Feature Selection and the Preservation of Infrequent and Highly Significant Attributes in the Context of Arabic Text Mining

Saeed Raheel
Department of Computer Science
American University of Science and Technology
Beirut, Lebanon

ABSTRACT
Effective feature selection is a key component for building an efficient automatic document classifier. We regularly encounter in the Arabic literature- especially the scientific one- infrequent non-Arabic words that are eliminated by practice during the pre-processing phase. Although infrequent, those words are highly pertinent to their documents and, thus, can contribute to build a more efficient classification model and enforce the subjectivity of the decision taken by the classifier. Therefore, we propose in this paper four different feature selection solutions that allow both preserving a maximum number of those words and getting satisfactory classification accuracy.

Keywords
Arabic Text mining, Machine Learning, Dimensionality Reduction, Automatic Classification

1. INTRODUCTION
A usual and frequently recurring phenomenon in the Arabic literature, especially the one dealing with subjects such as Science, Medicine, and Technology, is the inclusion of non-Arabic terms. In all of the literature written on the automatic classification of Arabic documents, researchers consider them as noisy terms and deliberately remove them during the pre-processing phase. Moreover, even if kept during that phase, the majority of those terms (if not all) will be statistically eliminated during the dimensionality reduction (henceforth, DR) phase and thus fail to contribute the prediction model built during the learning phase. We argue that, although infrequent, those terms are highly relevant to their documents and categories and many of them deserve to be kept throughout the classification cycle. For that purpose, we propose in this paper four dimensionality reduction techniques capable of accomplishing this task. In order to confirm their validity, the proposed methods have undergone more than 99 experiments in which they were tested against six of the very well-known DR methods in the literature: (Chi Square (χ^2), Odds Ratio (OR), Document Frequency (DF), Information Gain (IG), Gain Ratio (GR), and Mutual Information (MI)). They have shown to outperform them in both the number of foreign terms preserved terms and, sometimes, in the performance of the classifiers based on their resulting datasets..

2. FOREIGN WORDS: AN INTEGRATION AND PERFORMANCE PROBLEM
In order to be able to perform an automatic classification of Arabic documents, we have used an Arabic dataset containing 7,043 documents (a total of 31,333 words). As a first attempt to reduce the number of Arabic words, we decided to keep only those having a frequency bigger than 3 as well as all of the foreign terms. The result was a dataset composed of 8,027 words out of which 1,430 are foreign (17.81%).

By examining the foreign words in the new dataset, we realized that, although they are highly significant to their documents, they are sparse and have a very low frequency. Based on that and given that the DR phase is based mainly on the frequency of the attributes1, this will lead to the elimination of most of the foreign words and will hamper them from participating in the learning phase and contributing to the construction of the prediction model. In short, even if we reconsider the practice of eliminating them during pre-processing phase they will be, unfortunately, eliminated statistically. Therefore, it is necessary to find the proper solutions that allow the maximum number of them to bypass the preprocessing phase and participate in the next phases.

In this paper, we propose more than one solution that we compare against the six very well-known DR methods mentioned earlier. Table 1 and figure 1 display the number of foreign words preserved by each of the very well-known methods where the first n attributes were preserved, ($n \in \{1000,1500,2000,2500\}$). If we consider, for example, the case where we are looking to only preserve the first 1000 attributes, we conclude that very few of the 1,430 foreign words are preserved by 3 methods whereas the other 3 fail to preserve any. By increasing the threshold to 1500 attributes things improve slightly but still the number of foreign terms preserved is not enough. It is only by selecting the first 2000 attributes that we start getting enough foreign terms. Unfortunately, the performance of the classifiers based on the datasets generated by these methods was not quite satisfactory. Most of the methods generated datasets that led to mediocre classification results and only a few (such as $\chi 2$ or IG) were capable of preserving the most relevant attributes and helped the classifier in getting comparable results.

Table1. Number of preserved foreign words by the known DR methods

#of preserved attributes	χ^2	OR	DF	IG	GR	MI
1000	10	56	0	0	31	0
1500	53	105	56	3	31	1
2000	56	129	129	24	82	361
2500	172	221	221	62	413	716

[1] The attributes of a document are the words in their original or reduced form. The reduced forms of the words used in this paper were obtained using the tools of the Arabic morphological analyzer of the computerized dictionary DIINAR.1 developed by SILAT (http://silat.univ-lyon2.fr). It should be noted that the foreign words were kept as is and no further reduction in their forms was performed.

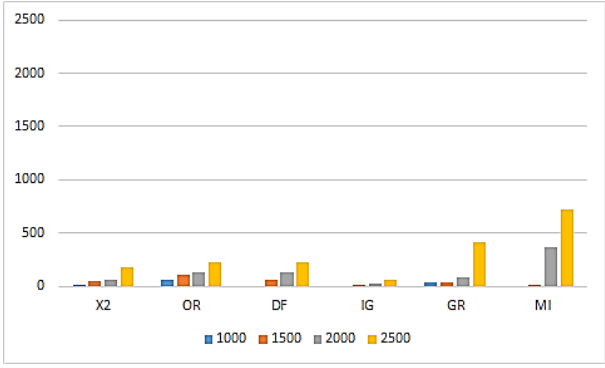

Fig. 1 No of preserved attributes by the known DR methods

In order to alleviate this problem we present in the next paragraph several solutions capable of both preserving a the most relevant Arabic attributes as well as the maximum number of foreign words and then generate from them datasets capable of leading to satisfactory classification results. Based on these two factors, we argue that a good DR method used with Arabic documents is one that satisfies the following two criteria:

Criteria#1: The preservation of the maximum number of foreign words.

Criteria#2: Insures a satisfactory classification performance.

In order to evaluate the results obtained by the classifiers, we used the F1 and accuracy measures. The accuracy is defined as the number of documents correctly classified in each category. In addition, we propose in this paper an Arabic-tailored evaluation measure (called "Eval_Multi") that takes into consideration the two previous criteria (i.e. it combines the F1 and Accuracy measures with the percentage of foreign words preserved). As such, the evaluation measure of a certain machine learning algorithm (denoted by Alg) is:

$$\boldsymbol{Eval\_Multi(Alg)} = F_1 + \frac{Accuracy}{100} + \frac{\#latin\ words\ preserved}{\#latin\ words}$$

where #latin words is the total number of foreign words present in the whole dataset.

That evaluation measure is valid since:

1. Its three components are compatible i.e. their values are between 0 and 1.
2. The "F1" and "Accuracy" measures are two very widely known and used methods in the literature. We could have relied on one of them only. However, let us consider the following example taken from our experiments using two datasets generated by the DF and IG methods and presented to the same machine learning algorithm. Table 2 shows a snapshot of results obtained. We can see clearly that the F1 measure is the same for both. This hampers us from deducing which of them did help the classifier get the better performance. However, by examining the values of the accuracy measure we can see that the method DF contributed to a slightly better performance. Therefore, by using both measures the problem is solved.

Table 2. A sample from the experiments

DR Method	DF	IG
# of foreign words preserved	0	0
Accuracy (%)	86.2525	86.2383
F-1 Measure	0.862	0.862

Finally, according to criteria #1, for a DR method to be considered a good one it should preserve the maximum number of foreign words. The degree of satisfaction to such a criterion can be measured by calculating the percentage of the foreign words preserved. Consequently, it is now unlikely to have two or more DR methods having the same value of *Eval_Multi*(). The method that is capable of contributing to getting the better performance and preserving the highest number of foreign words will now come first.

The main contribution behind this paper is that it proposes several solutions to the aforementioned problem and all of them are capable of satisfying the two previously stated criteria. We present each one of them in details in the next paragraph followed by the various experiments that confirm their validity.

3. OUR PROPOSED SOLUTIONS

Each proposed solutions is based mainly on the frequency of the attributes but adopts a different approach in calculating their scores. In what follows, we present the first one.

3.1 Cumulative Categorical Coefficient (« 3C »)

This method measures the frequency of an attribute in all of the categories. It is defined as:

$$3C(t_j) = \sum_{i=1}^{|C|} \frac{A_{ij}}{A_{ij} + C_{ij}}$$

where,

A_{ij} is the number of documents belonging to category c_i and containing the attribute t_j,

C_{ij} is the number of documents belonging to category c_i but not containing the attribute t_j.

and |c| is the total number of categories.

Once applied, we sort the attributes in their descending order of scores and then select the first n attributes.

Advantages:
This method is simple and easy to implement and use.

Its speed for calculating the scores is acceptable.

It is always capable of preserving more foreign words than the six known methods.

Weakness:
The simplicity and acceptable speed of this method made it a good competitor to the six well-known methods. However, we observed during the experiments that the classifiers using the datasets generated by this method were not frequently enough among the best 3 performing algorithms. A variant of this method called "3C-CS", which we present in the next paragraph, tries to rectify this problem by adopting a composite selection strategy rather than a sequential one. We propose as well another method called "CDFR-C" that relies

partially on the "3C" method to do the calculation of the scores of a subset of the attributes. Both methods prove to be better variants.

3.2 Cumulative Categorical Coefficient with a Composite Selection (« 3C-CS »)

This method is similar to the previous one but changes the selection strategy i.e. instead of selecting the top *n* attributes sequentially, we select them in a composite manner. In other words, once the scores are calculated, we select the first *na* Arabic attributes and then revisit the same list to select the first *nf* attributes such that:

$$n = na + nf$$

where *n* is the total number of attributes to be preserved by the DR method. This subtle modification proved to be very effective since the classifiers based on the datasets generated by this method were better performers than with the datasets generated by the original one. With the latter method, the algorithms were most of the time between the 5th and the 7th position and were only capable of being first 2 times whereas this variant placed them more frequently among the best 3 and even in the first position 3 times.

3.3 The Categorical Document-Frequency Ratio with Conformity (« CDFR-C »)

An alternative strategy to boost the chances of foreign attributes (as well as Arabic ones) during the DR phase was to use more than one method to calculate the scores. Therefore, we used:

1. The "3C" method to calculate only the scores of Arabic attributes and then sort them in their descending order of scores.
2. Another method based on the Categorical Document-Frequency Ratio ("CDFR") is used to calculate the scores of the foreign words. As such, to calculate the score of a foreign attribute a_j in a certain category c_i, the "CDFR" method normalizes the total number of documents in c_i that contain the attribute a_j by dividing it by the total number of documents remaining in c_i i.e.

$$CDFR(a_j) = \frac{\sum_{i=1}^{|c|} A_{ij}}{\sum_{i=1}^{|c|} C_{ij}} \qquad (3)$$

where,

A_{ij} is the total number of document in c_i that contain the attribute a_j,

C_{ij} is the remaining number of documents in c_i that do not contain the attribute a_j, and |c| is the total number of categories in the dataset.

Again, once the scores are calculated the attributes are sorted in their descending order of scores. As it was the case with "3C-CS", once the scores of the Arabic and foreign attributes are calculated, the first *na* Arabic attributes and the first *nf* foreign attributes are selected from these lists such that *n = na + nf*.

The advantage sought throughout this strategy is that, by using two separate methods to calculate the scores of Arabic and foreign attributes respectively, then each one focuses on the specificities of its attributes and chances are that the most significant ones will pass the DR phase.

Advantages:
This method is simple and easy to implement and use.

Its speed for calculating the scores is acceptable.

It is always capable of preserving more foreign words than the six known methods.

Weakness:
Consider two foreign attributes t_1 and t_2 having the same frequency *f* within the dataset but not the same distribution i.e. they both exist *f* times but t_1 appears *f* times in one category and zero times in the others while t_2 appears fd_i times in each category such that $f = \sum_{i=1}^{|c|} fd_i$ and |c| is the total number of categories in the dataset. During the experiments, we noticed that the "CDFR" method gives the same score to both t_1 and t_2 and thus both will be at the same position once the scores list is sorted. Given that the foreign attributes are scarce yet highly pertinent to their categories the method fails to reflect the notion of pertinence in its formula. In other words, t_1 is more pertinent to its category than t_2 and thus should be placed significantly higher than t_2 in the scores list.

Consequently, we introduce the notion of pertinence in "CDFR" by integrating a conformity measure – which was firstly proposed by [30] and calculated via the "Inverted Conformity Frequency" (ICF) method. By definition, the ICF of an attribute t_j according to a category c_i is calculated as:

$$ICF_{ij} = \frac{A_{ij}}{A_{ij}+C_{ij}} . \left(\log_2 \frac{A_{ij}}{A_{ij}+C_{ij}}\right) \forall A_{ij} = 0 \rightarrow ICF_{ij} = 0$$

Where,

A_{ij} is the total number of documents in c_i that contain the attribute t_j,

C_{ij} is the remaining number of documents in c_i that do not contain the attribute t_j.

Therefore, the "CDFR" method becomes "CDFR-C", where the suffix C stands for the conformity. This method is the third solution proposed in this paper and it's computed as follows:

$$CDFR - C(t_j) = CDFR(t_j) + \sum_{i=1}^{|c|} ICF_{ij}$$

$$CDFR - C(t_j) = \frac{\sum_{i=1}^{|c|} A_{ij}}{\sum_{i=1}^{|c|} C_{ij}} + \sum_{i=1}^{|c|} \frac{A_{ij}}{A_{ij}+C_{ij}} . \left(\log_2 \frac{A_{ij}}{A_{ij}+C_{ij}}\right)$$

As we have mentioned previously, this method is solely used to calculate the scores of foreign attributes whereas the "3C" method is used to calculate the scores of Arabic ones. When both scores' lists are ready, we use the composite selection to choose the best attributes from each set to form the new dataset with a reduced size.

The allocation of score calculation to two methods has proven to be a very good and effective choice since the experiments have shown that the algorithms using the datasets generated by this method were more often among the best 3 performers (where they were once in the first position).

The next paragraph presents the fourth and last solution. Unlike the previous ones, this one applies the idea of a composite selection to the already existing and known DR methods

3.4 Modifying the selection strategy of the known DR methods

The results obtained by the previous methods were very encouraging. This is why we decided to apply the composite selection strategy to the already existing methods. We have chosen to apply this strategy to two very well-known and reputed methods, mainly χ2 and IG. The modified versions of these two methods were called « χ2-CS» and « IG-CS ».

The experiments showed that the classifiers using the datasets generated by these two methods gave the best performance and accuracy i.e. they were 67% of the time in the first position (with 83% of the time the ones based on datasets generated by «χ2-CS» occupying the first position).

We have presented in the preceding paragraphs new and different solutions capable of both preserving a considerable number of foreign attributes and at the same time contribute in obtaining satisfactory classification accuracy. In what follows, we present the experiments conducted and the results obtained that confirm the validity of these solutions.

4. Experiments and Results

To confirm the validity of the proposed solutions, we conducted a large number of experiments using the following parameters:

- No of datasets: 99 (7,043 docs/dataset). Each uses the same set of documents but is represented by a different number of attributes.
- Type of attributes: Stems for the Arabic words (in [24] we found out that the best classification results in Arabic were obtained when using stems.) and the foreign words in their original form (i.e. without any morphological analysis)
- Data Mining Software: Weka.
- ML algorithms: Support Vector Machines (SVM) and Naïve Bayes Machines (NBM).
- Validation measure: 10-fold cross validation.
- Evaluation measures: F_1, Accuracy, and Eval_Multi

In order to make sure that none of ML algorithms nor the DR methods is biased by an inappropriate number of representative attributes, we followed an incremental strategy where we started the experiments with 250 attributes (foreign and Arabic) representing the 7,034 documents and increased that number gradually by a factor of 250 attributes/increment until we reached an upper bound of 2000 attributes.

4.1 Results from Criteria#1's Perspective

As a recall, Criteria#1 focuses on the number of foreign words preserved by the different DR methods. Table 3 and Figure 2 summarizes the results from this perspective:

Table 3. No of foreign attributes preserved by each DR method

n / DR Method	250	500	750	1000	1250	1500	1750	2000	2500
χ^2-CS	15	50	50	75	100	100	100	350	300
IG-CS	15	50	50	75	100	100	100	300	300
3C-CS	15	50	50	75	100	100	100	175	200
CDFR-C	15	50	50	75	100	100	100	175	200
3C	-	-	50	75	100	5	12	56	58
χ^2	-	-	2	10	26	53	54	56	172
OR	6	30	54	56	100	105	105	129	221
DF	-	-	-	-	-	3	9	24	62
IG	-	-	-	-	-	5	10	26	58
GR	21	21	21	31	31	31	37	82	413
MI	-	-	-	-	-	1	150	361	716

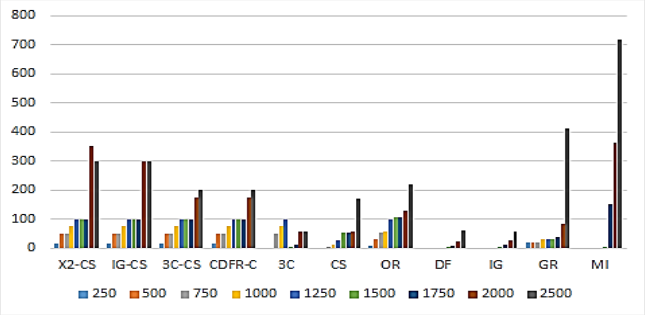

Fig. 2 No of foreign attributes preserved by each DR method

Observation #1:
The method χ^2 wasn't capable of preserving foreign attributes except when $n \geq 750$, knowing that the number of foreign attributes preserved is negligible. It is only when $n \geq 750$ that this method started to preserve an acceptable number of those attributes. In addition, the classification results obtained with this method were inferior to the ones obtained by the proposed methods (the classifiers always occupied the 5th position).

Observation #2:
The «MI» method wasn't capable of preserving foreign attributes except when $n \geq 1500$. The number of foreign attributes preserved increased significantly in the following datasets and was always greater than the other methods. Although that method preserved more foreign attributes than our methods, the classification results obtained with it were very bad (the classifiers always occupied the last position).

Observation #3:
The methods « IG » and « DF » were pretty much alike the « MI » method where they only started preserving foreign attributes when $n \geq 1500$. The classification results obtained with these methods were always inferior to the ones obtained with our methods.

Observation #4:
Unlike the other well-known methods, « OR » and « GR » were always capable of preserving foreign attributes regardless of the value of *n*. Similarly to the «MI» method, the classification results obtained with those methods were very bad (the classifiers always occupied the 9th and 10th positions).

Conclusion:
Unlike the results obtained with the very well-known methods, the proposed solutions had a stable behavior and were always capable of preserving a significant number of foreign attributes regardless of the value of n[2]. That number was always greater than the one preserved by the other methods and varied between 15 and 294!

Therefore, it is clear that the proposed methods satisfy criteria #1.

4.2 Results from Criteria#2's Perspective

Tables 4 and 5 show the positions of the classifiers based on the datasets generated by each DR method. We are going to focus on the first five positions in our discussion. As such, it is clearly seen that our methods outperform the others in approximately 80% of the time.

We can clearly see that the classifiers using the datasets generated by our methods have occupied the first 5 positions far more times than the ones based on datasets generated by the known methods (78 times against only 11 times). The « χ^2-CS » method contributed to the best classification results since the classifiers using its generated datasets were in the first position 10 times. In addition, the classifiers based on the datasets generated by the known methods never occupied the first position while the ones based on the datasets generated by our methods occupied it 18 times.

Table 4. The positions obtained by the classifiers based on the datasets generated by our dr methods

Method	Among the first 5 positions	The first	The positions detailed
3C	8	2	5,5,5,1,5,4,1,5
3C-CS	18	3	3,1,1,3,2,4,3,2,2,3,3,1,2,4,3,3,3,3
CDFR-C	18	1	2,2,3,1,3,3,2,3,3,4,4,2,3,2,4,4,4,4
χ^2-CS	16	10	4,4,1,5,1,5,4,1,1,4,1,1,1,1,1,1
IG-CS	18	2	1,4,2,2,4,2,4,1,5,2,2,3,4,3,2,2,2,2

Table 5. The positions obtained by the classifiers based on the datasets generated by the known dr methods

Method	No of times among the first 5 positions	No of times in the first position	The positions detailed
IG	3	0	4,3,5
DF	1	0	5
χ^2	7	0	5,5,5,5,5,5,5
OR	0	0	
GR	0	0	
MI	0	0	

[2] Except for the « 3C » method which failed to preserve foreign attributes unless $n \geq 750$.

Even though the previous discussions are based on the Eval_Multi() measure, the drawn conclusions hardly change if they were to be based only on the « F1 » and/or « Accuracy » measures while disregarding the number of foreign attributes preserved. This is due to the fact that the difference, in terms of the "F1" measure, between the performance of the classifiers based on the datasets generated by our methods and the ones based on the datasets generated by the known methods (in the case where the latter ones outperform the former ones) is negligible since it does not exceed 0.04 and this only happens a very rare number of times. The same is true for the "Accuracy" measure where the difference is very small and does not exceed 0.3% a very rare number of times.

Consequently, we conclude that our solutions satisfy criteria #2 body.

5. CONCLUSION

The decision to assign a category to a given document is subjective i.e. it is totally based on its content. Eliminating important elements of a document, as it is the case with foreign words, decreases the degree of subjectivity of the decision taken by the classifier. The foreign terms have a very low frequency which is why many of the DR methods based on statistics give them low scores. We presented in this paper a number of DR methods capable of solving this problem. Following a large series of conducted experiments, we concluded that our solutions allow for both the preservation of a considerable number of foreign attributes as well as the contribution in obtaining satisfactory classification results.

A remarkable advantage of the proposed solutions is that they can be used not only with Arabic but also with any other language where the writing practices are similar. Moreover, they are good potential candidate solutions to any other scenario where one might need to push forward a minority of attributes to get past the DR process and contribute in the learning phase.

6. REFERENCES

[1] Abbès, Ramzi et Dichy, Joseph, « Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1 » in : Heiden, Serge et Bénédicte Pincemain, Actes des JADT 2008, 9esjournées internationales d'analyse statistique des données textuelles (Proceedings of JADT 2008, 9th International Conference on Textual Data statistical Analysis).

[2] Cornuéjols, A. et Miclet, L. Apprentissage Artificiel : Méthodes et Algorithmes. Eyrolles 2002.

[3] Dichy J., Braham A., Ghazali S., Hassoun M., "La base de connaissances linguistiques DIINAR.1 (DIctionnaire INformatisé de l'Arabe, version 1)", paper presented at the International Symposium on The Processing of Arabic, Tunis (La Manouba), 18-20 April 2002.

[4] Fawcett, T. An Introduction to ROC Analysis. In ROC Analysis in Pattern Recognition, Vol. 27, No. 8. (June 2006), pp. 861-874.

[5] Feng, S. L. and Manmatha, R. 2005. Classification Models for Historical Manuscript Recognition. In Proceedings of the Eighth international Conference on Document Analysis and Recognition (August 31 - September 01, 2005). ICDAR. IEEE Computer Society, Washington, DC, 528-532.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Forman G. Computational Methods of Feature Selection. CRC Press/Taylor and Francis Group. 2007.

[8] Yoav Freund. Boosting a weak learning algorithm by majority. Information and Computation, 121(2) :256–285, 1995.

[9] Freund, Y., et Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139.

[10] Freund, Y., et Shapire, R. E. A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence, 14(5) :771-780, September, 1999.

[11] Jalam, R. (2003) "Apprentissage automatique et catégorisation de textes multilingues". Thèse de doctorat, Université Lumière Lyon 2.

[12] Joachims, T. (2002) Learning to Classify Text Using Support Vector Machines : Methods, Theory and Algorithms. Kluwer Academic Publishers.

[13] Khreisat, L. Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study, Proceedings of the 2006 International Conference on Data Mining. Las Vegas, USA, 2006, pp. 78-82.

[14] Kim, Y., Hahn, S., and Zhang, B. 2000. Text filtering by boosting naive Bayes classifiers. In Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Athens, Greece, July 24 - 28, 2000). SIGIR '00. ACM, New York, NY, 168-175.

[15] Kotsiantis S. Supervised Machine Learning : A Review of Classification Techniques, Informatica Journal 31 (2007) 249-268.

[16] Mitchell, T. M. (1997). Machine Learning Computer Science. McGraw-Hill. New York.

[17] István Pilászy. Text Categorization and Support Vector Machines. In the Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence

[18] Plantié M., Roche M., Dray G., EGC 2008 : Un système de vote pour la classification de textes d\\opinion, Laboratoire LGI2P, Laboratoire LIRMM.

[19] Quinlan, J. R. Induction of decision trees. Machine Learning, 1(1) :81–106, 1986.

[20] Quinlan, J. R. (1993) C4.5 : Programs for Machine Learning. Morgan Kaufmann Publishers Inc.

[21] Raheel, S. L'organisation des Connaissances et la Recherche d'Information Textuelles par l'Application des Méthodes Statistiques. 7ème colloque du chapitre français de l'ISKO. Lyon, France.

[22] Raheel S., J. Dichy, M. Hassoun. The Automatic Categorization of Arabic Documents by Boosting Decision Trees.In the proceedings of the 5th International IEEE/ACM Conference on Signal-Image Technology and Internet-Based Systems , IEEE CS Press, Marrakech, Morocco, November, 2009.

[23] Raheel, S., and Dichy J. Reducing Data Sparsity in a Language Dependent Automatic Classification of Arabic Documents. In the proceedings of the 3rd. IEEE International Conference on Information Systems and Economic Intelligence, Sousse, Tunisia, 2010. Pages : 37-46.

[24] Raheel, S., and Dichy, J. An Empirical Study on the Feature's Type Effect on the Automatic Classification of Arabic Documents**.** In the proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics. Iaşi, Romania. 2010. Springer LNCS 6008, Pages : 673-686.

[25] Rakotomalala R., "Arbres de Décision", Revue MODULAD, n°33, pp. 163-187, 2005.

[26] Schapire, R. E. et Singer, Y. (2000). BOOSTEXTER : a boosting-based system for text categorization. Machine Learning, 39(2/3) : 135-168.

[27] Sebag, M. et Gallinari, P. 2002. Apprentissage Artificiel: Acquis, Limites et Enjeux. In J. Le Maître, editor, *Assises 2002* : *Information - Interaction - Intelligence*. Cépaduès, 2002.

[28] Sebastiani, F. (1999). A tutorial on automated text categorization. Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR, 1999, pp. 7-35.

[29] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1) : 1.47.

[30] Shannon C.E., The communication theory of secrecy systems, Bell System Technical Journal 28 (1949) (4), pp. 656–715.

[31] Witten, I. H. and Frank, E. (2005) Data mining : practical machine learning tools and techniques. (second ed). Morgan Kaufmann, San Francisco, CA.

[32] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information Retrieval, 1 (1/2) : 60-69.

[33] Yang Y., Pederson J., 1997. A comparative study on feature selection in text categorization. In J. D. H. Fisher, editor, The Fourteenth International Conference on Machine Learning (ICML'97), page 412-420. Morgan Kaufmann.

[34] Zighed, D. A. et Rakotomalala, R. (2000). Graphes d'induction. Apprentissage et Data Mining. Hermes Science Publication, Paris.