other Objects corresponding to Section 12.6, so that the value associated to Q1 is 0.30.

- The response to Q3 is 'Rather', so that the value associated to Q3 is 0.5.

- Term 'virtual' is tied to term 'user' so that the value associated to Q4 is 0.7.

- With all the values as inputs for the fuzzy logic engine, we obtain a term weight of 0.52. We can see the difference with the corresponding term weight obtained with the TF-IDF method, but it is exactly what we are looking for: not only the desired object must be retrieved, but the most closely related to it.

5 TESTS AND RESULTS

Tests have been done on the University of Seville web portal. This web portal has 50,000 daily visits, what qualifies it into the 10% most visited University portals – there are more than 4,000 -. As there is much information in it, 253 objects grouped in 12 Topics were defined. All these groups were made up of a variable number of Sections and Objects. 2107 standard questions surged from these 253 Objects, but slightly more than the half of them were eliminated for these tests because of being very similar to others. Eventually, tests consisted of 914 user consultations.

To compare results, we considered the position in which the correct answer appeared among the retrieved answers, according to fuzzy engine outputs. The first necessary step to follow is to define the overcoming thresholds for the fuzzy engine. This way, Topics and Sections that are not related with the Object to identify are eliminated. We also have to define low enough thresholds, in order to be able to obtain also related Objects. We suggest to present between 1 and 5 answers, depending on the number of related Objects.

The results of the consultation were sorted in 5 categories:

- Category Cat1: the correct answer is retrieved as the only answer or it is the one that has a higher degree of certainty between the answers retrieved by the system.

- Category Cat2: The correct answer is retrieved between the 3 with higher degree of certainty -excluding the previous case -.

- Category Cat3: The correct answer is retrieved between the 5 with higher degree of certainty - excluding the previous cases -.

- Category Cat4: The correct answer is retrieved, but not between the 5 with higher degree of certainty.

- Category Cat5: The correct answer is not retrieved by system.

The ideal situation comes when the desired Object is retrieved as Cat1, though Cat2 and Cat3 would be reasonably acceptable. The obtained results are shown in Table 9. Though the obtained results with the TF-IDF method are quite reasonable, 81.18 % of the objects being retrieved between the first 5 options - and more than as Cat1, the FL based method turns out to be clearly better, with 92.45 % of the desired Objects retrieved - and more than three quarters as the first option -.

6 CONCLUSIONS

A FL based Term Weighting method has been presented as an alternative to classical TF-IDF method.

The main advantage for the proposed method is the obtention of better results, especially in terms of extracting not only the most suitable information but also related information.

This method will be used for the design of a Web Intelligent Agent which will soon start to work for the University of Seville web page.

REFERENCES

Aronson, A.R, Rindflesch, T.C, Browne, A. C., 1994. *Exploiting a large thesaurus for information retrieval.* Proceedings of RIAO, pp. 197-216.

Kosala, R., Blockeel, H., 2002. *Web Mining Research: A Survey*. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, Vol. 2 (2000).

Kwok, K. L., 1989. *A neural network for probabilistic information retrieval*. Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval. Cambridge, Massachusetts, United States.

Lee, D.L., Chuang, H., Seamons, K., 1997. *Document ranking and the vector-space model*. IEEE Software, Vol. 14, Issue 2, p. 67 – 75.

Lertnattee, V., Theeramunkong, T. 2003. Combining homogenous classifiers for centroid-based text classification. *Proceedings of the 7th International Symposium on Computers and Communications, pp. 1034-1039.*

Liu, S., Dong, M., Zhang, H., Li, R. Shi, Z., 2001. *An approach of multi-hierarchy text classification* Proceedings of the International Conferences on Info-tech and Info-net, 2001. Beijing. Vol 3, pp. 95 – 100.

Lu, M., Hu, K., Wu, Y., Lu, Y., Zhou, L., 2002. *SECTCS: towards improving VSM and Naive Bayesian classifier.* IEEE International Conference on Systems, Man and Cybernetics, Vol. 5, p. 5.

Ropero J., Gomez, A., Leon, C., Carrasco, A. 2007. *Information Extraction in a Set of Knowledge Using a Fuzzy Logic Based Intelligent Agent*. Lecture Notes in Computer Science. Vol. 4707, pp. 811-820.

Ruiz, M.E., Srinivasan, P., 1998. *Automatic Text Categorization Using Neural Networks.* Advances in Classification Research vol. 8: Proceedings of the 8th ASIS SIG/CR Classification Research Workshop. Ed. Efthimis Efthimiadis. Information Today, Medford:New Jersey. 1998. pp 59-72.

Salton, G., Buckley, C., 1996. *Term Weighting Approaches in Automatic Text Retrieval*. Information Processing and Management, Vol.32 (4), pp. 431-443.

Xu, J., Wang, Z. 2003. TCBLSA: A new method of text clustering. *International Conference on Machine Learning and Cybernetics. Vol. 1, pp. 63-66.*

Zhao, Y., Karypis, G., 2002. *Improving precategorized collection retrieval by using supervised term weighting schemes.* Proceedings of the International Conference on Information Technology: Coding and Computing, 2002. pp 16 – 21.