N-grams in Texts Categorization

Zakaria Elberrichi & Badr Aljohar*

Computer Science Department, College of Engineering,
University of Sidi Bel-Abbes, Algeria
*Computer Science & IT college, King Faisal University,
Al-Hasa, Saudi Arabia

Abstract:

This paper deals with automatic classification of documents; this is performed by a supervised classification since it operates on a set of preset classes. The suggested approach is original since it is based on a vector representation of the documents centred not on the words but on the n-grams of characters for n varying from 2 to 5.

Considering the significant number of the n-grams generated for each class, we used in our work the law of χ^2 to reduce the number of the characteristic n-grams of each class. The weighting of the vectors was done by using the measurement of the TFIDF, and for the calculation of the distance between two vectors, we used the method of the Cosine. The experiments were done on two well-known corpora in the community of categorization, the Reuter 21578 and the 20Newsgroups. Evaluation of the approach was performed by using a function combining both precision and recall.

The results obtained show that the technique of the n-grams is very effective in the field of the categorization of texts.

Key words: Text categorization, n-grams, the law of χ^2 , method of the Cosine, TFIDF, Reuters21578, 20 Newsgroups.

Introduction :

Automatic Text categorization is gaining popularity with the growing interest and usage of text data available as well on the world wide web as within enterprises and because, the manual realization of this classification is extremely expensive in term of time because of the appalling growth of the number of the numerical documents available. We distinguish in the field from automatic classification two types of approaches supervised classification and not supervised classification. These two methods differ on the way in which the classes are generated. In the case of not supervised classification, the groups of documents (categories) are generated automatically by the machine, while they are, in the supervised approach, defined by an expert. In this last case, it is interesting to represent the documents and the classes using the same formalism and the one used generally is a vector space.

In this article, we will be interested in categorization, i.e. supervised classification and more particularly to show the influence of the n-grams method of presentation of the documents on the results of the latter.

The article is organized in the following way. In section 2, we present problems of the categorization of texts. Section 3 presents the approach suggested with all the stages. So as to show the effectiveness of our approach, we describe in the section 4 some experiments carried out on corpora of evaluation (Reuters21578, 20 news groups) as well as a reading of these results.

Categorization of texts :

Text categorization (T.C) is the process which consists in assigning one or more categories among a preset list to a document. The manual realization of this task is extremely expensive in term of time because, it is necessary to attentively read each document to be able to decide.

In other words, the categorization of text consists in seeking a working joint (model of prediction) between a whole of texts and a whole of categories (labels, classes), one of several well-known techniques in information retrieval. [MaN2004]

Sebastiani, F defines the T.C as being the process which consists in seeking a working joint between a set of texts and a set of categories (labels, classes). Formally, the categorization of text consists in associating a Boolean value to each pair (D_J C_I *in* $D \times C$ where D is the set of the texts and C is the set of the categories.

The value T (Truth) is then associated to the couple (D_J C_I if the text D_J belongs to the class C_I while the value F (False) will be associated to it in the contrary case. The goal of the categorization of text is thus to build a procedure (model, classifier) F: $D \times C!$ $\{V, F\}$ which associates one or more labels (categories) a document D_J such as the decision given by this procedure " *coincides as much as possible* " with the function E: $D \times C!$ $\{V, F\}$ the true function which turns over for each vector D_J a value C [Seb2002].

The design of a system of text categorization comprises several stages. Firstly it is necessary to choose a model of representation of the documents and categories that is exploitable by the machine, the model most usually used in this field is the vectorial model. The second stage is that of the training in which we try to find a mathematical model able to represent, for then comparing the semantics of the texts. All the methods of training resulting from the artificial leaning (AL) community can be applied to text categorization applications. The third stage is that of the classification [YMi2005] in which we assign a text to a category based on the model found in the preceding stage and which is the stage of learning]. A last stage is necessary to evaluate the performances of the system. To measure the accuracy of prediction of the system, various measures are used in the continuation of this article.

N-Grams :

A n-gram is a sequence of n consecutive characters. The set of the n-grams that can be generated is the results obtained by moving a window of n boxes on the body of text. This displacement is done by stages, a stage corresponds to a character. In our work, we used several lengths for the n-grams (n=1,2..., 5). we replaced the space character by the character " - ". For example, the text " you and you " gives the following n-grams:

- Bi-grams: yo,ou,u-,-a,an,nd,d-,…etc.
- Tri-grams:you,ou-,u-a,-an,and,nd-,... etc.
- Quadri-grams : you-,ou-a,u-an,-and,... etc.

The n-grams have several advantages:

- *Automatic capture of the roots of the most frequent words.*
- *Independence towards the document language*. Contrary to other techniques which require the use of specific dictionaries ((feminine-masculine; singular-plural; conjugations; etc.) for each language. Moreover, with the n-grams, we do not need preliminary segmentation of the text in words; this is interesting for the processing of languages in which the borders between words are not strongly marked, like Chinese for example.
- *Tolerances with the spelling mistakes and the deformations* For example, it is possible that the word " *chapter* " is written like " *clapter* ". A system based on the words will have difficulties to recognize the word " *chapter* " since the word is badly spelled. On the other hand, a system based on the n-grams is able to take into account the others n-grams (parts) like " *apte* ", " *pter* ", etc

Considering the importance of these advantages, the n-grams are used in several fields.

4. the law of χ^2

If the n-grams offer several advantages, the number of the generated n-grams disadvantage their uses in the field of categorization of texts. To minimize this problem, we used the law of χ^2 like a means of reduction of the number of generated n-grams [Fern2000], [Bekk2002]

The statistics of χ^2 measure the variation with independence between a descriptor *T* and a topic *C.* There exist two alternatives for this measurement, the first measures independence in term of absence/presence of a descriptor in the documents associated with a topic; Calculation requires to build the table of contingency (2×2) for each descriptor *T* of the corpus and each class *C* (table 1).

Table (1)

Table of contingency for descriptors of the corpus.

	term t_k present	term t_k absent	
term c_i present	*a*	*c*	*a+c*
term c_i absent	*b*	*d*	*b+d*
	a+b	*c+d*	*N=a+b+c+d*

The statistics of χ^2 can be put in the form

$$\chi^2_{uni}(t_k c_i) = \frac{N(ad-cb)^2}{(a+c)(b+d)(a+d)(c+d)}$$

This first measurement called *univariate* is used for the selection of the descriptors in [Zhang01], [Cav94] , and [Yan97].

The second alternative called multivariate, is a supervised method allowing the selection of terms while taking into account, not only their frequencies in each class, but also the interaction of the terms between-them and the interactions terms/classes [He2000] , [Jal2002].

The idea consists in using the contributions of the cells (*t, c*) to the χ^2 associated to the global cross table, where NR $_{ki}$ is the number of times where the term t $_K$ is present in the documents of the class c $_I$ The stages are described in algorithm 1.

In all our experiments, we chose to use multivariate χ^2 method because :

- It is supervised since it is based on the information brought by the categories.
- It is multivariate because it evaluates on an overall basis the role of a term compared to the others.
- It takes account of the interaction terms/classes because it makes it possible to choose, for each category, the terms which contribute more to their discrimination.

Algorithm:

```
Input:  $C = \{c_1\ C_2\ C_{3...,}\ C_C\}$ / / the list of the categories preset.
        Train corpus//texts for learning for each category.
        n.// size of the window of the n-grams.
        k / / size of the profiles.
        Test corpus / / texts for testing for each category;

Begin
        For(i=0;i<c;i++)
            { generate the n-grams using the texts for learning of category i;
            C ompute the number of frequency of each n-grams; }
        Build table $N_{ij}$ of occurrence of the n-grams j in category i;

        Compute the value $\chi^2$ representing independence between the n-grams j and category i;
        Sort the table in a decreasing order ;
        For(i=0;i<c;i++)
        { Profil[i]=Vector containing K first n-grams table $\chi^2_{i.;}$
        }// End of the learning phase
        For(i=0;i<c;i++)
        { For each $d_i \in$ Test corpus do
        { generate the n-grams and build the vector of the document $d_i$
            Compute the distance cosine between the document vector and the
            profiles of the categories;
            associate the document $d_i$ to the category to which its profiles is the
            closest. }
                    }
End
```

- Algorithm of categorization by using the n-grams -

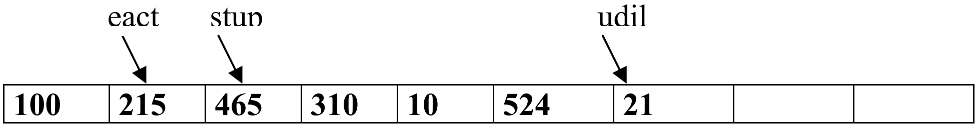
An approach of categorization based on the n-grams

The majority of the approaches of classification are centred on pre-linguistics processing such as the deletion of the blank words, the lemmatisation and the stemming. These pre-processing require a preliminary knowledge of the language in which are transcribed the documents. In other words, these approaches suffer from a strong dependency towards the language of the documents, which limit their applications.

The approach studied in this article is an approach based on the n-grams, an approach which has the advantage of being independent towards the language of the documents, and operates without any linguistic pre-processing.
[Yang,99], [Yang97].

5.1- Generation of the n-grams

At this stage, it is a matter of representing each category in the form of a vector whose each descriptor represents a n-gram, with each n-gram in the vector we associate his number of occurrence in the category.



100	**215**	**465**	**310**	**10**	**524**	**21**		

The example of the figure presents a vector of a category in which the quadri-gram " eact ", " stup ", " udi " are repeated on the latter with respectively 100,215,524 occurrences.

Selection of the characteristic n-grams

In this second stage, it is a matter of generating a profile for each category, A profile of a category contains all the n-grams which characterizes it, this compared to the other categories. There are several methods to discriminate the classes, we chose to use the law of χ^2 multivariate to discriminate the categories.

The stages of the algorithm are to detail as follows:
Firstly, a matrix χ^2 [i,j] occurrences of the n-grams i in the category j should be built. Thereafter it is necessary to calculate the value χ^2[i,j] which represents the independence between the n-grams *i* and the category *j* then sort the table in the decreasing order.

$$\chi_{ij}^{2} = \frac{(N_{ij} - \frac{N_{i} X N_{j}}{N})^{2}}{\frac{N_{i} X N_{j}}{N}} .sign(N_{ij} - \frac{N_{i} X N_{j}}{N})$$

With:

N_{ij} *Number of* occurrence of the n-grams i *in* the category j.

$N_{i.}$ *Number of* occurrence of the n-grams i *in* all the learning corpus .

N_{j} *Number of* occurrence of all the n-grams in the category *j.*

N *Number of* occurrence of all the n-grams in all the learning corpus.

The profiles of each category will thus contain K first n-grams. The influence of the value of the parameter K is studied in the experimentation part.

Classification:

In this stage, it is a question of calculating a distance between the profile of a document to be categorized and the profiles of the categories. To do this, it is first of all necessary to balance the n-grams constituting the profiles of the categories [Bekk2002]. There exist many measurements to balance the vectors, the most used is measurement are TFIDFs.

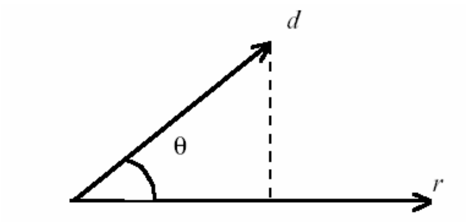
$$TFIDF(w,d) = TF_{w,d}.IDF_{w,d} = TF_{w,d}\{(\log_{2} \frac{N}{DF_{w}}) - 1\}$$

With:

TF $_{w,d}$ Number of occurrence of the n-grams *W* in the profile *d.*
DF $_{W}$ Number of profiles containing the n-grams *w.*
There are several methods to calculate the distance between two vectors, the most used in this field and the method of the Cosine.

$$Co\ \sin\ e(d,r) = \frac{\sum_{w \in d \mathrm{O} r} TFIDF_{w,d} . TFIDF_{w,r}}{\sqrt{(\sum_{w \in d} TFIDF^{2}_{w,d}) . \sum_{w \in r} TFIDF^{2}_{w,r})}}$$

with: *w* a n-grams, *d* the document to be categorized, *r the* profile of a category, *TFIDF* $_{w,d}$ the weight of the n-grams w in the document d and *TFIDF* $_{w,r}$ that of *w* in category r.
The figure presents a geometrical view of this distance.



Experiments :
Data preparation :

For the experiments, we used the two most used corpora in this field : the Reuters-21578 and the 20Newsgroup corpora.
- the corpus Reuters –21578:

The Reuters-21578 corpus is a set of financial dispatches emitted during the year 1987 by the Reuters agency, in English, and available free on the Web [1]. This corpus is very often used for evaluation in publications, as in [Sha98] comparing their AdaBoost *algorithm* with the formula of Rocchio, or in [Joa98], [Seb2002], and [Dum98] evaluating the performances of the machines with vectors supports. [Yan99] also used this corpus to compare various algorithms (machines with vectors supports, networks of neurons, decision trees, networks Bayesians). This corpus is composed of a certain number of categories each one comprising a learning set and a testing set.

In our experiments, we used only the 10 categories the most represented within the version" ModApte " of this corpus. Table 2 shows the distribution of the documents on the two set (learning set and test set).

Table (2)

Word distribution of the document Reuters-21578

	Category	Learning set	Test set
1	earn	2877	1087
2	acq	1650	719
3	money-fx	538	179
4	grain	433	149
5	crude	389	189
6	trade	369	118
7	interest	347	131
8	wheat	212	71
9	ship	197	89
10	corn	182	56

- the corpus 20Newsgroup:

It is a corpus developed in CMU which consists of 20 000 electronic messages of 20 newsgroup (1000 by group). Within the framework of our experiments, we took only 10 categories out of the 20 categories present. The 1000 documents of each category are divided into a learning set and a test set. Table 3 shows the distribution of the documents.

Table (3)

Word distribution of the document 20Newsgroups.

	Category	Learning set	Test set
1	alt.atheism	666	350
2	misc.forsale	677	333
3	rec.autos	670	333
4	rec.motorcycles	667	333
5	rec.sport.baseball	666	333
6	sci.electronics	675	333
7	sci.med	667	333
8	soc.religion.christian	664	333
9	talk.politics.mideast	670	333
10	sci.crypt	667	333

Measurements of performance :

Currently, deciding what measure decide if a categorization is correct or not is in itself an issu. The evaluation of a categorization is thus made empirically on two criteria which are most significant, the effectiveness which measures the calculating time and the memory size, and the accuracy of prediction which measures if the categorization carried out is correct or not. In our experiments, the accuracy of prediction is the criterion which imports us more. Measurement the most used to measure the accuracy of a prediction is the couple precision and recall developed initially for IR (Information Retrieval).

[1] http://www.daviddlewis.com/resources/testcollections/Reuters21578 /
[1] http://www.ai.mit.edu / ~~jrennie/20Newsgroups

Définition1: Recall and Precision:

$$\pi_i = \frac{VP_i}{VP_i + FP_i}$$

$$\rho_i = \frac{VP_i}{VP_i + FN_i}$$

With VP_I FP_I FN_I respectively defining the well classified texts, the texts assigned by error as well as the texts omitted by the classifier (for a category i *)* .

To evaluate a categorization, one cannot measure only the recall or the precision because these two measurements do not have any significance one without the other.
To take into account at the same time the recall and the precision, the formula *F_βs used most of the time.*
Definition 2: F_β

$$F_\beta = \frac{(\beta^2 + 1).\pi_i.\rho_i}{\beta^2 \pi_i + \rho_i}$$

When β> 1, the precision plays a more significant role than the recall and measurement F_β support the classifiers with a good precision. Inversely, when β< 1, the recall is more preferred. When there is not a priori a value β= 1 is used.

Results:

In the whole of our experiments, we tried to evaluate the method on the two corpora Reuters21578 and 20 newsgroup while showing the influence of several parameters on the results.

Table (4) shows the results obtained for n=2,3,4,5 and n=2+3+4+5 while taking for each value of N, various values of K

Table (4)

N	*2*		*3*		*4*		*5*		*2+3+4+5*	
Corpus	*Reuters*	*news*	*Reuters*	*news*	*Reuters*	*news*	*Reuters*	*news*	*Reuters*	*news*
K=100	0.447	0.386	0.649	0.705	0.698	0.769	0.709	0.786	0.689	0.747
K=200	0.451	0.394	0.648	0.726	0.705	0.791	0.707	0.805	0.701	0.779
K=400	0.451	0.394	0.652	0.734	0.702	0.817	0.704	0.824	0.706	0.804
K=600	0.451	0.394	0.654	0.736	0.699	0.825	0.703	0.830	0.706	0.821
K=800	0.451	0.394	0.653	0.736	0.698	0.829	0.703	0.835	0.706	0.830
K=1200	0.451	0.394	0.653	0.736	0.698	0.829	0.703	0.835	0.698	0.834

The results presented in the table affirm several prepositions. Indeed, we can note that:

1. for the value of N:
   - 1.1- The best performances were obtained with the quint-grams (n=5).
   - 1.2- While decreasing the size of the window (the value of n) the performances deteriorate more and more.
   - 1.3- The bi-grams gives the worst results ,and are the closest to the random.
   - 1.4- Combining the n-grams (2+3+4+5) didn't bring a considerable improvement.

2. for the value of K:
   - 2.1- By increasing the value of K(vector size), the performances increase, then is stabilized for a value ranging between 600 and 800.

3. The approach is more powerful on the corpus 20Newsgoups than on the corpus Reuters. On this point, we find the conclusions of many authors on the disadvantages of the corpus Reuters. Because the corpus Reuters comprises some very close categories which makes their discrimination more difficult. To show that, we calculated the Cosine distance between the profiles of the categories for the two corpora. The results show that the grain, corn and wheat categories are very close in the corpus Reuters, while the ten categories of the corpus 20Newsgroup are well discriminated.

Conclusion:

The approach suggested in this article is different from several ones existing in the literature, it uses a vectorial representation based on the n-grams; an approach which has the advantage of being independent of the language in which the documents are transcribed. We made experiments on two corpora that best illustrate the interest of this approach in improving performance in text categorization. Further works may emphasize on how to take advantage of this method of text categorization in information Retrieval and data mining [Benn2005].

References :

1. Ron Bekkerman, (2002), "Distributional Clustering of Words for Text Categorization", M.Sc. Thesis, August the 8th, , SIGIR'01, USA 2001.
2. Bennett P., J. Carbonell, (2005),"Beyond Bag-of-Words Workshop", at the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 63-70, Salvador, Brazil, August 15 - August 19,. ACM Press.
3. Cavnar W and Trenkl J . (1994), *NR* - Gram Based Text Categorization. In Symposium one Document Analysis and Retrieval Information Las Vegas, , pp140-148.
1. 4. Dumais S., J Platt, D. Heckerman, Mr. Sahami, (1998), "Inductive Learning: Algorithms and Representations for Text Categorization", Proceedings of the seventh International Conference one Information and Knowledge Management (CIKM' 98) 148-155.
4. Fernanda M. , (2000), 'Statistical Phrases in Automated Text Categorization", B4-007, 2000.
5. He, J, Tan, A.H., Tan and, C.L. (2000), "With comparative study one Chinese text categorization methods", in PRICAI Workshop one Text and Web Mining.
6. Jalam, R. and Chauchat, J.H. (2002), "Why the N-grams make it possible to classify texts? Search for relevant keywords using N grams characteristics", in Morin, A. and Sébillot, P., editors, 6th International Conference one Textual Dated Statistical Analysis, volume 1.
7. Joachims, T.(1998), "Text categorization with support vector machines: Depending Learning with many features", in proceedings of the European conference one Machine learning 1998.
8. McNamee P.,J.Mayfield, (2004), "Character N-Gram Tokenization for European Language Text Retrieval", in Information retrieval 7(1-2),pp 73-97.
9. Miao Y., (2005) ,"Document Clustering using character n-grams: A comparative evaluation with term-based and word-based clustering", technical report, Faculty of Computer Science, Dalhousie University,. www.cs.dal.ca/research/rechreports/2005/
10. Schapire, R. E., Freund, Y., Bartlett, P., Lee, W S. (1998), "Booster pump the Margin: new explanation for the effectiveness of Voting has".
11. Sebastiani, F. (2002), "Machine learning in automated text categorization", Acm Computing Surveys., pp45-49.
12. Sebastiani, F., (2003), " Machine learning in automated Text Categorization". ACM Computing Surveys, 34(1):1-47.

13. Yang, Y. and Pedersen, J O (1997). With comparative study one feature selection in text categorization. In Fisher, D. H., editor, Proceedings of ICML-97, 14th International Conference one Learning Machine

2. 15. Yang Y., Liu X, (1999), "With Re-examination of text categorization methods", Proceedings of the 22 Nd Annual International Conference one Research and Development in Information Retrieval (SIGIR ' 99) 42-49.

16. Zhang T., Oles F.J., (2001), "Text Categorization Based on Regularized Linear Classification Methods, Information Retrieval", vol.4.

تصنيف أوتوماتيكي للنصوص باستعمال تقنية ال"ن-غرامس"

زكريا البريشي و بدر عبداللطيف الجوهر*

قسم الحاسب الآلي،كلية الهندسة، جامعة سيدي بلعباس، الجزائر

*كلية علوم الحاسب و تقنية المعلومات، جامعة الملك فيصل،

الأحساء ، المملكة العربية السعودية

الملخص :

هذه الورقة تتناول التصنيف الأوتوماتيكي للنصوص والذي يعتمد على الإرشاد في اختيار التصنيف الملائم بناءً على عدد من جزئيات الكلمات المحددة مسبقاً. الطريقة المقترحة في هذه الورقة تعتمد على التمثيل الشعاعي للوثيقة أو النص بناءً على جزئيات الكلمات (ن غرامس) وليس على الكلمات. وقد استخدم المعامل من ٢ حرف إلى ٥ حروف لكل صنف ليتم احتساب جزئيات كل صنف بناءً على عدد مرات تكرار كل جزئية في الوثيقة أو النص. يتم بعدها إنتاج جزئيات كل صنف ومن ثم تقلص عدد هذه الجزئيات باستخدام القانون الإحصائي (كاي ٢). جميع التصنيفات المرشحة تعطى أوزان نسبية باستخدام مقياس (تي أف آي دي أف) ومن ثم يحتسب الفارق بين كل صنف وآخر باستخدام طريقة (الكوساين).

أخيراً تضمنت الورقة نتائج تجارب أجريت على مدونات تحتوي على نصوص جمعت من وكالة رويترز ونصوص جمعت من مجموعات اخبارية ، لتقييم مدى قوة وفاعلية الطريقة المقترحة. وقد استخدم في التقييم دالة تجمع بين الدقة في التصنيف وإمكانية إعادة الاستعمال ، حيث أظهرت النتائج أن الطريقة المقترحة حققت أداء جيد في تصنيف النصوص.

الكلمات الأساسية: تصنيف النصوص، "ن- غرامس"، قانون ال"كاي χ_2"، *طريقة "كوساين"*، *TFIDF*، "روينر٢١٥٧٨"، مجموعة " نيوز ٢٠".