

# New Fuzzy Decision Tree Model for Text Classification

Ben Abdessalam Wahiba and Ben El Fadhl Ahmed

**Abstract** In this paper, a supervised automatic text documents classification using the fuzzy decision trees technique is proposed. Whatever the algorithm used in the fuzzy decision trees, there must be a criterion for the choice of discriminating attribute at the nodes to partition. For fuzzy decision trees two heuristics are usually used to select the discriminating attribute at the node to partition. In the field of text documents classification there is a heuristic that has not yet been tested. This paper tested this heuristic.

## 1 Introduction

Supervised classification is performed to assign automatically and independently one or more documents to one or more predefined categories [1]. There are various techniques for supervised classification, among the most known: the Bayesian networks, support vector machines, k-nearest neighbors, decision trees, etc. Among these techniques, only the decision trees easily generate a set of rules justifying the generated classification decisions. Other techniques generate in a more difficult and complicated way such set of rules.

Despite the wide spread of decision trees, this technique suffers from a problem that may affect its effectiveness: the problem of continuous value attributes. Let us take the example of a tree that will classify two men according to their sizes.

---

B.A. Wahiba  
Kingdom of Saudi Arabia, Taif University, Taif, Saudi Arabia

B.E.F. Ahmed (✉)  
Higher Institute of Management, Tunis University, Tunis, Tunisia  
e-mail: benelfadhl.ahmed@yahoo.fr

The first has a height of 181 cm; the second has a height of 180. The tree classifies a man as tall, if he has a height strictly larger than 180. In this example the tree will classify the first man as tall, but not the second, despite the invisible difference between the two sizes in the real world. One of the solutions used to solve this type of problem is the integration of fuzzy set theory with decision trees. This theory describes the phenomena of the real world in a graduated way closer to the reality [2].

A fuzzy decision tree is a good choice to use in the field of text classification to manage a big problem which is the uncertainty and the ambiguity necessarily related to the use of human language terms in the documents to be classified.

Different models have been developed in the literature to construct fuzzy decision trees. Most of these models are based on the fuzzy ID3 algorithm [3], which is an extension of the ID3 algorithm [4].

For the discrimination attribute selection in the fuzzy decision tree, two heuristics have been used in literature: The first is based on the minimization of the fuzzy entropy; the second is based on minimizing the classification ambiguity [5]. In the area of text classification with fuzzy decision tree, in our literature search, only the first heuristic has been implemented and tested [6]. Concerning the second heuristic based on the minimization of the classification ambiguity, it has not been yet implemented, nor tested.

Minimizing ambiguity has been used by [7] in their model with sample classification on sport to practice according to the state of the climate described by four attributes. We will study and apply this heuristic for text classification.

The rest of this paper is organized as follows: the second section presents the calculation of the classification ambiguity in a document classification context. The third section details our fuzzy decision tree model. In the fourth section we present the results of our model experiments. We end this paper with an analysis and discussion of the results found in the experimental section, and some perspectives.

## 2 Calculation of the Classification Ambiguity

Before starting the details of the classification ambiguity calculus, we have to prepare Table 1 as an example of the learning space that we will use in our model. For attributes, symbols L, M and H respectively designate the following linguistic terms “Low”, “Medium”, and “High”. These terms are the definition domain of the fuzzy variable “weight of the attribute in a document.”

In Table 1, the lines present the learning space documents. The columns present the attributes describing documents and the classes of these documents. We have transformed each attribute weight with continuous values into a fuzzy variable with values as one of the three linguistic terms: low, medium or high.

**Table 1** An example of the learning space

	Attributes												Classes of documents					
	A1			A2			A3			A4			C1	C2	C3	C4		
	L	M	H	L	M	H	L	M	H	L	M	H						
D1	0.9	0.1	0.0	1.0	0.0	0.0	0.4	0.5	0.1	0.1	0.0	0.0	0.4	0.6	0.1	0.3	0.2	0.6
D2	0.8	0.2	0.0	0.0	0.6	0.4	0.6	0.3	0.1	0.1	1.0	1.0	0.0	0.0	0.5	0.7	0.2	0.1
D3	0.0	0.7	0.3	0.5	0.2	0.3	0.1	0.7	0.2	0.2	0.0	0.0	0.5	0.5	0.3	0.2	0.4	0.2
D4	0.2	0.7	0.1	0.3	0.7	0.0	0.0	0.2	0.8	0.3	0.6	0.6	0.3	0.1	0.0	0.5	0.1	0.3
D5	0.0	1.0	0.0	1.0	0.0	0.0	0.3	0.4	0.3	0.3	0.0	0.0	0.0	1.0	0.7	0.1	0.4	0.5
D6	0.2	0.6	0.2	0.0	0.3	0.7	1.0	0.0	0.0	0.0	0.1	0.1	0.7	0.2	0.5	0.3	0.4	0.1
D7	0.9	0.1	0.0	0.5	0.5	0.0	0.0	0.3	0.7	0.7	0.3	0.3	0.4	0.3	0.1	0.8	0.5	0.4

The numerical values crossing documents with the attributes are calculated through membership functions which convert the continuous value weight of each attribute into a set of membership degrees. The details of these membership functions are presented in Sect. 3.2.

The method used to calculate the weight of each attribute is  $TF \times IDF$  method. The numerical values crossing documents with classes present the membership degrees calculated for each document to each class. In Sect. 2.4 we present our proposed method to calculate these degrees. According to [7], to calculate the classification ambiguity, we have to calculate some pre values:

- The truth degree of a fuzzy rule.
- The classification possibility with fuzzy evidence.
- The ambiguity of the evidence and fuzzy classification with fuzzy partitioning.

## 2.1 The Truth Degree of a Fuzzy Rule

A fuzzy rule is generally of the form “if A then B”, which defines a fuzzy relation between the two fuzzy sets A and B. According to Yuan and Shaw [7] a fuzzy rule is true, means that A implies B ( $A \rightarrow B$ ). They define the implication of A and B based on the principle of “subsethood”, that is A is a subset of B. In fuzzy logic, A is a subset of B, if for every element u belonging to the universe of discourse U the membership degree of u to A is less than the membership degree of u to B ( $\mu_A(u) \leq \mu_B(u)$ ). Often this condition is not satisfied for all u belonging to U, in this case we speak about partial “subsethood” or “the truth degree of the rule”. To calculate the truth degree of a rule of the form “if A then B”, denoted S (A, B), we use the following formula:

$$S(A, B) = \frac{M(A \cap B)}{M(A)} = \frac{\sum_{u \in U} \min(\mu_A(u), \mu_B(u))}{\sum_{u \in U} \mu_A(u)} \quad (1)$$

With, M is a function measuring the cardinality of a set.

As an example, let us calculate the truth degree of the following rule: “if A1 is L (low) then C = C3” denoted S (A1, C3).

$$S(A1, C3) = (0.2 + 0.2 + 0.0 + 0.1 + 0.0 + 0.2 + 0.5) / (0.9 + 0.8 + 0.0 + 0.2 + 0.0 + 0.2 + 0.9) = 0.4$$

## 2.2 The Classification Possibility with Fuzzy Evidence

Yuan and Shaw define the fuzzy evidence as a condition (simple or compound) using fuzzy variables (variables with linguistic terms as values). Example, “A3 is M” is a fuzzy evidence.

The possibility  $\Pi$  of the classification of an object in a class  $C_i$ , knowing a fuzzy evidence  $E$  is calculated using the following formula:

$$\Pi (C_i|E) = S (E, C_i) / \max_j S (E, C_j) \tag{2}$$

With, the index  $i$  is used to define a particular class  $C_i$ ,  $j$  is a counter used to browse the various classes. Let us consider the fuzzy evidence  $E1 = \text{“A3 is M”}$ . The calculation of  $\Pi (\text{Class}|E1)$ , requires the calculation of the 4 following values:  $S (E1, C1)$ ,  $S (E1, C2)$ ,  $S (E1, C3)$ ,  $S (E1, C4)$ .  $S (E1, C1) = 0.5$ ,  $S (E1, C2) = 0.58$ ,  $S (E1, C3) = 0.66$ ,  $S (E1, C4) = 0.7$ .  $\Pi (\text{Class}|E1) = \{0.5, 0.58, 0.66, 0.7\}$ , so after applying normalization (by dividing all the values by the biggest one) and sorting we will have:  $\Pi (\text{Class}|E1) = \{1, 0.94, 0.82, 0.71\}$ .

## 2.3 Classification Ambiguity with Fuzzy Evidence and Fuzzy Partitioning

In their model [7], Yuan and Shaw provide the formulas for calculating the classification ambiguity, first, with the verification of one fuzzy evidence, second, with the verification of several fuzzy evidences.

**In the First Case:** Knowing a fuzzy evidence  $E$ , the classification ambiguity noted  $G (\text{Class}|E)$  can be defined as follows:

$$G (\text{Class}|E) = \sum_{i=1}^n (\Pi (C_i|E) - (\Pi C_{i+1} | E)) * \ln (i) \tag{3}$$

With,  $\Pi (C_i|E) > \Pi (C_{i+1}|E)$  and  $n$  presents the number of the classes. As an example let us calculate the classification ambiguity with the fuzzy evidence  $E1 = \text{“A3 is M”}$  noted  $G (\text{Class}|E1)$ :

$$G (\text{Class}|E1) = (1 - 0.94) * \ln (1) + (0.94 - 0.82) * \ln (2) + (0.82 - 0.71) * \ln (3) + (0.71 - 0) * \ln (4) = 1.18$$

How to calculate the classification ambiguity if then we will partition by the attribute  $A2$ ?

**In the Second Case:** Let  $F$  be a fuzzy evidence,  $P$  a set of  $k$  fuzzy evidences  $\{E1, \dots, Ek\}$  set in the universe of discourse  $U$ . Yuan and Shaw define the fuzzy partition of  $P$  in  $F$  as:  $PIF = \{E1 \cap F, \dots, Ek \cap F\}$  To calculate  $G (P | F)$ , we use the following formula:

$$G(P|F) = \sum_{i=1}^k w(E_i|F) * G(E_i \cap F) \quad (4)$$

With,  $w(E_i|F) = M(E_i \cap F) / \sum_{j=1}^k (E_j \cap F)$ .

In the first case of this section we calculated  $G(\Pi(\text{class}|E1))$ , we will continue the partition with the attribute A2. Consider the following fuzzy evidences: E2: "A2 is L", E3: "A2 is M" and E4: "A2 is H". To calculate the value of the classification ambiguity after partitioning by the attribute A2, we have to calculate  $G(A2|E1)$ .

$$G(A2|E1) = w(E2|E1) * G(E2 \cap E1) + w(E3|E1) * G(E3 \cap E1) + w(E4|E1) * G(E4 \cap E1).$$

To calculate  $w(E2|E1)$ , we use the following formula:  $w(E2|E1) = M(E2 \cap E1) / (M(E2 \cap E1) + M(E3 \cap E1) + M(E4 \cap E1))$ . To calculate  $M(E2 \cap E1)$  it is sufficient to calculate the sum of the minimum of the two degrees belonging to E1 and E2 (which are the columns M of the attribute A3 and L of the attribute A2) for all documents in learning space. The same method is used to calculate  $w(E3|E1)$  and  $w(E4|E1)$ .

To calculate  $G(E2 \cap E1)$ , we have to calculate  $G(\Pi(C_i|E2 \cap E1))$ .

$\Pi(C_i|E2 \cap E1) = S(E2 \cap E1, C_i) / \max_j S(E2 \cap E1, C_j)$ .  $S(E2 \cap E1, C_i) = M(E2 \cap E1 \cap C_i) / M(E2 \cap E1)$ . We have to calculate the sum of the minimum of the three membership degrees of each document for the two evidences E2, E1, and the class  $C_i$ , divided after by the sum of the minimum of the two membership degrees for both E1 and E2 evidences. The same calculation is applied to calculate the entire formula  $G(A2|E1)$ .

## 2.4 The Calculation of Document Membership Degrees to the Different Classes

To calculate the membership degree of each document to each class in Table 1 we relied on Bayesian classifiers. The calculation of membership degree of a document  $D_i$  to a class  $C_j$  can be seen as the calculation of the value of  $P(C_j|D_i)$ . According to Bayes' theorem and after some simplifications we have [8]:

$$P(C_j|D_i) = (P(D_i|C_j) * P(C_j)) \quad (5)$$

To calculate  $P(D_i|C_j)$ , it is sufficient to calculate  $P(A_{i1}|C_j) * P(A_{i2}|C_j) * \dots * P(A_{in}|C_j)$ , with  $A_{ie}$ , is an attribute belonging to the document  $D_i$ ,  $e$  varies between 1 and  $n$ .  $n$  presents the number of attributes without repetition in the document  $D_i$ . To calculate  $P(A_{i1}|C_j)$  it is sufficient to calculate the number of documents belonging to the class  $C_j$  and containing  $A_{i1}$ , after, divide the result by the total number of documents [9]. Having a single attribute  $A_{ie}$  whose probability  $P(A_{ie}|C_j)$

equals to zero will affect the probability of the entire document. To solve this problem, simply [10] add the value one for each numerator and n in the denominator. So,  $P(A_{i1}|C_j) = (q + 1)/(n_i + \text{total number of document})$ . With q presents the number of documents belonging to the class  $C_j$  and containing the attribute  $A_{i1}$ .  $n_i$  is the total number of attributes without repetition in the document  $D_i$ . The calculation of  $P(C_j)$  is done by dividing the number of documents of class  $C_j$  by the total number of documents in all the classes, so:

$$P(C_j) = |C_j| / \sum_{i=1}^m |C_i| \tag{6}$$

m is the number of classes in the learning space.

### 3 The Proposed Model

Figure 1 shows the various stages of our fuzzy decision tree model for the text documents classification. Underlined steps are ones we have modified to fit the context of text documents classification with fuzzy decision tree using the minimization of the classification ambiguity.

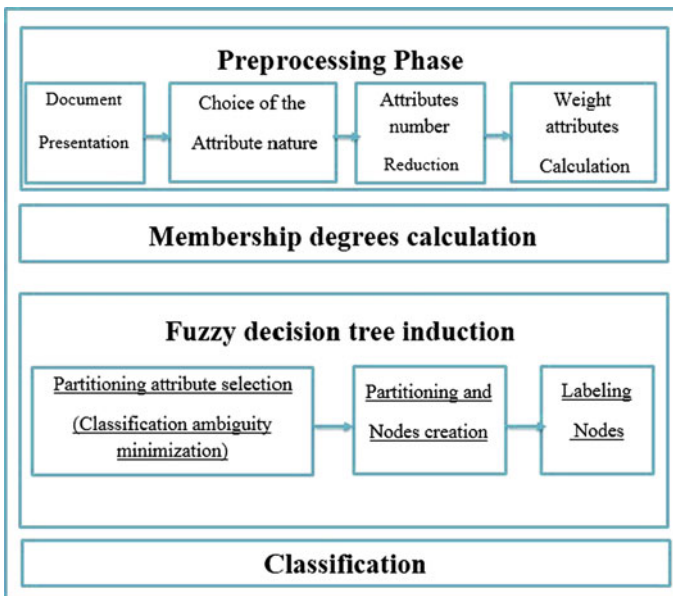


Fig. 1 The proposed model

### 3.1 Preprocessing Phase

For the document presentation we have chosen the vector format and the bag-of-words model. We chose the words as attributes to present the documents. For the attributes number reduction we first eliminated the stop words. Secondly we used a digital processing which orders the attributes based on a given measure then keeps only the first  $m$  attributes. The value of  $m$  depends on the experiments. We used two measures: the document frequency and CHI2 measure.

### 3.2 Membership Degrees Calculation

We used the triangular membership functions presented by [6]. (In the example we used in Sect. 2, we set the number of membership functions at 3 to simplify calculations). These functions are defined as follows:

$$T_1(x) = \begin{cases} 1 & \text{if } x < C_1 \\ (C_2 - x)/(C_2 - C_1) & \text{if } x \in [C_1, C_2] \\ 0 & \text{if } C_2 < x \end{cases}$$

$$T_i(x) = \begin{cases} 1 & \text{if } x > C_{i+1} \\ (C_{i+1} - x)/(C_{i+1} - C_i) & \text{if } x \in [C_i, C_{i+1}] \\ (x - C_i)/(C_i - C_{i-1}) & \text{if } x \in [C_{i-1}, C_i] \\ 0 & \text{if } x < C_{i-1} \end{cases}$$

$$T_k(x) = \begin{cases} 1 & \text{if } x > C_k \\ (x - C_k)/(C_k - C_{k-1}) & \text{if } x \in [C_{k-1}, C_k] \\ 0 & \text{if } x < C_{k-1} \end{cases}$$

To determine the number of the membership functions and their centers, [6] used an iterative algorithm that is based on a density distribution function and a measure called the F-statistic.

### 3.3 Induction of Fuzzy Decision Tree

Generating the tree is based on the algorithm proposed by [7]. This algorithm is detailed below. Before starting the details of the algorithm, a parameter must be set:  $\beta$ : This setting allows managing the evolution of the size of the tree. (In our experiments we set  $\beta$  to 0.7 as did Yuan and Shaw)



Below the general algorithm for the induction of fuzzy decision tree.

**Begin Algo\_Fuzzy\_Tree**

1. Search among the list of the attributes the one having the smallest classification ambiguity.
2. Use the found attribute to partition the current node.
3. For each child node Do
  - i. Calculate the truth degree  $T_i$  for each class  $C_i$ .
  - ii. **If** there is  $T_i > \beta$   
The node is a leaf, its class is that with the highest truth degree.

**Else**

- a. Calculate initial classification ambiguity  $G_i$  of the child node.
- b. Search among the remaining attributes if there is one with an ambiguity  $G$  after partition  $< G_i$ .
- c. **If** there is one  
Partition the current node and go to step 3.

**Else**

Consider the node as a leaf and assign it the class with the highest truth degree.

**End Alg\_Fuzzy\_Tree**

### 3.4 The Classification with Fuzzy Decision Tree

The classification of a document passes through two sub-steps: the conversion of the tree into a set of rules and the use of these rules for the classification.

For the first sub-step, each branch path leading from the root to a leaf is a rule. The set of attributes visited from the root to the leaf form the condition part of the rule. In the leaf, the class with the highest truth degree is the conclusion part of the rule.

After rules generation, we can classify the new documents. For classical decision trees, a new document undergoes one rule. With fuzzy decision trees, a document to classify may undergo several rules at once, so a document can be classified into different classes with different degrees. Classification in this case is obtained by the following steps:

1. For each rule, calculate the membership degree of the condition part of the rule. Calculating this membership degree is done by multiplying the membership degrees of the different attributes used in the condition part, the membership degree of the conclusion is the same as the membership degree of the condition.
2. If two rules are used to classify a document in the same class with two different degrees, retain the rule with the greatest degree.
3. If the rules give different classes for the same document, retain the class of the rule with the highest membership degree.

**Table 2** Comparison of three different algorithms using the document frequency

Number of used attributes	50	100	150	200	228
Original model of Yuan and Shaw	11.78	11.78	11.78	11.78	11.78
Proposed model	23.27	28.36	33.73	41.95	45.18
Fuzzy ID3	26.59	30.84	35.62	44.29	48.67

**Table 3** Comparison of three different algorithms using the measurement CHI2

Number of used attributes	50	100	150	200	228
Original model of Yuan et Shaw	11.78	11.78	11.78	11.78	11.78
Proposed model	22.34	26.12	30.88	39.54	43.24
Fuzzy ID3	25.27	28.52	34.57	43.12	45.52

## 4 Experimentation and Results

To evaluate our model we used a known measure: the macro F1.

The used data set to test our fuzzy decision tree system is the set “Reuters”. Initially this set contains 21578 documents [11, 12]. Different versions (or subsets) were generated later from the original version to adapt the needs of researchers.

In our experimentation, we used the subset R8 of the set “Reuters”. It consists of the 8 most frequent categories counting in total 7674 documents, divided in 5485 learning documents and 2189 test documents. The performed tests concern three models: the original model of Yuan and Shaw, the model we have proposed, and the fuzzy ID3 algorithm with 5 different values for the number of used attributes. The tests are operated following two scenarios. The first uses the frequency of the attribute in each document to reduce the number of attributes (Table 2). The second uses the index CHI 2 (Table 3). We used these two indices to reduce the number of attributes because it has been shown that these two indices offer the best classification results [13].

In both scenarios we chose the following values for the attributes number to consider in the learning phase: 50, 100, 150, 200, and 228. Over than 228 attributes the system launches a RAM saturation exception.

## 5 Analysis and Discussion

The application of the Yuan and Shaw model in its original version (Tables 2 and 3) gives very poor results (a rate of 11, 78 % for the measurement F1). In their model Yuan and Shaw gave directly the membership degrees of documents to different classes without specifying how to calculate these degrees. Applying the model of Yuan and Shaw directly for the documents classification, is assigning binary

degrees of belonging to different classes. If the document in the training set belongs to a class  $i$ , the membership degree to this class will be equal to one, zero if else. With such a distribution of membership degrees we get poor results because these membership degrees interfere in the calculation of the truth degree, the criterion that controls stopping the tree development. Applying our model significantly improves the classification results but these results are still not as good as those achieved by the fuzzy ID3 algorithm. We note that the maximum value of the macro-F1 in the different used models did not exceed 48 %. We recall here that the purpose of our experiments is not to measure the maximum efficiency for our model, but to compare its effectiveness with the famous fuzzy ID3 algorithm. We note that our fuzzy tree is not optimized, and the results of the classifications of our model can be further improved by optimizing the decision tree. This optimization can be done by testing other types of attributes (n-grams, the lemmas, the stems, etc.), by changing the membership function, optimizing the pruning of the tree (in our experiments we only used the pre pruning without implementing a post-pruning technique), increasing the number of used attributes, testing different values for  $\beta$  parameter used in the tree induction algorithm, etc. For the index used to reduce the number of attributes, we note that the frequency of the attribute offer better results than those found with CHI2 index. This is explained by the nature of each index. Indeed, the frequency of the attribute, promotes the most common attributes in the training set. An attribute with a non-zero frequency will have a greater opportunity to have a non-zero degree later in the fuzzification phase. The membership degrees having a zero value affect the performance of the model of Yuan and Shaw because they present absorbing elements in some formulas used in this model. Specifically the formula that calculates the truth degree of each classification rule and the formula that calculates the degree of total belonging to a rule. These two formulas are based on the two operators “min” and “product”.

## 6 Conclusion

In this paper, we proposed a new fuzzy decision tree model for the text documents classification. We tested a new heuristic based on the minimization of the classification ambiguity for the choice of discriminating attribute in the tree nodes partition step. Also we have proposed a method based on Bayes theorem to calculate the membership degrees of the training set documents for the different classes. In the fuzzification step, we used a method proposed by [6] to calculate the membership degrees of the attributes used in the tree induction phase. As it has been verified by its authors, this method has the advantage to autonomously find the appropriate number of linguistic terms for each attribute. In the experimentation results, our system gives better results than those presented by the original model of Yuan and Shaw. But compared with the famous fuzzy ID3 algorithm, our system

cannot exceed the results of this algorithm, at the same time it presents close results. In future works we will try to surpass even ID3 fuzzy algorithm, by testing other values for the parameter  $\beta$ , testing other operators different from the used “product” and “min”, and trying various other interpretations for the involvement operation.

## References

1. Sebastiani, F.: Machine learning in automated text categorization. *ACM computing survey*. **34** (1), 1–47 (2002)
2. Janikow, C.Z., Kawa, K.: Fuzzy decision tree FID. In: *Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society*, pp. 379–384 (2005)
3. Matiasko, K., Bohacik, J., Levashenko, V., Kovalik, S.: Learning fuzzy rules from fuzzy decision tree. *J. Inf. Control Manage. Syst.* **4**(2), 143–154 (2006)
4. Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986)
5. Wang, X., Chen, B., Qian, G., Ye, F.: On the optimization of fuzzy decision trees. *Fuzzy Sets Syst.* **112**, 117–125 (2000)
6. Wang, Y., Wang, Z.O.: Text categorization rule extraction based on fuzzy decision tree. In: *Proceedings of International Conference on Machine Learning and Cybernetics*, vol. 4, pp. 2122–2127 (2005)
7. Yuan, Y., Shaw, M.J.: Induction of fuzzy decision trees. *Fuzzy Sets Syst.* **69**, 125–139 (1995)
8. Raheel, S.: L'apprentissage arti ciel pour la fouille de donnees multilingues: application pour la classification automatique des documents arabes. Ph.D. thesis defended on October 22, 2010. Higher National School of Information and Communication Sciences, University of Lyon 2 (2010)
9. Rehel, S.: Categorisation automatique de textes et cooccurrence de mots provenant de documents non etiquetes. Faculty of Science and Engineering, University LAVAL, QUEBEC (2005)
10. Witten, I.H., Frank, E.: *Data mining: Practical Machine Learning Tools and Techniques*, 2nd edn, p. 2005. Morgan Kaufmann, San Francisco, CA (2005)
11. Reuters: <http://www.cs.umb.edu/smimarog/textmining/datasets> (2007)
12. Cardoso-Cachopo, A., Oliviera, A.L.: Semi supervised single label text categorization using centroid-based classifiers. In: *SAC'07 11–15 March 2007, Seoul, Korea* (2007)
13. Yang, Y., Pederson, J.: A comparative study on feature selection in text categorization. In: Fisher, J.D.H. (ed.) *The Proceedings of the Fourteenth International Conference on Machine Learning (ICML' 97)*, pp. 412–420 (1997)