Experimenting N-Grams in Text Categorization

Abdellatif Rahmoun and Zakaria Elberrichi
Faculty of Computer and Information Technology, University of King Faisal, KSA

Abstract: *This paper deals with automatic supervised classification of documents. The approach suggested is based on a vector representation of the documents centred not on the words but on the n-grams of characters for varying n. The effects of this method are examined in several experiments using the multivariate chi-square to reduce the dimensionality, the cosine and Kullback&Liebler distances, and two benchmark corpuses the reuters-21578 newswire articles and the 20 newsgroups data for evaluation. The evaluation was done, by using the macroaveraged F_1 function. The results show the effectiveness of this approach compared to the Bag-Of Word and stem representations.*

Keywords: *Text categorization, n-grams, multivariate chi-square, cosine measure, reuters21578, 20 news groups.*

Received April 5, 2006; accepted June 1, 2006

1. Introduction

Text categorization (TC) consists in assigning a text to one or more categories among a preset list [15]. The majority of the existing methods for TC are based on the relations between three parameters (the class, the document, the term). Indeed, all these methods are based on:

1. A relation of membership of a document to one or more classes.
2. A relation of importance of a term in one or more documents.

These methods are confronted with many problems:

1. The cost of the treatment, because the number of the terms intervenes in the expression of the complexity of the algorithm.
2. The weak and strong frequency of certain terms: we cannot build reliable rules starting from some occurrences in the training set. In the other hand, it was observed that the most frequent terms also do not bring important information since they are present everywhere.
3. More the number of documents per class is high, more space memory as well as computational time is needed.
4. The majority of the existing methods are based on the comparison between the document to classify and all the documents pre-classified. Therefore, more the number of documents is significant, more the response time will be significant.
5. The documents are generally of unspecified size, this size can influence on the performances of a categorization. Indeed, there are methods that support large size documents, as there are also methods that support small size documents.

In this article, we present a new method for text categorization, which makes it possible to mitigate the disadvantages just quoted. This method rests on:

1. A direct relation "term-class", instead of passing by the two relations: "term-document" and "document-class". This will reduce the response time and eliminate the influence of the documents size in the performances of text categorization.
2. A reduction of dimensionality by preserving only the terms that characterize best a class compared to the other classes. This will reduce the capacity of the used memory and neglect the most frequent terms that do not bring any information as well as the terms of weak frequencies.

In section 2 we detail the approach suggested with all its stages. In section 3 we present the experiments that make the evaluation of the relevance of this approach possible. We will conclude in section 4, by indicating the main characteristics of the approach.

2. Our Approach

In this approach, we use the N-grams method as a method of representation of the data and the multivariate χ_2 method for the selection of the characteristic terms. Figure 1 illustrates the approach with all its stages.

We distinguish in this approach two phases, a training phase and a classification phase.

2.1. Training Phase

The first issue that needs to be addressed in TC is how to represent texts so as to facilitate machine manipulation but also to retain as much information as needed. The commonly used text representation is the Bag-Of-

Words, which simply uses a set of words and the number of occurrences of the words to represent documents and categories [12]. Many efforts have been taken to improve this simple and limited text representation. For example, [9] uses phrases or word sequences to replace single words. In our approach, we use the N-grams representation.

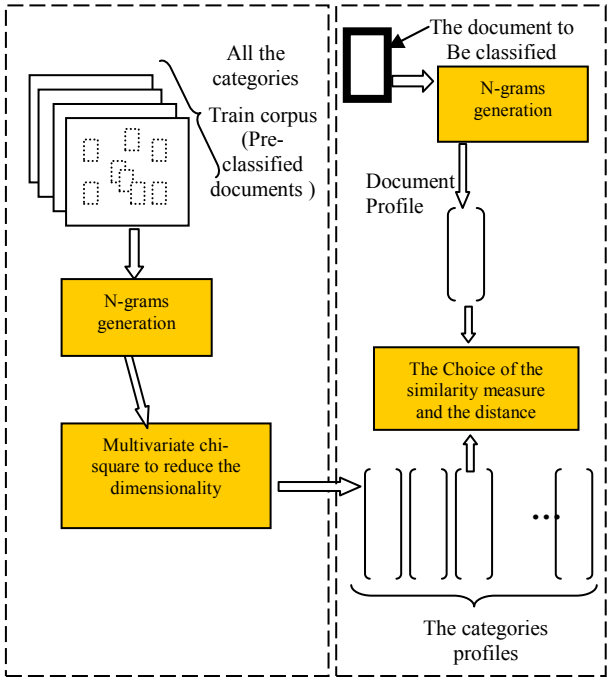

Figure 1. The N-grams based approach.

2.1.1. N-Grams Generation

An N-gram is a sequence of N consecutive characters. In a text, we locate all the n-grams present, and then we count their frequencies. We replace the space character by the character "_", to facilitate detection. This technique, purely statistical, does not require any knowledge of the document language. Another advantage of the N-grams is the automatic capture of the most frequent roots [7]: for example, thanks to this technique, we find the common root of: to nourish, nourished, nourishes, nourishing, nourishment…etc. The tolerance to spelling mistakes and deformations is also a significant property [11]. Lastly, this technique does not need to eliminate the stop words or to proceed to the lemmatisation, or Stemming [6, 14, 16].

This first step consists in representing each category by a vector where figure all the n-grams generated with their number of occurrence.

2.1.2. Selection of the Characteristic N-Grams

In the second step, we generate a profile for each category. The profile of a category must contain all N-grams that characterize this category compared to the other categories. To build the profiles of the categories, it is necessary to use a method of term selection. There are several methods of term selection.

In our work, we chose to use the multivariate χ_2 method to discriminate the categories.

The χ_2 statistic measures the degree of association between a term and the category. Its application is based on the assumption that a term whose frequency strongly depends on the category in which it occurs will be useful for discriminating among the categories. For the purpose of dimensionality reduction, terms with small χ_2 values are discarded.

The χ_2 multivariate, noted χ^2 multvariate is a supervised method allowing the selection of terms by taking into account not only their frequencies in each category but also the interaction of the terms between them and the interactions between the terms and the categories. The principle consists in extracting K better features characterizing best the category compared to the others, this for each category.

With this intention, the matrix (term-categories) representing the total number of occurrences of the p features in the m categories is calculated as shown in Fgure 2. The total sum of the occurrences is noted N. The values N_{jk} represent the frequency of the feature X^j in the category e_k. Then, the contributions of these features in discriminating categories are calculated as indicated in equation (1), then sorted by descending order for each category. The evaluation of the sign in the equation (1) makes it possible to determine the direction of the contribution of the feature in discriminating the category. A positive value indicates that it is the presence of the feature which contribute in the discrimination while a negative value reveals that it is its absence which contribute in it.

	e_1		e_k		e_m	
X^1	N_{11}	…	N_{1k}	…	N_{1m}	$N_{1.}$
⋮	⋮		⋮		⋮	
X^j	N_{j1}	…	N_{jk}	…	N_{jm}	$N_{j.}$
⋮	⋮		⋮		⋮	
X^n	N_{n1}	…	N_{nk}	…	N_{nm}	$N_{n.}$
	$N_{.1}$		$N_{.k}$		$N_{.m}$	$N = N_{..}$

Figure 2. Matrix of features frequencies in categories.

$$C_{jk}^{\chi^2} = N \frac{(f_{jk} - f_{j.} f_{.k})^2}{f_{j.} f_{.k}} \times sign\,(f_{jk} - f_{j.} f_{.k}) \quad (1)$$

with, $f_{jk} = \frac{N_{jk}}{N}$ representing the relative frequencies of the occurrences.

The principal characteristics of this method are:

1. It is supervised because it is based on the information brought by the categories.

2. It is a multivariate methode because it evaluates the role of the feature while considering the other features.
3. It considers interactions between features and categories.

2.2. Classification Phase

In this step, we compare the profile of the document to be categorized with the profiles of the categories that were already calculated in the training phase. This comparison initially consists in weighting each term (N-grams) in each profile (categories and new document), then calculating the distance between the profile of the document and the profile of each category. The document will be assigned to the category to which its profile is closest.

2.2.1. Weighting

Weighting makes it possible to represent the importance of the term (N-grams) in the category concerned. The number of occurrences of the term in the category is the simplest way to calculate this value, but is not very satisfactory because it does not take into account the other categories, or we need to compare.

In our experiments, we used the standard *tfidf* function, defined as:

$$tfidf\ (t_k, c_i) = tf\ (t_k, c_i) \times Log\left(\frac{|C|}{df\ (t_k)}\right) \quad (2)$$

where:

- $tf\ (t_k, c_i)$ denotes the number of times feature t_k occurs in category c_i.
- $df\ (t_k)$ denotes the number of categories in which feature t_k occurs.
- $|C|$ denotes the number of categories.

2.2.2. The Distance Calculation

After having weighted the terms (N-grams), it is necessary to calculate the distance between the categories profiles and the profile of the document to be categorized. For that, several measures of similarity can be used. For our experiments, we chose to use the Cosine measurement and the Kullback & Liebler in order to study the influence of the similarity measure in the performance of a system of categorization.

2.2.2.1. COSINE

The dominant similarity measure in information retrieval and text classification is the *cosine similarity* between two vectors. Geometrically, the cosine similarity evaluates the cosine of the angle between two vectors d_1 and d_2 and is, thus, based on angular distance. This allows us to abstract from varying vector length. The cosine similarity can be calculated as the normalized:

$$S_{i,j} = \frac{\sum_{w \in i \cap j} TFIDF_{w,i} \times TFIDF_{w,j}}{\sqrt{\sum_{w \in i} TFIDF_{w,i}^2} \times \sqrt{\sum_{w \in j} TFIDF_{w,j}^2}} \quad (3)$$

with:

- w: a feature, I and J: the two vectors (profiles) to be compared.
- TFIDF w,I the weight of the term w in I and TFIDF w,j the weight of the term w in J.

This can be translated in the following way:

"More there are common features and more these features have strong weightings, more the similarity will be close to 1, and vice versa"

2.2.2.2. KULLBACK & LIEBLER

Kullback and Liebler studied in 1951 a statistical measurement of information called function of discrimination by taking into account two probability distributions. The Kullback & Liebler measurement also known under the name of the relative entropy calculates the divergence between two probability distributions. The divergence between two probabilities P and Q on a finished set X is defined as follows:

$$D\ (P, Q) = \sum_{x \in X} P\ (x) \times \log \frac{P\ (x)}{Q\ (x)} \quad (4)$$

It should be noted that this divergence is not symmetrical $(D(P,Q) \neq D(Q,P))$; therefore it cannot be used like measure of distance.

In our work, we chose to use the symmetrical divergence Kullbak&Liebler i. e. the distance Kullback&Liebler defined as follows:

$$D(P,Q) = \sum_{x \in X} \left((P(x) - Q(x)) \times \log \frac{P(x)}{Q(x)} \right) \quad (5)$$

This similarity measurement was used in various fields such as the treatment of natural languages [2], parole recognition [4] as well as information retrieval and themes identification [1].

With regard to the field of the text categorization, this measurement is used to calculate the distance between the profile of the document and the profile of the category as follows:

$$KLD(c_i, d_j) = \sum \left\{ \left(P(t_k, c_i) - P(t_k, d_j)\right) \times \log\left(\frac{P(t_k, c_i)}{P(t_k, d_j)}\right) \right\} \quad (6)$$

In its calculation, four cases are taken into account:

1. $(t_k \in d_j)$ and $(t_k \in c_i)$ i. e.: the t_k term appears in the category profile and the document profile.

2. ($t_k \in d_j$) and ($t_k \notin c_i$) i. e.: the t_k term appears in the document profile but does not appear in the category profile.
3. (tk ∈ dj) and (tk ∉ ci) i. e.: the tk term does not appear in the document profile but appears in the category profile.
4. (tk ∉ dj) and (tk ∉ ci) i. e.: the tk term appears neither in the document profile nor in the category profile.

The probability of appearance of a t_k term in a category profile is defined as follows:

$$\begin{cases} P(t_k, c_i) = \dfrac{tf(t_k, c_i)}{\sum_{x \in c_i} tf(t_x, c_i)} & \text{If the term } t_k \text{ appears in the profile of } c_i \\ \varepsilon \text{ (epsilon)} & \text{If not.} \end{cases} \quad (7)$$

In the same way, the probability of appearance of a t_k term in the document profile is defined as follows:

$$\begin{cases} P(t_k, d_j) = \dfrac{tf(t_k, d_j)}{\sum_{x \in c_i} tf(t_x, d_j)} & \text{If the term } t_k \text{ appears in the profile of } d_J \\ \varepsilon \text{ (epsilon)} & \text{If not.} \end{cases} \quad (8)$$

where:

- $P(t_k, c_i)$ is the conditional probability of a term in a category with $\sum_{x \in d_j} tf(t_k, d_j) = 1$
- ε is a probability granted to the terms which do not appear neither in the document, nor in the category.

For each category, it is necessary to standardize the distance, because the categories are of different sizes. By consequence, we will use the standardized distance Kullback & Liebler:

$$KLD^*(c_i, d_j) = \frac{KLD(c_i, d_j)}{KLD(c_i, 0)} \quad (9)$$

where KLD (c_i, 0) represents the distance between the category and an empty document.

Finally, after having calculated the distance $KLD^*(C_i, d_j)$ between the document to be categorized and all the categories, the document will be assigned to the closest category:

$$H_{KLD^*}(d_j) = \arg \min_{c_i \in C} KLD^*(c_i, d_j) \quad (10)$$

3. Experimental Results

This section describes our experiments. The approach suggested is tested on the two corpora most used by the researchers in this field, the Reuters 21578 and the 20Newsgroups by using the similarity measurements COSINE and Kullback & Liebler.

After having analysed these results, we will carry out a comparison between the N-grams representation and the Bag-Of-Words and stem representations.
Experimental results reported in this section are based on the so-called "F_1 measure", which is the harmonic mean of precision and recall.

$$F_1(recall, precision) = \frac{2 \times recall \times precision}{recall + precision} \quad (11)$$

In the above formula, precision and recall are two standard measures widely used in text categorization literature to evaluate the algorithm's effectiveness on a given category where:

$$precision = \frac{true\ positive}{(true\ positive) + (false\ positive)} \times 100$$

$$recall = \frac{true\ positive}{(true\ positive) + (false\ negative)} \times 100 \quad (12)$$

We also use the macroaveraged F_1 to evaluate the overall performance of our approach on given datasets. The macroaveraged F_1 compute the F_1 values for each category and then takes the average over the per-category F_1 scores. Given a training dataset with *m* categories, assuming the F_1 value for the i-th category is $F_1(i)$, the macroaveraged F_1 is defined as :

$$\text{macroveraged } F_1 = \frac{\sum_{i=1}^{m} F_1(i)}{m} \quad (13)$$

3.1. The Effect of the Choice of N in the Results of the Approach

The representation based on the N-grams is dependent on an essential parameter : the value of N, i. e. : the number of characters that each N-gram will contain.
In this section, we try to answer two interesting questions:

- Which value of N gives the best results?
- To combine N-grams is or is not an improvement?:

To answer the first question, we tried out the approach on the two corpora Reuters and 20Newsgroups for values of N ranging between 2 and 7 with K (the profile size) varying. Tables 1 and 2 respectively have the results obtained.

By analysing the results of tables 1 and 2, we notice that the performances improve by increasing the value of N until N=5 and start degrading from N=6.

By comparing the results of table 3 with those of tables 1 and 2, we realize that the combination of the n-grams does not result in any noticeable improvement in the performances; the results are quasi similar. On this point, we find the conclusions of many authors in particular in [3, 6, 10].

Table 1. Results comparison (MacroAveraged F_1) on reuters 21578

	N=2	N=3	N=4	N=5	N=6	N=7
K=100	0.458	0.643	0.701	0.704	0.680	0.629
K=200	0.462	0.650	0.702	0.702	0.681	0.626
K=300	0.462	0.648	0.703	0.707	0.685	0.626
K=400	0.462	0.646	0.701	0.704	0.686	0.624
K=500	0.462	0.649	0.700	0.703	0.685	0.621
K=600	0.462	0.648	0.699	0.702	0.685	0.620
K=700	0.462	0.648	0.697	0.701	0.685	0.622
K=800	0.462	0.648	0.695	0.701	0.685	0.622

Table 2. Results comparison (MacroAveraged F_1) on the 20 newsgroups.

	N=2	N=3	N=4	N=5	N=6	N=7
K=100	0.265	0.524	0.639	0.677	0.667	0.641
K=200	0.265	0.542	0.659	0.699	0.694	0.669
K=300	0.265	0.549	0.666	0.707	0.705	0.676
K=400	0.265	0.550	0.670	0.714	0.711	0.678
K=500	0.265	0.550	0.674	0.716	0.713	0.680
K=600	0.265	0.550	0.674	**0.717**	0.714	0.681
K=700	0.265	0.550	0.676	0.717	0.715	0.681
K=800	0.265	0.551	0.676	0.717	0.716	0.681

Table 3. Results (MacroAveraged F1) of the combination of N-grams.

Ngrams	2+3+4+5+6		3+4+5+6		4+5+6	
Corpus	R	N	R	N	R	N
K=100	0.679	0.606	0.680	0.608	0.690	0.616
K=200	0.707	0.640	0.705	0.640	0.708	0.643
K=300	0.706	0.651	0.704	0.652	0.702	0.654
K=400	0.703	0.661	0.703	0.659	0.705	0.662
K=500	0.707	0.668	0.707	0.667	0.707	0.668
K=600	0.707	0.671	0.707	0.671	**0.709**	0.672
K=700	0.705	0.678	0.705	0.677	0.706	0.678
K=800	0.705	0.680	0.704	0.680	0.706	**0.681**

3.2. Influence of the Size of the Profiles

By analysing the results of tables 1,2 and 3, we note that by increasing the size of the profiles (K), the performances improve, and then stabilize for a certain value that varies according to the corpus. This value is between 200 and 400 for the Reuters corpus, while it is between 600 and 800 for the 20Newsgroups corpus.

These observations are made on the total results of the two corpora but for studying well the influence of the size of the profiles, a detailed analysis concerning the categories was necessary.

Table 4 has the results obtained for some categories of the Reuters corpus, for a value of N=5. The results presented are uneven on the optimal value of the profiles size. Indeed, for the categories whose training sets are rather significant (Earn, Acquisition), the performances improve by increasing the size of the profiles up to the K=500 value. On the other hand for the categories whose training sets are less significant (Corn, Trade, Ship), we note no improvement by increasing the size of the profiles.

Concerning the 20Newsgroups corpus, we noted an improvement for all the categories by increasing the size of the profiles.

Within sight of these results, we can affirm that the best size for a profile is dependent on the training corpus. Indeed, for a rather rich and well-balanced training corpus, the profiles must have a significant number of features (N-grams) in order to be able to discriminate well the categories.

Table 4. Detailed results of some categories of the corpus reuters.

	Acquisition	Corn	Earn	Trade	Ship
K=100	0.938	0.483	0.971	**0.821**	**0.681**
K=200	0.945	0.477	0.974	0.792	0.664
K=300	0.949	**0.480**	0.976	0.803	0.664
K=400	0.952	0.463	**0.978**	0.794	0.657
K=500	**0.956**	0.463	0.978	0.791	0.648
K=600	0.954	0.463	0.978	0.787	0.651
K=700	0.953	0.459	0.978	0.786	0.651
K=800	0.952	0.459	0.976	0.792	0.651

3.3. Influence of the corpus

By comparing the results presented in tables 2 and 3, we note, that the best performances were obtained with the corpus 20Newsgroups.

In order to argue this observation, we have to carry out an additional experiment to highlight the influence of the quality of the training corpus in the performances. This experiment quite simply consisted in calculating the COSINE distance between the profiles of the categories. Tables 5 and 6 have the results of this experiment.

The results of table 5 indicate that the categories Corn, Grain and Wheat are very close to each other, that makes difficult their discrimination. On this point, we find the conclusions of many authors in particular in [5] on the difficulties to categorize the Reuters corpus.

These results give a good explanation to the results presented in table 4. Indeed, the categories whose training corpus is rather significant were well distanced compared to the other categories and consequently they are well discriminated.

On the other hand, the categories of the corpus 20Newsgoups were better distanced than the categories of the Reuters corpus as table 6 shows. It is due to the richness and the equitable distribution of the 20Newsgroups corpus.

With an aim of being more argumentative, we carried out another experimentation that consisted in evaluating our approach on the corpus used by Gongde Guo. This corpus contains only the 7 most significant and well-discriminated categories of the Reuters corpus and which are *Acquisition, Corn, Crude, Earn, Interest, Ship, and Trade.*

Table 7 has the results of our approach on this corpus. It shows an improvement in the performances when eliminating the categories Grain, and Wheat too close to corn, and Moneyfx close to interest. The results were a lot better and gave a good explanation to the results of table 5. The MacroAveraged F1 obtained by Gongde guo on this corpus is 0.860 [8].

Table 5. COSINE distance between the category profiles of reuters corpus, for N=5, K=500.

	Acq	Corn	Crude	Earn	Grain	Interest	Moneyfx	Ship	Trade	Wheat
Acq	1	0.013	0.103	0.231	0.039	0.055	0.075	0.037	0.074	0.015
Corn	0.013	1	0.012	0.005	**0.850**	0.005	0.009	0.149	0.415	**0.669**
Crude	0.103	0.012	1	0.069	0.045	0.044	0.055	0.103	0.061	0.012
Earn	0.231	0.005	0.069	1	0.014	0.031	0.034	0.014	0.028	0.005
Grain	0.039	**0.850**	0.045	0.014	1	0.028	0.044	0.178	0.098	**0.849**
Interest	0.055	0.005	0.044	0.031	0.028	1	**0.682**	0.009	0.131	0.009
Moneyfx	0.075	0.009	0.055	0.034	0.044	**0.682**	1	0.015	0.279	0.018
Ship	0.037	0.149	0.103	0.014	0.178	0.009	0.015	1	0.042	0.159
Trade	0.074	0.415	0.061	0.028	0.098	0.131	0.279	0.042	1	0.051
Wheat	0.015	**0.669**	0.012	0.005	**0.849**	0.009	0.018	0.159	0.051	1

Table 6. COSINE distance between categories profiles of corpus 20 newsgroups, for N=5 et K=500.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
C1	1	0.015	0.003	0.007	0.013	0.018	0.008	0.012	0.009	0.009	0.019	0.017	0.009	0.041	0.024	0.375	0.050	0.056	0.069	0.381
C2	0.015	1	0.072	0.085	0.110	0.238	0.079	0.027	0.014	0.011	0.009	0.069	0.069	0.029	0.080	0.011	0.016	0.007	0.022	0.013
C3	0.003	0.072	1	0.071	0.048	0.114	0.037	0.013	0.020	0.006	0.004	0.009	0.016	0.006	0.015	0.003	0.008	0.002	0.005	0.003
C4	0.007	0.085	0.071	1	0.359	0.064	0.134	0.043	0.015	0.010	0.010	0.028	0.086	0.015	0.016	0.007	0.017	0.007	0.012	0.007
C5	0.013	0.110	0.048	0.359	1	0.074	0.135	0.047	0.014	0.016	0.013	0.030	0.090	0.019	0.031	0.007	0.013	0.009	0.014	0.026
C6	0.018	0.238	0.114	0.064	0.074	1	0.050	0.023	0.011	0.012	0.012	0.043	0.066	0.025	0.044	0.015	0.020	0.012	0.021	0.014
C7	0.008	0.079	0.037	0.134	0.135	0.050	1	0.064	0.040	0.048	0.050	0.017	0.069	0.022	0.027	0.013	0.020	0.009	0.026	0.009
C8	0.012	0.027	0.013	0.043	0.047	0.023	0.064	1	0.098	0.018	0.014	0.024	0.058	0.018	0.039	0.011	0.045	0.014	0.034	0.011
C9	0.009	0.014	0.020	0.015	0.014	0.011	0.040	0.098	1	0.024	0.012	0.007	0.039	0.016	0.015	0.009	0.018	0.008	0.018	0.007
C10	0.009	0.011	0.006	0.010	0.016	0.012	0.048	0.018	0.024	1	0.320	0.005	0.025	0.023	0.026	0.015	0.021	0.010	0.027	0.017
C11	0.019	0.009	0.004	0.010	0.013	0.012	0.050	0.014	0.012	0.320	1	0.005	0.024	0.031	0.031	0.012	0.030	0.010	0.034	0.020
C12	0.017	0.069	0.009	0.028	0.030	0.043	0.017	0.024	0.007	0.005	0.005	1	0.059	0.023	0.036	0.013	0.067	0.026	0.083	0.019
C13	0.009	0.069	0.016	0.086	0.090	0.066	0.069	0.058	0.039	0.025	0.024	0.059	1	0.031	0.060	0.015	0.029	0.013	0.027	0.007
C14	0.041	0.029	0.006	0.015	0.019	0.025	0.022	0.018	0.016	0.023	0.031	0.023	0.031	1	0.038	0.035	0.031	0.019	0.067	0.034
C15	0.024	0.080	0.015	0.016	0.031	0.044	0.027	0.039	0.015	0.026	0.031	0.036	0.060	0.038	1	0.018	0.029	0.018	0.059	0.019
C16	0.375	0.011	0.003	0.007	0.007	0.015	0.013	0.011	0.009	0.015	0.012	0.013	0.015	0.035	0.018	1	0.035	0.052	0.060	0.619
C17	0.050	0.016	0.008	0.017	0.013	0.020	0.020	0.045	0.018	0.021	0.030	0.067	0.029	0.031	0.029	0.035	1	0.068	0.134	0.061
C18	0.056	0.007	0.002	0.007	0.009	0.012	0.009	0.014	0.008	0.010	0.010	0.026	0.013	0.019	0.018	0.052	0.068	1	0.0064	0.055
C19	0.069	0.022	0.005	0.012	0.014	0.021	0.026	0.034	0.018	0.027	0.034	0.083	0.027	0.067	0.059	0.060	0.134	0.064	1	0.082
C20	0.381	0.013	0.003	0.007	0.026	0.014	0.009	0.011	0.007	0.017	0.020	0.019	0.007	0.034	0.019	0.619	0.061	0.055	0.082	1

Table7. Results (Macro Averaged F_1) on the limited reuters 21578.

Taille de profils	Résultats
K=100	0.870
K=200	0.861
K=300	0.865
K=400	0.867
K=500	0.868
K=600	0.867
K=700	0.869
K=800	0.869

3.4. Adaptation of the Size of the Profiles According to the Corpus of Training

Considering the preceding results, we can affirm that there is a great dependence between the corpus of training and the size of the profiles. Indeed, for a corpus of training that suffers from a no equitable distribution of the categories, the ideal profile size changes from one category to another.

From this point, we carried out an additional experiment. The idea of this experiment consists in varying the size of the profile from one category to another according to the corpus of training. For each category, plus the size of training set is signify cant plus the size of the best profile is large.

Table 8 presents the results of this experiment on the Reuters corpus, for N=5.

The results presented in table 8, show that the best performances were obtained with profiles of varied size.

3.5. The Influence of the Measure of Similarity

In this section, it is a question of comparing various measurements of similarity in order to study the influence of the measurement of similarity on the performances of a categorization.

In our work, we evaluated the approach by using the two distances described in section 2.2.2.

Table 9 presents the results of the categorization using these two distances.

Table 8. The results of the variation of the profile size on reuters corpus.

	Taille de profil variée		Taille de profil constante	
	Taille	Résultat	k=50	k=350
Acquisition	k=350	0.953	0.801	0.952
Corn	k=50	0.497	0.478	0.463
Crude	k=350	0.773	0.670	0.773
Earn	k=350	0.978	0.872	0.978
Grain	k=50	0.436	0.344	0.446
Interest	k=350	0.735	0.599	0.730
Moneyfx	k=350	0.749	0.534	0.747
Ship	k=350	0.673	0.521	0.664
Trade	k=350	0.797	0.738	0.800
Wheat	k=50	0.524	0.503	0.503
MacroAveraged F_1	**0.711**		0.606	0.706

Table 9. Results (MacroAveraged F1) of different similarity measurements.

	COSINUS		KULLBAK&LIEBLER	
	Reuters	20News	Reuters	20News
K=100	0.704	0.677	0.642	0.350
K=200	0.702	0.699	0.635	0.668
K=300	0.707	0.707	0.638	0.589
K=400	0.704	0.714	0.632	0.584
K=500	0.703	0.716	0.631	0.576
K=600	0.702	0.717	0.630	0.578
K=700	0.701	0.717	0.627	0.577
K=800	0.701	0.717	0.624	0.569

The results presented in table 9 shows that the best performances were obtained with the COSINE distance.

3.6. Comparison Between the "N-Grams" Representation and other Representations

In this section, we will compare the N-grams representation with the two types of representation most used in this field: the representation based on the words and the representation based on the stemmes. This comparison will be valid only within the framework of our approach.

The use of these two types of representation (words and stemmes) requires carrying out pre-treatment on the data. The textual documents must be cleaned as much as possible from the useless information that they contain so that the operation of categorization is the most effective possible, because the information not withdrawn is as relevant as it may be. Indeed in textual documents many words bring little (see no) information on the document concerns. The algorithms known as of "stopwords" deal with eliminating them. Another pre-treatment named "stemming" also makes it possible to simplify the texts while increasing their informative characters.

a. *The "stop words"*

The "stop words" are the words which have only a low semantic importance and which are often very frequent. Their elimination, during the pre-processing of the document makes it possible thereafter to gain much time and effectiveness when modelling and analysing the document.

Linguists have drawn up a list of these words for most languages.

b. *The "Stemming"*

The method presented by [13] makes it possible to gather the words resulting from the same root. That makes it possible for the processes of categorization to imitate what a human being naturally does when reading a text that contains words with common roots: if for example it reads the words "walking", "Walker" and "walk", it will deduce naturally that this document strongly evokes the topic of walk. Whereas, for a document "not stemmized", a topic by word can be deduced.

This pre-processing consists of a succession of rules that exploit the way in which are formed the words present in the vocabulary of a language in order to determine the common roots. The Porter algorithm is adapted to the language of Shakespeare, but following the success that it gained, it was adapted to other languages.

Our experimentation consists in testing our approach by using the words and stemmes in the place of the n-grams. Each category will be represented by a profile that will contain the words or the stemmes that characterize it compared to the other categories. The results of this experiment on the Reuters corpus are transcribed in Table 10.

The results presented in table 10 show that the N-grams representation is more powerful in text categorization (for our approach) compared to the other representations.

These results confirm the advantages of use of the N-grams as a representation technique.

Table 10. Comparing the three types of representation on the corpus reuters.

	5-grams		Mots		Stemmes	
	Reuters	News	Reuters	News	Reuters	News
K=100	0.704	0.677	0.637	0.643	0.640	0.651
K=200	0.702	0.699	0.646	0.659	0.642	0.670
K=300	**0.707**	0.707	**0.649**	0.665	0.655	0.677
K=400	0.704	0.714	0.646	0.666	0.656	0.681
K=500	0.703	0.716	0.643	0.666	0.657	0.682
K=600	0.702	**0.717**	0.643	**0.667**	0.658	**0.683**
K=700	0.701	0.717	0.646	0.667	0.659	0.683
K=800	0.701	0.717	0.646	0.667	**0.660**	0.683

4. Conclusions

In this article, we presented the approach suggested with all its stages, an approach that benefits from the use of the n-grams method to represent the data, and of the multivariate χ_2 as a method for categories profiles construction.

We analyzed the results of this approach on the two corpora most widely used in the field of the text

categorization: the Reuters 21578 corpus and the 20Newsgroups corpus.

The experiments carried out led to the following observations:

1- The choice of the value N influences on the results of the approach. Indeed, the experiments showed that the quint-grams are ideal for the two corpuses. This value can change for other corpus.
2- To combine the n-grams does not make any improvement in the performances; the results are quasi similar to the best N-grams without combination. On this point, we find the same conclusions of many authors in particular in [3, 6, 10]
3- The ideal size of profiles changes from one corpus to another and one category to another. Indeed, for the categories whose training sets are rather significant, the performances improve by increasing the size of the profiles. On the other hand for the categories whose training sets are less significant, we note no improvement by increasing the size of the profiles.
4- The adaptation of the size of the profiles to the training corpus size gave better results compared to the approach based on a constant profile size.
5- The quality of the training corpus influences on the results of the approach. Indeed, more the corpus of training is rich and equitably distributed, more the performances improve.
6- The representation based on the n-grams is more powerful than the other representations (words and stemmes).

References

[1] Bigi B., De Mori R., El-Bèze M. and Spriet T. "A Fuzzy Decision Strategy for Topic Identification and Dynamic Selection of Language Models," *Signal Processing Journal*, vol. 6, no. 6, pp. 1085- 1097, 2000.

[2] Carpinto C., De Mori R., Romano G., and Bigi B., "An information Theoretic Approach to Automatic Query Expansion," *ACM Transactions on Information Systems*, vol. 19, no. 1, pp. 1-27, 2001.

[3] Cavnar W. and Trenkl J., "N-Gram Based Text Categorization," *Symposium on Document Analysis and Information Retrieval*, Las Vegas, 1994.

[4] Dagan I., Lee L., and Pereira F., "Similarity Based Models of Word Co-Occurrence Probabilities," *Machine Learning*, vol. 34, no. 1-3, pp. 43-69, 1999.

[5] Debole F. and Sebastiani F., "An Analysis of the Relative Hardness of Reuters-21578 Subsets," *Technical Report 2003-TR-49*, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, Submitted for publication, 2003.

[6] Fürnkranz J., "A Study Using n-gram Features for Text Categorization," *Technical Report OEFAI-TR-98-30*, Austrian Research Institute for Artificial Intelligence, Austria, 1998.

[7] Grefenstette G., "Comparing Two Language Identification Schemes," *in Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95)*, Italy, 1995.

[8] Guo G., Wang H., and Bell D., "A K-nn Model-Based Approach and its Application in Text Categorization, Computational Linguistics, and Intelligent Text Processing," *in Proceedings of the 5th International Conference, CICLin,* Korea, vol. 2945, pp. 559-570, 2004.

[9] Hofmann T., "A Probabilistic Approach for Mapping Large Document Collections," *Journal for Intelligent Data Analysis,* vol. 4, no. 2, pp. 149-164, 2000.

[10] Lelu A. and Hallab M., *Consultation "floue" de grandes listes de formes lexicales simples et composées : un outil préparatoire pour l'analyse de grands corpus textuels*, vol. 1, pp. 317–324, Lausanne, 2000.

[11] Miller E., Shen D., Liu J., and Nicholas C., "Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System," *Journal of Digital Information*, vol. 1, no. 21, pp. 1-25, 1999.

[12] Peng X. and Choi B., "Document Classifications Based on Word Semantic Hierarchies," *in Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, vol. 1, pp. 362-367, 2005.

[13] Porter M. F., "An Algorithm for Suffix Stripping," *Program*, pp. 130–137, 1980.

[14] Sahami M., "Using Machine Learning to Improve Information Access," PhD Thesis, Computer Science Department, Stanford University, 1999.

[15] Sebastiani F., "Machine Learning in Automated Text Categorization,"*ACM Computing Surveys*, vol. 34, pp. 1-47, 2002.

[16] Zhou S. and Guan J., "Chinese Documents Classification Based on N-grams," *in Proceedings 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico, vol. 2276, pp. 405-414, 2002.

Abdellatif Rahmoun received his BSc degree in electrical engineering, University of Science and Engineering of Oran, Algeria, his Master degree in electrical engineering and computer science from Oregon State University, USA, and his PhD degree in computer engineering, Algeria. Currently, he is a lecturer in Computer Science Department, Faculty of Planning and Management, King Faisal University, Kingdom of Saudi Arabia. His areas of interest include fuzzy logic, genetic algorithms and genetic programming, neural networks and applications, designing ga-based neuro fuzzy systems, decision support systems, AI applications, e-learning, electronic commerce and electronic business and fractal image compression using genetic tools

Zakaria Elberrichi received his Master degree in computer science from the California State University in addition to PGCert in higher education. Currently, he is a lecturer in computer science and a researcher at Evolutionary Engineering and Distributed Information Systems Laboratory, EEDIS at the university Djillali Liabes, Sidi-belabbes, Algeria. He has more than 17 years of experience in teaching both BSc and MSc levels in computer science and planning and leading data mining related projects. The last one called “new methodologies for knowledge acquisition”. He supervises five master students in e-learning, text mining, web services, and workflow.