Arabic documents classification using N-gram

Majed SANAN
Paris 8 university
Paris – France
Sinane80@hotmail.com

Mahmoud RAMMAL
Lebanese university
Beirut-Lebanon
mrammal@ul.edu.lb

Khaldoun ZREIK
Paris 8 university
Paris - France
zreik@univ-paris8

Abstract- This paper is aimed to study and explain the useful classification of Arabic text using N-gram as indexing method. The Lebanese official journal documents can be categorized into several classes. Supposing that we know the class(es) of some documents(called learning texts), this can help us to know the candidate words of each class by segmenting the documents.

In our approach we have used N-gram as a representation method in the case of 3 characters.

In this paper, we study the effect of using n-grams (sequences of words of length n) for text categorization.

Keywords: Classification, learning method, N-gram, Arabic, Categorization.

I. INTRODUCTION

The rapid growth of the Internet has increased the number of online documents available. This has led to the development of automated text and document classification systems that are capable of automatically organizing and classifying documents. Text . classification (or categorization) is the process of structuring a set of documents according to a group structure that is known in advance. There are several different methods for text classification, including statistical-based algorithms, Bayesian classification, distance-based algorithms, knearest neighbors, decision tree-based methods….

Text classification techniques are used in many applications, including e-mail filtering, mail routing, spam filtering, news monitoring, sorting through digitized paper archives, automated indexing of scientific articles, classification of news stories and searching for interesting information on the web

The majority of these systems is designed to handle documents written in non Arabic language, Developing text classification systems for Arabic documents is a challenging task due to the complex and rich nature of the Arabic language. The Arabic language consists of 28 letters. The language is written from right to left. It has very complex morphology, and the majority of words have a tri-letter root. The rest have either a quadletter root, penta-letter root or hexa-letter root.

In our approach we will use only the similarity measures and compare the results in order to know the convenient measure in classification using N-grams. And because that classification is one method of text mining we will explain in the following paragraph the steps of text mining, then we will see the preprocessing and indexing of texts before to be classified. At the next paragraph, we will explain the different similarity measures that we will use in our approach, and then the effectiveness measure used to calculate the precision and recall of each class. At paragraph 6 we will explain our approach and experiments, and finally we will see the conclusion and future approaches.

II. TEXT MINING

A. Definition

Text mining is defined [1] as the non trivial extraction of implicit, previously unknown, and potentially useful information from (large amount of) textual data.

Text Mining is the process of applying automatic methods to analyze and structure textual data in order to create useable knowledge from previously unstructured information.

B. Text mining methods

There are many text mining applications or methods. Four of these methods are the following:

- Information Retrieval

This method consists of indexing and retrieval of textual documents

- Information Extraction

It means extraction of partial knowledge in the text

- Web Mining

It consists on indexing and retrieval of textual documents and extraction of partial knowledge using the web

- Classification

Given: a collection of labelled documents (*training set*), the goal is to find a model for the class as a function of the values of the features

C. *Text mining steps*

These steps concern principally the manner in which a text is represented (or structured), the choice of predicted algorithm to use, and then how to evaluate the obtained results to guarantee a good generalization of the model applied.

1) *Representation of the information*

In this step we have to segment the unstructured information and put the units segmented into a table. But we have to choose the descriptors (important terms in documents) which can be chosen as words , lemmas, stemmas, or n-grams (characters or words or phrases).

And finally in some cases we have to think how to reduce the dimension of this textual space.

2) *Automatic categorization-classification of documents*

This is the second step, the text categorization can be defined as the process that permit to associate a category (ies) or class(es) to a text (or document), in function of information contained in this text .

This association is very long and expensive then we think about the automation of this process.

The functional link between a class and a document, that is called 'a prediction model', is estimated by a machine learning method.

The categorization of documents comports a choice of a learning technique (or classifier). The main classifiers used are the following:

- Discriminated factorial analysis [2].
- Neuronal network [3].
- K-Neighbors [4].
- Decision Tree [5].
- Bayesian network [6].

...

3) *Validation method*

In this final step we have to evaluate the obtained results to guarantee a good generalization of the model applied.

III. TEXT PREPROCESSING AND INDEXING

All text documents went through a preprocessing stage. This was necessary due to the variations in the way text can be represented in Arabic. The preprocessing was performed for the documents to be classified and the training classes themselves. Preprocessing consisted of the following steps:

1) Convert text files to UTF-8 encoding.
2) Remove punctuation marks, diacritics, non letters, stop words. The definitions of thesewere obtained from the Khoja stemmer.
3) Replace initial أ إ ,أ with . ا
4) Replace final ى followed by ء with

A. *Spelling Normalization and Mapping*

Arabic orthography is highly variable. For instance, changing the letter YEH (ي) to ALEF MAKSURA (ى) at the end of a word is very common. (Not surprisingly, the shapes of the two letters are very similar.) Since variations of this kind usually result in an "invalid" word, in our experiments we detected such "errors" using a stemmer (the Buckwalter Stemmer) and restored the correct word ending.

A more problematic type of spelling variation is that certain glyphs combining HAMZA or MADDA with ALEF (e.g. إ ,أ and آ) are sometimes written as a plain ALEF (ا), possibly because of their similarity in appearance. Often, both the intended word and what is actually written are valid words.

This is much like confusing "résumé" with "resume" in English. Since both the intended word and the written form are correct words, it is impossible to correct the spellings without the use of context.

We explored two techniques to address the problem.

1) with normalization technique, we replace all occurrences of the diacritical ALEFs by the plain ALEF.

2) with the mapping technique, we map a word with the plain ALEF to a set of words that can potentially be written as that word by changing diacritical ALEFs to the plain ALEF. In this

absence of training data, we will assume that all the words in the set are equally probable.

Both techniques have pros and cons. The normalization technique is simple, but it increases ambiguity. The mapping technique, on the other hand, does not introduce additional ambiguity, but it is more complex.

B. Arabic Stemming

Arabic has a complex morphology. Most Arabic words (except some proper nouns and words borrowed from other languages) are derived from a root. A root usually consists of three letters. We can view a word as derived by first applying a pattern to a root to generate a stem and then attaching prefixes and suffixes to the stem to generate the word [7]. For this reason, an Arabic stemmer can be either root-based or stem-based.

C. Character N-grams

Broken plurals are very common in Arabic. There is no existing rule-based algorithm to reduce them to their singular forms, and it seems that it would be not be straight-forward to create such an algorithm. As such, broken plurals are not handled by current Arabic stemmers.

One technique to address this problem is to use character n-grams. Although broken plurals are not derived by attaching word affixes, many of the letters in broken plurals are the same as in the singular forms (though sometimes in a different order). If words

are divided into character n-grams, some of the n-grams from the singular and plural forms will probably match.

This technique can also handle words that have a stem but cannot be stemmed by a stemmer for various reasons. For example, the Buckwalter stemmer uses a list of valid stems to ensure the validity of the resulting stems. Although the list is quite large, it is still not complete. N-grams in this case provide a fallback where exact word match fails.

In previous work [8], experiments have been made with n-grams created from stems as well as n-grams from words. N-grams were created by applying a shifting window of *n* characters over a word or stem. If the word or stem has fewer than *n* characters, the whole word or stem was returned. The following table shows some results of these experiments. Two methods of creating n-grams were tried: from words and from stems. Retrieval scores in Table 1 show that stem-based n-grams are better than word-based n-grams for retrieval. The probable reason is that some of the word-based n-grams are prefixes or suffixes, which can cause false matches between documents and queries

	Bigrams	Trigrams	4-grams
Words	0.1461	0.2990	0.2900
Stems	0.1655	0.3365	0.3165

Table 1 *Retrieval results using n-grams*

However, character level n-gram models offer the following benefits and have been successfully used in many information retrieval problems:

1- Language independence and simplicity: Character level n-gram models are applicable to any language, and even non-language sequences such as music or gene sequences.

2- Robustness: Character level n-gram models are relatively insensitive to spelling variations and errors, particularly in comparison to word features.

3- Completeness: The vocabulary of character tokens is much smaller than any word vocabulary and normally is known in advance. Therefore, the sparse data problem is much less serious in character n-gram models of the same order.

IV. SIMILARITY AND MATCHING MEASURES IN VECTOR SPACE MODEL

In this paragraph we will explain some similarity and matching measures in vector space model often used in Information Retrieval, but we will use them in similarity between a document and a class.

Now that we have the document in a form that minimizes the information we need to consider when matching documents to classes we have to do some matching.

Under the vector and probabilistic models, the document is initially indexed in the same way as the classes.

*A. TF*ICF weight*

In fact I have used the TF*ICF and apply it to the class, then the query will be replaced by the document to be classified and the document will be replaced by the class, then I have defined the new weight TF*ICF.

In TF*ICF, ICF stands for inverse Class frequency and TF stands for term frequency. ("*" indicates multiplication.).

The *term frequency* in the given class is simply the number of times a given term appears in that class. This count is usually normalized to prevent a bias towards longer classes (which may have a higher term frequency regardless of the actual importance of that term in the class) to give a measure of the importance of the term t_i within the particular class.

$$\mathrm{tf_i} = \frac{n_i}{\sum_k n_k}$$

where n_i is the number of occurrences of the considered term, and the denominator is the number of occurrences of all terms.

The *inverse class frequency* is a measure of the general importance of the term (obtained by dividing the number of all classes by the number of classes containing the term, and then taking the logarithm of that quotient).

$$ICFi = \log \frac{|C|}{|\{c : ti \in c\}|}$$

with

- $|C|$: total number of classes in the corpus
- $|\{c\ ti \in c\}|$: number of classes where the term t_i appears (that is $n_i \neq 0$).

Then

TF ICF= TF*ICF

A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

ICF is defined as log (no of classes in the collection/no of classes containing this document in the collection). This reflects the fact that uncommon documents are more likely to useful in narrowing down the selection of class than very common documents. TF is defined as log (frequency of term in this class) This reflects the fact that if a keyword occurs multiple times in a class, that class is more likely to be relevant than a class where the keyword occurs just once.

B. *Dice's coefficient*

Now that the documents are both represented as vectors, the vector space model considers the similarity of them to be based on the angle between the two vectors in space. Up until this point (and with the probabilistic model), the vector has simply been a convenient mathematical model for storing a list of terms and their weights. The vector space model then makes the jump to processing them as if they were real geometrical vectors in a space with thousands of dimensions. Although this seems rather strange initially, it is based on an extension of a simple matching routine for non-weighted indexes. Consider non-weighted indexes for the above document and the sample class, these are basically a list of four words for the document and fourteen for the class. A rough measure of matching strength is the number of terms they have in common, in this case two. This doesn't take into account how large each class is and would have a tendency to match larger documents, so we could divide by the number of terms in total between the document and the class. This leads to *Dice's coefficient*:

$$m = 2 * \frac{|D \cap C|}{|D| + |C|}$$

where |*D*[[intersection]]*C*| is the number of terms common to the document and class, |*D*| is the number of terms in the document, |*C*| the number in the class and *m* the matching value - the fraction is doubled to give a maximum value, for matching a class with itself, of 1 instead of 0.5.

C. *Cosine Coefficient*

When considering weighted terms, like those we indexed, it is not possible to simply count the number of terms in common. Instead the vector space model multiplies the term weights together. For the vast majority of terms either the document or the class will have a zero weight, hence the resulting weight is zero. These individual new weights are then summed to give the top line of the matching algorithm. For a document D of N terms and for a class vector C, this leads to:

$$m = \frac{\sum_{i=1}^{N} D_i C_i}{\| D \| . \| C \|}$$

where $\|D\|$ is the length of the document (= total number of terms in document D), Di is the weight of term i in vector D, and N is the total number of individual terms (the dimensionality of C). In geometry this equation is used to calculate the cosine of the angle between the two vectors, hence this matching routine is known as the *cosine coefficient*.

Although quite simple to understand this approach has no sound bases in information theory there is no theoretical reason for this to be a good matching algorithm. The cosine coefficient does, however, perform well in practice, is reasonably easy to code and is used in many retrieval systems.

V. CLASSIFICATION EFFECTIVENESS (PRECISION AND RECALL)

Precision and recall are defined in [9] as follows:

$$\text{Precision} = \frac{CC}{TCF}$$

$$\text{Recall} = \frac{CC}{TC}$$

Where,

CC :number of correct categories(classes) found.

TCF : total number of categories found

TC : total number of correct categories

Ever since the 1960's information retrieval (IR) effectiveness is evaluated using the twin measures of *recall* and *precision* [10]

To combine the two measures (Precision and Recall) in a single value, the F-measure is often used. The F-measure reflects the relative importance of recall versus precision. When as much importance is granted to precision as it is to recall we have the F1-measure which is an estimation of the breakeven point where precision and recall meets if classifier parameters are tuned to balance precision and recall.

Then we can also use a single-number measure for the effectiveness as follows:

F1 = 2PR / (P+R) ... where F1 as a harmonic mean of precision and recall. [10]

For this study relevance has been defined conceptually as:

VI. OUR APPROACH

A. Corpus

We work on the meeting minutes of Lebanon parliament. Our Data base content all minutes from 1922 until 2005. The meetings are riche in different kind of information (economical, political, juridical, social, etc). In other ways, we have a data base for the official journal that content all laws and decrees. The texts of official journal are classified manually. Our work is to apply classification of official journal to parliament minutes in which Lebanese official journal documents form the main part. Then our corpus is the Lebanese official journal documents for year 2002 that form about 2667 classified (labeled) documents

B. Methodology

In our approach, we have chosen the N-gram method to represent information. And then to categorize the texts we will use the learning method in which we supposed that we have some categorized texts (learning texts) that we used to find the prediction method using the n-gram technique.

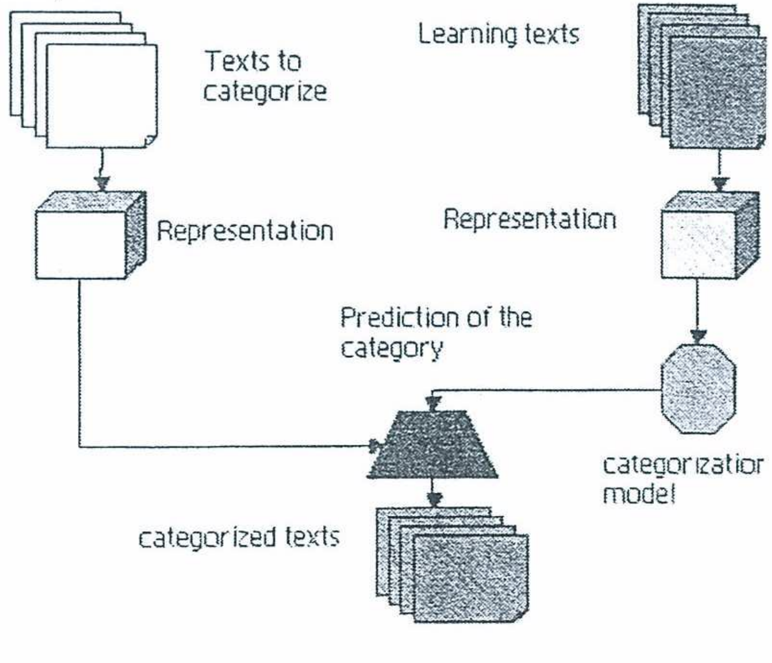


FIG. 1.1 : *Text categorization process*

In our experiment the learning texts will be the 2667 Lebanese official journals

(for the year 2002).

These documents are classified and each document belongs to one or multiple predefined classes.

We have two levels of classification, then each document belongs to a class of Level 1 and then to a subclass of a level 2.

In general we have three main classifications (Level 1):

1- Administrative classification.
2- Juridical classification.
3- Thematic classification.

And in each classification we have different classes, for example in administrative classification we have 137 classes. Table 2 shows some of these classes.

ID	Descriptior
1	مجلس الخدمة المدنية
3	المجلس العدلي
4	رئاسة مجلس الوزراء
5	المجلس الاسلامي العلوي
7	الهيئة العليا للاغاثة
9	مجلس الجنوب
10	المصرف الوطني للانماء الصناعي والسياحي
11	وزارة العدل
13	المديرية العامة للتنظيم المدني
14	المديرية العامة لرئاسة مجلس الوزراء
16	المديرية العامة للطرق والمباني
18	المديرية العامة لوزارة العدل
20	وزارة الخارجية والمغتربين
22	المديرية العامة للاحوال الشخصية
24	وزارة الزراعة
27	المديرية العامة للبريد

Table 2 *Administrative classes.*

Using these classified documents we will segment them by using N-gram method (three characters)and then segment the two level classes.

Then we will try to find the candidate words for each class (level 1 and level 2) and for each document using the vector space model.

After that we will apply the similarity and matching measures on documents and classes to classify automatically the pre classified documents.

Then we will conclude which is the convenient measure using the precision and recall parameters. Knowing the convenient measure we can then in the future use it to classify new documents.

C. Experimental Software

To segment the 2667 documents that form the corpus, using the n-gram method, I have made a program (using VB.net) that use the n-gram with 3,4 and 5 characters and give the result in a table (top 50).

D. Experience

Generating the N-gram profile consisted of the following steps:

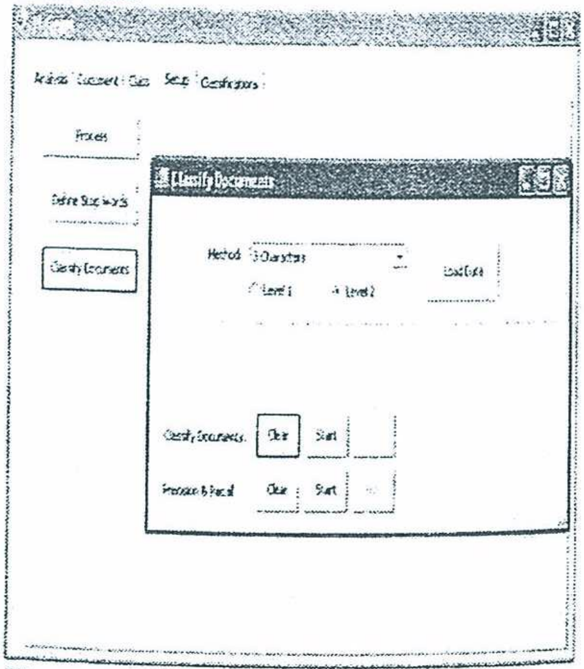
1) Split the text into tokens consisting only of letters. All digits are removed.

2) Compute all possible N-grams , for N=3 (Tri-grams)

3) Compute the frequency of occurrence of each N-gram.

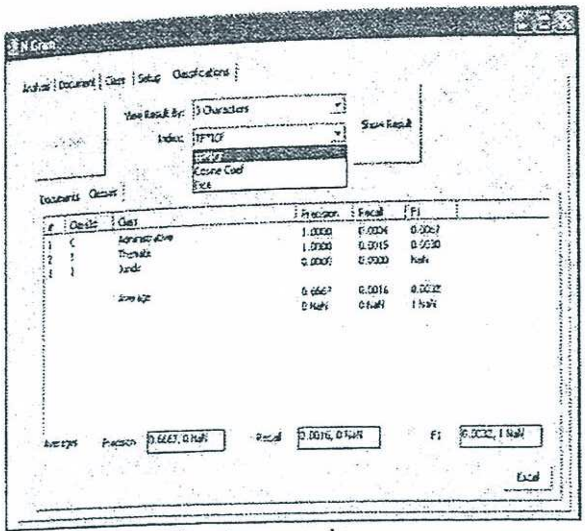
4) Sort the N-grams according to their frequencies from most frequent to least frequent. Discard the frequencies

5) This gives us the N-gram profile for a document. For training class documents, the N-gram profiles were saved in text files. Each document to be classified, went through the text preprocessing phase, then the N-gram profile was generated as described above. The N-gram profile of each text document (document profile) was compared against the profiles of all documents in the training classes (class profile) in terms of similarity. Specifically, two measures were used.



Method: 3 Characters
○ Level 1 ● Level 2

Figure 3: *Classify Documents using 3-grams*


N Gram
Analyse | Document | Class | Setup | Classifications
View Result By: 3 Characters
Index: TF*ICF
Show Result
TF*ICF
Cosine Coef
Dice
Documents | Classes
Excel

#	ClassId	Class	Precision	Recall	F1
1	C	Administrative	1.0000	0.0034	0.0067
2	T	Thematic	1.0000	0.0015	0.0030
3	J	Juridic	0.0000	0.0000	NaN
		Average	0.6667	0.0016	0.0032
			0 NaN	0 NaN	1 NaN

Averages Precision 0.6667, 0 NaN Recall 0.0016, 0 NaN F1 0.0032, 1 NaN

Figure 5: *Classifications result*

E. Results

Calculate the precision and Recall basing on similarity methods and then choose the convenient method.

Level 1: TF*ICF

#	ID	Class	Precision	Recall	F1
1	C	Administrative	1.0000	0.0034	0.0067
2	T	Thematic	1.0000	0.0015	0.0030
3	J	Juridic	0.0000	0.0000	NaN
		Average:	0.6667	0.0016	0.0032
			0 NaN	0 NaN	1 NaN

Table3 : *Level 1, TF*ICF Classification*

Cosine coefficient:

#	ID	Class	Precision	Recall	F1
1	C	Administrative	0.9980	0.5753	0.7299
2	T	Thematic	0.9985	0.2475	0.3966
3	J	Juridic	0.9979	0.1784	0.3027
		Average:	0.9981	0.3337	0.4764
			0 NaN	0 NaN	0 NaN

Table4: *Level1, Cosine coefficient Classification*

Dice:

#	ID	Class	Precision	Recall	F1
1	C	Administrative	0.0000	0.0000	NaN
2	T	Thematic	0.9985	1.0000	0.9992
3	J	Juridic	0.0000	0.0000	NaN
		Average:	0.3328	0.3333	0.3331
			0 NaN	0 NaN	2 NaN

Table 5: *Level 1, Dice Classification*

Level 2: TF*ICF:

#	ClassId	Class	Precision	Recall	F1
1	C107	تعاونية موظفي الدولة	1.0000	1.0000	1.0000
2	C104	مشيخة عقل الطائفة الدرزية	1.0000	0.5000	0.6667
6	……	…..	….	….	….
7	…….	…..	….	….	….
420	C112	وزارة السياحة	0.0000	0.0000	NaN
421	C110	مجلس تنفيذ المشاريع الكبرى لمدينة بيروت	0.0000	NaN	NaN
422	C109	المديرية العامة للصناعة	0.3333	1.0000	0.5000
		Average:	0.3677	0.3107	0.2591
			0 NaN	140 NaN	215 NaN

Table 6: *Level 2, TF*ICF Classification*

Cosine coefficient:

#	ClassId	Class	Precision	Recall	F1
1	C107	تعاونية موظفي الدولة	0.5714	1.0000	0.7273
2	C104	مشيخة عقل الطائفة الدرزية	0.3333	1.0000	0.5000
3	……	…..	….	….	….
4	…….	…..	….	….	….
420	C112	وزارة السياحة	1.0000	0.0707	0.1321
421	C110	مجلس تنفيذ المشاريع الكبرى لمدينة بيروت	0.0000	NaN	NaN
422	C109	المديرية العامة للصناعة	0.5000	1.0000	0.6667
		Average:	0.4603	0.3257	0.3099
			0 NaN	140 NaN	174 NaN

Table7:*Level2.Cosine Coefficient Classification*

Dice:

#	ClassId	Class	Precision	Recall	F1
1	C107	تعاونية موظفي الدولة	0.0000	0.0000	NaN
2	C104	مشيخة عقل الطائفة الدرزية	0.0000	0.0000	NaN
3	......	.....	....	....	....
4	....	.....	....	....	....
420	C112	وزارة السياحة	1.0000	0.0101	0.0200
421	C110	مجلس تنفيذ المشاريع الكبرى لمدينة بيروت	0.0000	NaN	NaN
422	C109	المديرية العامة للصناعة	0.0000	0.0000	NaN
		Average:	0.0283	0.0099	0.0083
			0 NaN	140 NaN	403 NaN

Table 8: *Level 2, Dice Classification*

F. *Discussion*

We remark that the cosine coefficient measure in the two levels has given us the best results between the three measures used. In level 1 the average F1 in case of using Cosine coefficient as similarity method is: **0.4764**, and in level 2 it is : **0.3099.**

Then it is the best between the three measures but it still insufficient.

In the results above you will see the term NaN which means Not a Number, it means that we have in the denominator a 0, then the number is not defined.

In level 1. We remark that the precision and Recall for Juridical class using TF*ICF and Dice coefficient is 0, that is because no correct classes or categories are found. We can explain by the expanding of juridical documents it means a juridical document can be considered as in Administrative or thematic class.

In level 2, the waste case was that of Dice coefficient in which contains 403 NaN may be because the Dice coefficient uses the intersection between the document and class divided by the sum of document and class.Then may be the ration very low for a document that belongs to a certain class.

VII. CONCLUSIONS AND FUTURE WORK

Arabic is one of the most widely used languages in the world, yet there are relatively few studies on the retrieval and classification of Arabic documents.

This paper presented the results of classifying Arabic text documents using the N-gram frequency statistics technique employing three similarity measures : TF*ICF, Cosine coefficient and Dice's measure of similarity.

Results showed that N-gram text classification using the cosine coefficient measure outperforms classification using the Dice's measure and TF*ICF weight.

This work evaluated a number of similarity measures for the classification of Arabic documents, using the Lebanese parliament documents and especially the Lebanese official journal documents Arabic corpus as the test bed.

We have proposed a segmentation method (N-gram) applied on Arabic documents, and our goal is to find the convenient similarity measure that gives us the powerful results when applied to Lebanese official journal documents.

the N-gram method is good, but still insufficient for the classification of Arabic documents, then we have to look at the future of a new approach like distributional or symbolic approach in order to increase the effectiveness.

REFERENCES

[1] Wikipedia Encyclopedia. http:// en.wikipedia.org/

[2] Lebart and Salem, 1994.

[3] Wiener et al., 1995, Schütze et al., 1995, Stricker, 2000.

[4] Yang and Chute, 1994, Yang and Liu, 1999.

[5] Lewis and Ringuette, 1994, Apté et al., 1994.

[6] Borko and Bernick, 1964, Lewis, 1998, Androutsopoulos et al., 2000,Chai et al., 2002, Adam et al., 2002

[7] Shereen Khoja and Roger Garside. Stemming Arabic text. Computer Science Department, Lancaster University, Lancaster,UK,http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps,1999.

[8] James Mayfield, Paul McNamee, Cash Costello, Christine Piatko, and Amit Banerjee, JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video, and Web Retrieval. In E. Voorhees and D. Harman (eds.), *Proceedings of the Tenth TextREtrieval Conference (TREC 2001)*, Gaithersburg, Maryland, July 2002

[9] Ahmed Abdelali ,Improving Arabic Information Retrieval Using Local variations in Modern Standard Arabic, New Mexico Institute of Mining and Technology, 2004.

[10] Victor Lavrenko. Center for intelligent Information Retrieval University of Massachusetts Amherst. *Hopkings IR workshop, 2005*