International Journal of Software Engineering and Its Applications
Vol.8, No.2 (2014), pp.127-144
http://dx.doi.org/10.14257/ijseia.2014.8.2.13

A Hybrid Method N-Grams-TFIDF with radial basis for indexing and classification of Arabic documents

Taher Zaki[1,2], Youssef Es-saady[1], Driss Mammass[1], Abdellatif Ennaji[2] and Stéphane Nicolas[2]

[1]*IRF-SIC Laboratory, University Ibn Zohr, Agadir, Morocco*
[2]*LITIS Laboratory EA 4108, University of Rouen, 76000 Rouen, France*
tah_zaki@yahoo.fr, y.essaady@uiz.ac.ma, mammass@univ-ibnzohr.ac.ma, abdel.ennaji@univ-rouen.fr, stephane.nicolas@univ-rouen.fr

Abstract

In this paper, we propose a hybrid system for contextual and semantic indexing of Arabic documents, bringing an improvement to classical models based on n-grams and the TFIDF model. This new approach takes into account the concept of the semantic vicinity of terms. We proceed in fact by the calculation of similarity between words using an hybridization of NGRAMs-TFIDF statistical measures and a kernel function in order to identify relevant descriptors. Terminological resources such as graphs and semantic dictionaries are integrated into the system to improve the indexing and the classification processes.

Keywords: *Arabic documents; classification; indexing; radial basis function; n-grams; tfidf*

1. Introduction

The great mass of textual information published in Arabic language on the net requires implementing effective techniques for the extraction of relevant information contained in large corpus of texts. The purpose of indexing is to create a representation to easily find and identify the needed information in a set of documents.

Arabic is one of the most used languages in the world, however there are only few studies looking for textual information in Arabic, so far. It is considered a difficult language to master in the field of automatic language processing, given its morphological and syntactic properties [3, 20].

The information retrieval in Arabic language, the object of our study, is a very delicate area taking into consideration its power and its wealth. However, research in this field faces major problems [8, 9].

Accordingly, we propose a new approach based on the model of n-grams and the TFIDF measure offering information extraction techniques based on portions of words. Therefore, this new method seeks to find the words which best describe the content of a document.

Hence, we are interested in the inclusion of explicit information around the text, namely the structure and distribution terms, as well as implicit information, *i.e.*, the semantics. However, the task is easier because the management of the ambiguity in the analysis of Arabic texts (inflected language, derivation, vowelization ...) is the challenge of all information retrieval systems in Arabic.

2. Related Works

Compared to other languages, Arabic has a rich morphological variation and inflectional syntactic characteristics extremely complex, which is one of the main reasons which [13, 29] explain the lack of research methods in the field of treatment of Arabic texts.

Indexing and text classification are important tasks of text processing. A typical process of text classification consists of the following steps: preprocessing, indexing, dimension reduction and classification [36].

A set of statistical models for classification and machine learning techniques have been applied to text classification, we include the linear regression model LLSF (linear least square fit mapping)[37], the K nearest neighbor [17, 1, 34], the decision tree [21], the Bayesian model [14], SVM model (Support Vector Machines) [15, 4], SVM combined with Chi-2 for feature extraction [22, 23], Maximum Entropy [30], the Rocchio algorithm [34], the distances-based classifiers [11, 12, 19], the knowledge-based classifiers as WordNet [6] and AdaBoost [27, 28].

It is difficult to compare the effectiveness of these approaches for various reasons. The first reason is that each author used different corpora. The second reason is that even those who have used the same corpus did not use the same documents for learning and testing their classifiers. The last reason is that each author used different evaluation measures: precision, recall and F-measure.

Al-Shalabi [2] uses a KNN to classify Arabic documents. They extract the keywords given by the unigrams and bigrams as features, then they apply the TFIDF measure as a selection method of these characteristics.

Thabtah [35] studied the different variants of the vector space model (VSM) using KNN algorithm through various weighting methods of the terms. These variants are the Jacaard similarity coefficient, the cosine coefficient and the Dice coefficient.

The results obtained on an Arabic database has indicated that the performance obtained by Dice-TFIDF and Jaccard-TFIDF, TFIDF surpass those obtained by the Dice based WIDF, Cosine based WIDF, Jaccard based WIDF, Cosine based TFIDF, Cosine based ITF, Dice based ITF, Jaccard based ITF, Cosine based log(1+tf), Dice based log(1+tf) and Jaccard based log(1+tf).

Zubi [38] made a comparison between the two classifiers KNN and NB applied on a set of 1562 documents classified into 6 categories and weighted using TFIDF measure. Experience has shown that KNN is more efficient.

Bawaneh [5] were compared between the two classifiers KNN and NB. The light stemmer was used as a characteristic and the TFIDF measure as a weighting of these characteristics. They have observed that KNN classifier was more efficient.

Khreisat [19] has built a classification system of Arabic text documents using frequency statistical technique N-grams and using 'Manhattan distance' as a measure of dissimilarity and the Dice operator as a measure of similarity. The Dice measure was used for comparison purposes. The results showed that the text classification using N-grams and Dice measure outperforms the classification based on N-grams and Manhattan measure.

3. Architecture of the Proposed System

3.1. Process Diagram

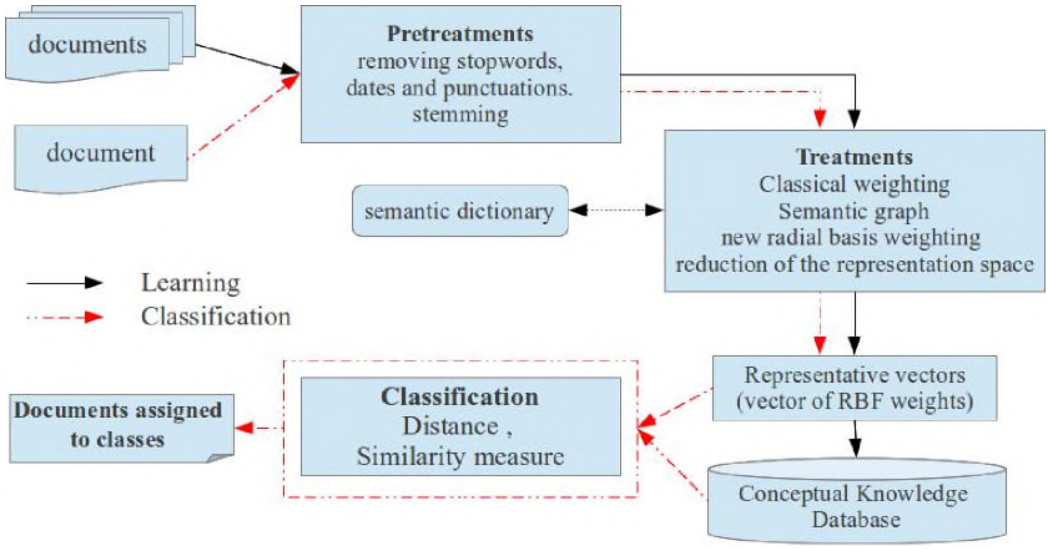


Figure 1. The Stages of the Proposed Indexing System

3.2. Used Corpus

During the learning phase we used a very reduced database of documents (initial corpus), labeled and representative of classes (sport, politics, economy & finances) sought to discriminate or to learn. The more this database is discriminating and representative the more our method becomes effective and showing better results.

To test our approach we used a corpus of Arabic-language press. This database is a collection of 5000 documents extracted from the Aljazeera[1] and Al Arabiya[2] sites.

Tables (1, 2, 3) show different results for each used measure. These results are expressed through the two criteria of recall and Precision. They show in particular the relevance of using our approach in comparison with known statistical approaches.

3.3. Preprocessing

The preprocessing phase starts by applying a noise filtering (stopwords elimination, punctuation, date) to the entire text which is followed by a morphological analysis (lemmatization, stemming) and concluded by the filtering of extracted terms. This treatment is necessary due to changes in the way in which the text can be represented in Arabic. The preparation of the text includes the following steps:

□ Converting text files in UTF-16 encoding.

□ Eliminating punctuation marks, diacritics and non-letters and stopwords.

□ Standardizing the Arabic text, this step is to transform some characters in standard form as "1, 1, إ" to "ا" and "c.j ,cs" to "is" and "ؤ" to "3"

[1] http://www.aljazeera.net/

[2] http://www.alarabiya.net/

3.4. Space of Representation

This step allows to adopt statistical vector representation using the selected terms to best represent the document. Then, to avoid the combinatorial problems related to the dimension of the space of representation [31, 7], we have adopted a frequency thresholding approach (Document Frequency Thresholding) and a principal components analysis to reduce this size.

For the choice of terms, we use a deductive method, which is to extract the vocabulary from the documents to be indexed. Therefore, we bring together a volume of documents believed to be representative of the domain, and we classify the extracted terms according to their weights.

Then we eliminate the terms deemed insignificant and out of considered domain. We distinguish thereafter between "descriptors" and "equivalent terms" (or synonyms). At the end of this phase, there is a glossary including usable descriptors and their equivalent terms for indexing and classification. Two ways for features extraction have been used. The Stemming of the terms is operated using the Khoja stemmer [18] and 3-grams as the optimal choice.

3.5. Descriptors Weighting by N-grams

The N-gram method offers the advantage of being a technique for a search based on a segment of Word. In fact, systems based on n-grams do not need preprocessing consisting in the elimination of stop- words, Stemming or lemmatization, which are essential for good performance in systems based on word search.

This phase generates a set of vectors whose elements are the 3-grams features and their appearance frequency in the document.

3.6. Descriptors Weighting by tfidf

Used in the vector model, the TFIDF (term frequency - inverse document frequency) is a statistical measure for assessing the importance of a word in a document, relatively to a documents collection or a corpus [33]. The weight increases proportionally with number of word occurrences in the document. It varies also according to the word frequency in the corpus. There are many variants of TF - IDF [32].

The basic data of this formula are f(d,t) which is the term frequency t in document d and df(t) which is the number of documents having at least one occurrence of the term t, the latter aims to give greater weight to the less frequent words which considered most discriminating. The functions TF reflect a growing monotony and IDF a decreasing monotony.

3.6.1. Problems: In the TF schema, the importance of a term t is proportional to its frequency in the document. This improves the recall but not always the precision, terms that are common are not discriminating that often leads to remove the most frequent terms: but what is the limit?

In the IDF schema, the words which appear in few documents are interesting and relevant. This scheme is intended to improve the accuracy.

n n

n t 2 2 n

3.6.2. Solution: Salton [26] has shown that the best results were obtained by multiplying TF and IDF. Finally, the weight is obtained by multiplying the two measures:

tfidf(t, d) = (tf(t, d) □ log( )) / □ tf(t , d) □ log ( )

□ □ □

n is the number of documents in the collection. n_t is the number of documents containing the term *t*.

4. Semantic Indexing with Kernel Function

We schematize the process diagrams as follows (Figure 2):

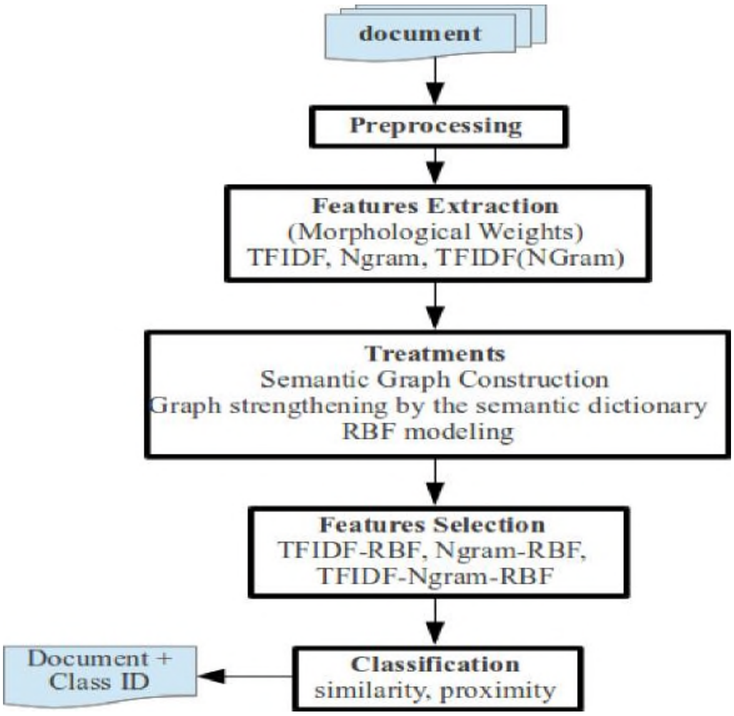


Figure 2. Process Mapping

Several studies have adapted the vector model by directly indexing concepts instead of terms. These approaches deal essentially with the synonymy by replacing the terms by their concepts. We treat most rich links between the terms by taking into account all types of semantic relations. This can solve the problem of synonymy, but also avoids the complications caused by other relations of specialization and generalization for example.

4.1. Calculation of the New Weights

Unlike existing methods, we do not restrict the use of concepts. Indeed, the terms are enriched, in terms of weighting if they are strongly connected to the neighbouring concepts and provide strong connectivity within the semantic graph.

To calculate the similarity between words, we define a radial basis function that assigns to each term a zone of influence characterized by the degree of semantic similarity and the relation between the core word and its neighbors. We have adapted our system to support any kind of semantic relations supposed existing between terms such as synonymy, meronymy, antonomy, *etc*., by a graph modeling and the use of a semantic dictionary which modelize these different relations. We chose to initially assign a weight unit (equal to 1) the semantic links in order to indicate the existence of some sort of relation between the two vertices of each edge (semantic relation or vicinity). The use of such a dictionary avoids the connectivity problems by ensuring a high connectivity within the graph and also increases the weight of the semantic descriptor.

After the preprocessing phase, we obtain three feature vectors using three different measures. TFIDF measure calculates the weight of the words' roots, and the N-grams calculates the occurrence frequency of each n-gram while the hybridization Ngram-TFIDF calculates the weights for each n-gram according to TFIDF Scheme.

In the following, we define the new statistical measures with radial basis function and we will see later how the weight of the indexing terms are enriched from the outputs of these measures.

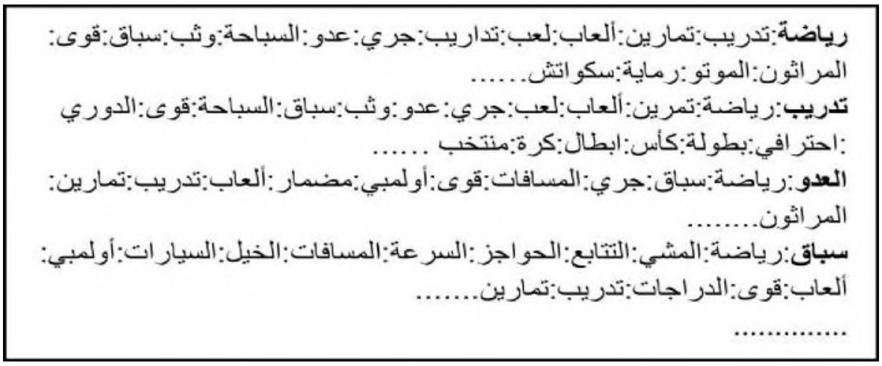
4.2. Semantic Resources

4.2.1. Auxiliary Semantic Dictionary: We developed an auxiliary semantic dictionary that is a hierarchy dictionary and containing a normalized vocabulary on the basis of generic terms and specific terms to domain. It incidentally provides definitions, relations between terms and their choice to outweigh the meanings. Relations commonly expressed in such a dictionary are:

- Taxonomic relations (of hierarchy).
- Equivalence relations (synonymy).
- Associate relations (relations of semantic proximity, close to, related to, *etc.*,).

4.2.2. Construction of the Dictionary: The dictionary is initially constructed manually based on the words found in the training set. But this dictionary can be enriched progressively during the training phase and classification to give more flexibility to our model.

Take for example the topic of sport and finance and economics, the built dictionary is shown in Figure 3 below:



رياضة:تدريب:تمارين:ألعاب:لعب:تداريب:جري:عدو:السباحة:وثب:سباق:قوى:
المراثون:الموتو:رماية:سكواتش......
تدريب:رياضة:تمرين:ألعاب:لعب:جري:عدو:وثب:سباق:السباحة:قوى:الدوري
:احترافي:بطولة:كأس:ابطال:كرة:منتخب ......
العدو:رياضة:سباق:جري:المسافات:قوى:أولمبي:مضمار:ألعاب:تدريب:تمارين:
المراثون........
سباق:رياضة:المشي:التتابع:الحواجز:السرعة:المسافات:الخيل:السيارات:أولمبي:
ألعاب:قوى:الدراجات:تدريب:تمارين.......
....

Figure 3. Example of Arabic Semantic Dictionary of the Sport Theme

Construction of the dictionaries is based on a set of dictionaries available on the web as "Almaany[3]" and "the free dictionary[4]". The semantic dictionary will be updated and fed progressively during the classification phase.

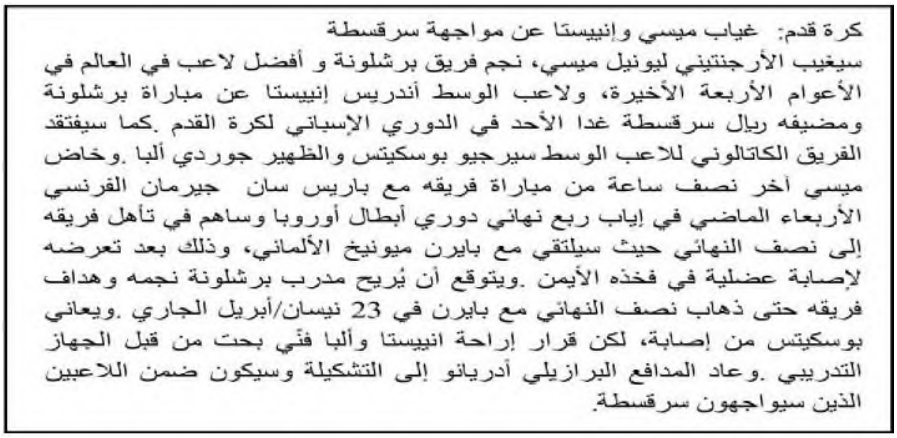
[3] http://www.almaany.com
[4] http://ar.thefreedictionary.com/

4.2.3. Semantic Networks: Semantic networks [25] were originally designed as a model of human memory. A semantic network is a labeled graph (more precisely a multigraph). An arc binds (at least) a start node to (at least) one arrival node. Relations between nodes are semantic relations and relations of part-of, cause-effect, parent-child, *etc.*.

The concepts are represented as nodes and relationships in the form of arcs. The links of different types can be mixed as well as concepts and instances.

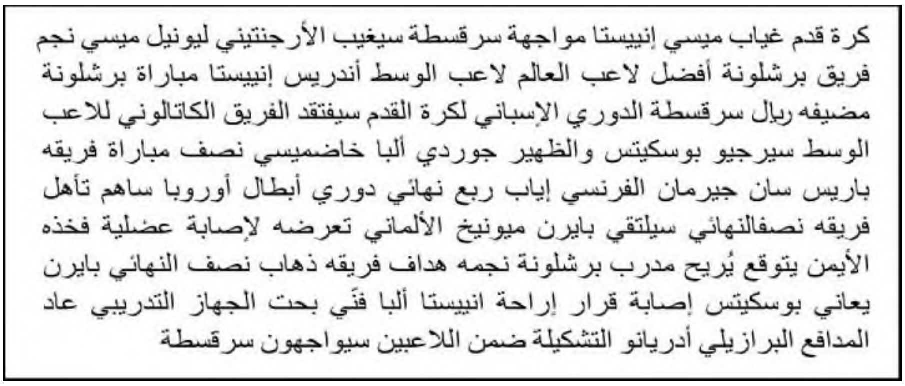
In our system, we used the concept of semantic network as a tool for strengthening of semantic graph outcome from the extracted terms of learning documents to improve the quality and representation of knowledge related to each theme of the document database.

4.2.4. The Graph Construction: It is important to note that the extraction of terminology descriptors is done in the order in which they appear in the document. Figure 6 and 7 illustrate this process for an example of the theme "Sport".



كرة قدم: غياب ميسي وإنييستا عن مواجهة سرقسطة
سيغيب الأرجنتيني ليونيل ميسي، نجم فريق برشلونة و أفضل لاعب في العالم في
الأعوام الأربعة الأخيرة، ولاعب الوسط أندريس إنييستا عن مباراة برشلونة
ومضيفه ريال سرقسطة غدا الأحد في الدوري الإسباني لكرة القدم .كما سيفتقد
الفريق الكاتالوني للاعب الوسط سيرجيو بوسكيتس والظهير جوردي ألبا .وخاض
ميسي آخر نصف ساعة من مباراة فريقه مع باريس سان جيرمان الفرنسي
الأربعاء الماضي في إياب ربع نهائي دوري أبطال أوروبا وساهم في تأهل فريقه
إلى نصف النهائي حيث سيلتقي مع بايرن ميونيخ الألماني، وذلك بعد تعرضه
لإصابة عضلية في فخذه الأيمن .ويتوقع أن يُريح مدرب برشلونة نجمه وهداف
فريقه حتى ذهاب نصف النهائي مع بايرن في 23 نيسان/أبريل الجاري .ويعاني
بوسكيتس من إصابة، لكن قرار إراحة انييستا وألبا فنّي بحت من قبل الجهاز
التدريبي .وعاد المدافع البرازيلي أدريانو إلى التشكيلة وسيكون ضمن اللاعبين
الذين سيواجهون سرقسطة.

Figure 4. Raw Text



كرة قدم غياب ميسي إنييستا مواجهة سرقسطة سيغيب الأرجنتيني ليونيل ميسي نجم
فريق برشلونة أفضل لاعب العالم لاعب الوسط أندريس إنييستا مباراة برشلونة
مضيفه ريال سرقسطة الدوري الإسباني لكرة القدم سيفتقد الفريق الكاتالوني للاعب
الوسط سيرجيو بوسكيتس والظهير جوردي ألبا خاضميسي نصف مباراة فريقه
باريس سان جيرمان الفرنسي إياب ربع نهائي دوري أبطال أوروبا ساهم تأهل
فريقه نصفالنهائي سيلتقي بايرن ميونيخ الألماني تعرضه لإصابة عضلية فخذه
الأيمن يتوقع يُريح مدرب برشلونة نجمه هداف فريقه ذهاب نصف النهائي بايرن
يعاني بوسكيتس إصابة قرار إراحة انييستا ألبا فنّي بحت الجهاز التدريبي عاد
المدافع البرازيلي أدريانو التشكيلة ضمن اللاعبين سيواجهون سرقسطة

Figure 5. Text after Preprocessing and Filtering

The construction of semantic graph takes into account the order of extraction and distribution of the terms in the document. Each term is associated with a radial basis function which determines the proximity to some vicinity (area of semantic influence of the term). We have adapted our system to support any kind of semantic relationship such as synonymy, meronymy, taxonomy, antonomy, *etc*. In addition, we initially assigned a unit weight to semantic links.

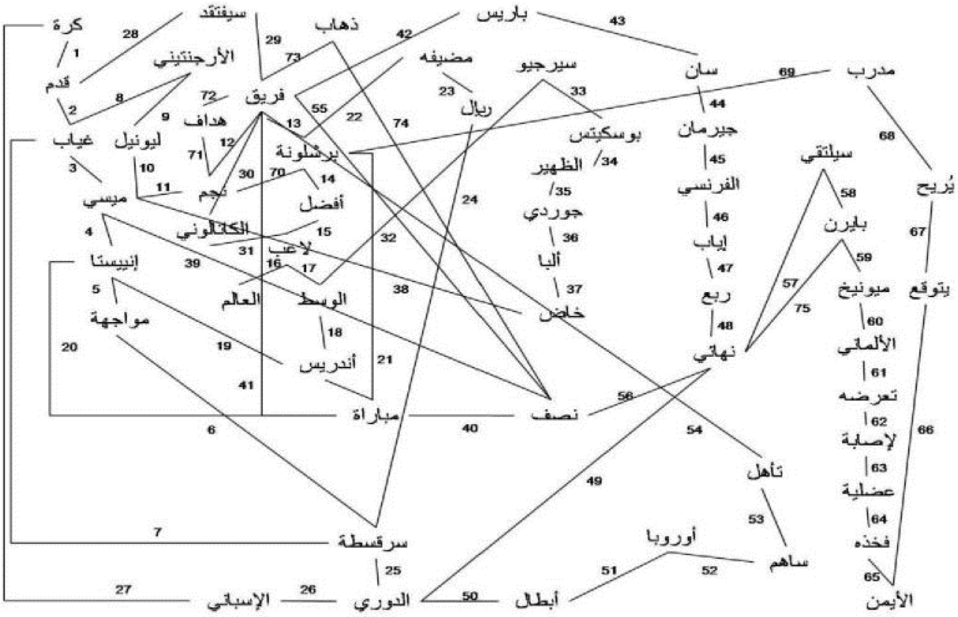


Figure 6. Semantic Graph Extracted from the Document

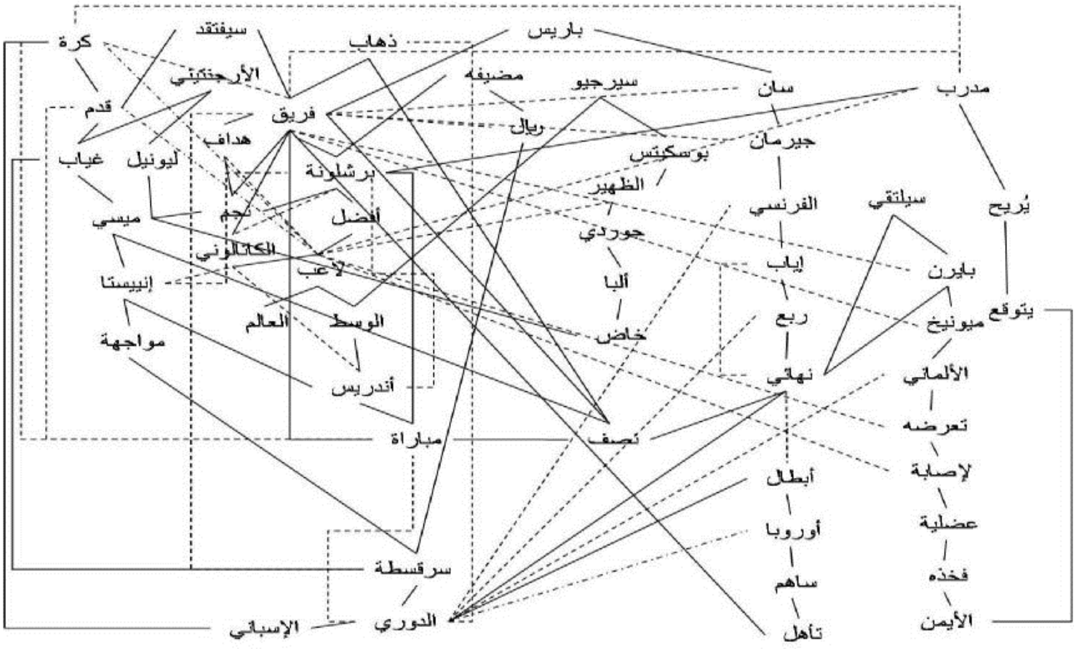


Figure 7. Strengthening of the Graph by Semantic Connections Extracted from the Auxiliary Dictionary

Then this graph is enriched by the helping semantic dictionary through adding connections whose weight is equal to 1. Such an approach allows to modelize the semantic relations supposedly existing between terms. This allows one hand to avoid connectivity problems so as to have strong network connectivity and secondly it increases the weight of the semantic descriptor terms thereafter. Unit weight means the existence of a kind of relation or a conceptual link between the corresponding terms.

Query-document matching is a projection of the query terms on the semantic graph. If these terms are in an area of strong semantic influence, then this document is relevant to this query.

In the following we will define our radial basis function and we will see the utility of the semantic graph to calculate the semantic proximity between the request and the document.

5. The New Weights with Radial Basis

The NGRAM-TFIDF with radial basis function (NGRAM-TFIDF-RBF) relies on on the determination of support in the representation space. However, unlike the classical NGRAM-TFIDF, the NGRAM-TFIDF -RBF may be fictional forms that are a combination of classical NGRAM-TFIDFs values, therefore we call them prototypes. They are associated with a zone of influence defined by a distance (Euclidean, Mahalanobis...) and a radial basis function (Gaussian, exponential,..). The discriminant function g of NGRAM-TFIDF-RBF with one output is defined from the distance of the form at the input to each of the prototypes and the linear combination of the corresponding radial basis functions:

$$(\qquad\qquad (2)$$

Where d(x,sup$_i$) is the distance between the input x and the support sup$_i$, {w$_0$, ..., w$_n$} is the weight of the combination and the radial basis function.

The NGRAM-TFIDF-RBFs prototypes represent the distribution of examples in representation space E (terms). In addition the multi-class problem management is easier in the NGRAM-TFIDF-RBFs.

The NGRAM-TFIDF-RBFs modeling is both discriminating and intrinsic. Indeed the layer of radial basis functions corresponds to an intrinsic description of the training data, then the combination layer at the output seeks to discriminate different classes.

In our system, a Cauchy function is used as a radial basis function:

$$x \quad)_1 \quad 1\frac{1}{\quad} \qquad\qquad (3)$$

We define two new operators:

(4)

Relw(c) is the relational weight of the concept c_r (root) and *degree(c)* is the number of incoming and outgoing edges of the vertex *c*. It therefore represents the connection density of the concept *c* in the semantic graph.

(5)

SemD (c_1, c_2) is the semantic density of the link (c_1, c_2). This is the ratio of the minimal semantic distance MinCost(c_1,c_2) between c_1 and c_2, calculated by Dijkstra's algorithm. This distance is calculated from the semantic graph, this latter is built from the document based on the minimal cost of the spanning tree (*i.e.*, the minimal cost tree by following all minimal paths from c1 to c2 through the other vertices of the semantic graph). This reflects the importance of the link (c1, c2) compared to all existing minimal paths. Subsequently we calculate the semantic distance (conceptual) as follows:

The proximity measure is a Cauchy function:

(6)

(7)

The contribution of these defined operators is that they give more importance to concepts which have dense semantic vicinity where they have good connectivity within the graph. This has also been verified during the validation of the prototype. The documents are represented by vector sets of terms. The weights of the terms are calculated according to their distribution in documents following three classical statistical measures, n-grams, TFIDF and TFIDF-Ngrams. The weight of a term is enriched by the conceptual similarities of the co-occurring terms in the same topic according to statistical measures improved with a radial basis namely Ngrams-RBF, TFIDF-ABR and TFIDF-Ngrams-RBF.

We also noticed that some terms, considered as significant for the documents indexing, were at the bottom of the ranking according to the classical weighting NGRAM and TFIDF separately. However, after the calculation of the NGRAM-TFIDF-ABR weighting these terms were better classified at the top of the rankings.

5.1. Radial Basis NGRAM

The use of N-gram method (with N = 3 number of characters) in information retrieval in Arabic documents is more efficient than the "keyword matching". The choice of statistical measures such as the trigrams seems relevant since the majority of Arabic words are derived from a root of 3 characters.

Unlike other works which proceed to the use of n-grams without the preliminary pretreatments such as the removal of stop-words, joints ... we are aware that this step is essential to minimize noise.

The use of N-gram method for documents indexing and classification remains insufficient to achieve good results for the Arabic language. For this we thought of adding semantic relevance to this measure taking into account the semantic vicinity of the extracted terms by combining N-gram with a kernel function. Thus, the formula becomes:

$$NGRAM-RBF(t,T) = NGRAM_0(t,T) + E\ NGRAM(t,T) \bullet (SemDist(t,T))$$

(8)

$NGRAM_0(t,T)$: the initial value of the occurrence frequency of trigrams t in the theme T

n

n

or simply, $NGRAM$ □ $_{RBF\ (t,T)}$ □ $NGRAM_{0\ (t,T)}$ □ □ $NGRAM\ (ti\ ,T)$ □ $_{\mathrm{Pr}\ oximity\ (ti,T}$)

i

With $_{Proximity\ (t,\ t,)}$ < *threshold* t e $T_{,,}$ as T. all n terms in the theme.

threshold: a value which sets the proximity to a certain vicinity (area of semantic influence of the term t), we set this value initially to the proximity between the concept of t and the general context (a concept that represents the theme).

calculated by classical n-grams.

(9)

5.2. Radial Basis TFIDF

We proceed to calculate the terms TFIDF for all of the themes of training basis to deduce the global relevance. Then we calculate local relevance through our radial basis function defined above by a combination with the classical TFIDF and accepting only the terms within the zone of influence. Noted that weight TFIDF-RBF (t) is calculated as follows:

(10)

Or simply,

(11)

With as all n terms in the theme.

threshold: a value which sets the proximity to a certain vicinity (zone of semantic influence of the term t), we set this value initially to the proximity between the concept of t and the general context (a concept that represents the theme).

: The initial value of the weight of term t (root) to the theme T calculated by the classical TFIDF.

5.3. Classification

In the classification phase, we adopted, in this preliminary version of our prototype, the KNN algorithm in order to assess the relevance of our choice of representation. Several metrics have been proposed in the literature, however we had to also choose a metric adapted to this context which is the Dice operator whose expression is:

Where,

$| P_i |$ is the number of terms in the profile P_i (vector representing the document i).

$| P_i \wedge P_j |$ is the number of terms of intersection between the two profiles P_i and P_j.

6. Complete Algorithms

6.1. NGRAM-RBF and TFIDF-RBF

Inputs: a textual database (whether pre-sorted or not) of documents representing three themes, sport, economy and politics.

semantic dictionaries of themes, sport , economy and politics.

Output: a set of indexed documents, weighted and classified, representing three themes, sport, economy and politics.

Algorithm:

1. Read each document, or documents of a class, from the dataset.

   do

2. Remove punctuation, stopwords, dates, and vowels;

3. For each word of d, calculate the TFIDF weighting according to the database;

4. Sliding window of N characters, which scans the document d by calculating the occurrence frequency of each N-gram in document d;

5. Build the lexicons of d (each lexicon corresponds to a measure) ;

6. Crossing the space of vector representation (generate two vectors of weights according to TFIDF and N-grams measures);

7. Construction of the semantic graph

where,

, is weighted by the semantic proximity

, build the neighbors set of

8. Strengthen the graph by the semantic link extracted from the dictionary

$$\frac{2\left|P_i \cap P_j\right|}{\left\|P_i\right| + \left|P_j\right\|} \tag{12}$$

calculate (see equations (9, 11))

with

with

9. Generation of new representation space ;

10. Reduction of space ;

11. Add ọt

$D = \square\ D_{sp} , D_{ec} , D_{po} \square$

6.2. Radial basis Hybrid Method

In this approach we follow the same approach as the previous algorithm, the difference in the calculation of the weighting and the construction of the graph. Indeed, after the preprocessing phase, we extract at first all n-grams then we calculate the weighting of each of them, using the TFIDF measure. The extracts N-grams are also used for the graph construction as explained by the following algorithm:

$Dic = \square\ Dic_{sp} , Dic_{ec} , Dic_{po} \square$

Algorithm

Inputs:

: a textual database (whether pre-sorted or not) of documents representing three themes, sport , economy and politics.

: a semantic dictionaries of themes, sport , economy and politics.

Output:

: a set of indexed documents, weighted and classified.

□ □

begin

1. Read each document, or documents of a class, from the database *do*

2. Remove punctuation, stopwords, dates, and vowels;

3. Sliding window of N characters, which scans the document d by calculating the occurrence frequency of each N-grams;

4. Calculation of the n-grams weighting which extracted from d according to database;

5. Build the lexicon of d;

6. Transition to the space vector representation (vector weighted by hybrid measure);

7. Construction of the semantic graph *where , , all words in d. extracting the set N_g of corresponding n-grams N_g ={n_k, n_k subword of lenght n extracted from t_i}*

8. Passage from to

V' ={x, x NG_d set of ngrams extracted from d}

E □

□

, is weighted by the semantic proximity

, built , the neighbors set of

9. Strengthen the graph by the semantic link extracted from the dictionary

Calculate (see equations (8 , 10)) with

m

$i=$

with

10. Generation of new representation space ;

11. Reduction of space ;

12. A d d 6 .

Results

Tables (1, 2, 3) show the different results obtained for each measure used. These results are expressed through the two criteria: the recall and Precision. They show in particular the relevance of the use of our approach in comparison with known statistical approaches.

Table 1. Results of TFIDF and TFIDF-RBF

Method	Corpus	Precision	Recall
TFIDF	Sport	0,83	0,73
	Politic	0,68	0,61
	Finance & economics	0,56	0,71
TFIDF-RBF	Sport	0,94	0,77
	Politic	0,78	0,67
	Finance & economics	0,59	0,71

Table 2. Results of NGRAM and NGRAM-RBF

Method	Corpus	Precision	Recall
NGRAM	Sport	0.78	0,68
	Politic	0.65	0,50
	Finance & economics	0.66	0,49
NGRAM-RBF	Sport	0.81	0,67
	Politic	0.60	0,53
	Finance & economics	0.62	0,51

Table 3. Results of NGRAM-TFIDF and NGRAM-TFIDF-RBF

Method	Corpus	Precision	Recall
NGRAM-TFIDF	Sport	0,84	0,78
	Politic	0,88	0,71
	Finance & economics	0,59	0,64
NGRAM-TFIDF-RBF	Sport	0,88	0,81
	Politic	0,86	0,76
	Finance & economics	0,67	0,87

From Tables, we can see that the best performances are recorded in the sport because the sport has a limited space compared to other domains. In addition, they shows that the economic and financial performances is low, this is due, on the one hand to the nature of newspaper articles in our possession which relate to the domain of finance and economy and on the other hand the involvement of politics in this domain which the most often generates an overlap of meaning.

7. Discussion and Conclusion

The preceding tables present the experimental results that we obtained on the indexation and classification of an Arabic corpus. We have chosen to apply statistical measures TFIDF and n-grams which are references in this domain. Then, we have developed a system for indexing and contextual classification of Arabic documents, based on the semantic vicinity of terms and the use of a radial basis modeling.

The use of semantic resources, namely semantic graphs and semantic dictionaries greatly improves the process of indexing and classification.

Subsequently, we have proposed new statistical measures with radial basis, taking into account the concept of semantic vicinity using a calculation of similarity between terms by combining the calculation of TFIDF and n-grams with a kernel function, for the evaluation and extraction of the indexing terms in order to identify the relevant concepts which represent best a document.

By comparing the obtained results, we find that the use of radial basis functions largely improves the performance of the measures with which they are combined. In particular, when they are combined with the TFIDF, however, they have shown less performance at the level of n-grams, although this method is widely invested on a number of text processing and information retrieval given its benefits regardless of the processed language. This may be caused by the choice of the optimal value of n that can cause quite a lot of noise by introducing some words which have no meaning in the lexicon. We thought to do a second filtering after extracting of the n-grams list, but that appears unnecessary since we will lose more semantic information which subsequently degrades the precision. However, these measures may also be combined together as in the case of n-grams-TFIDF hybridization which has improved the results as compared to the use of n-grams or N-grams-RBF all alone.

We noticed that the results of indexing contain exactly the keywords sorted by relevance. We also set a threshold for the semantic enrichment, which can lead to return some unwanted terms which are quite different from those sought.

Another point to take into account and which can degrade the precision of classical statistical methods is the presence of complex concepts. We proposed a partial solution to this scourge by attempting to model these complex forms within the semantic dictionary, nevertheless this solution is insufficient given the richness of the Arabic language and the puns used by this language. However, this point may be an interesting track to explore since the long concepts are generally less prone to ambiguity.

The calculation of semantic proximity during indexing alleviates the treatments during the search. Although this phase is costly in time but the results are very interesting. But despite the good outcomes, we noticed that the results of indexing contain exactly the sought keywords sorted by relevance. We have also set a threshold for the semantic enrichment, which can lead to return some fairly distant adverse terms of those sought.

Acknowledgements

This work is supported with grants from the “Action intégrée Maroco-Française” n° MA/10/233 and the AIDA project, program Euro-Mediterranean 3 +3: n°M/09/05.

References

[1] R. Al-Shalabi, G. Kanaan and M. Gharaibeh, “Arabic Text Categorization Using kNN Algorithm”, Proceedings of The 4th International Multiconference on Computer Science and Information Technology, Amman, Jordan, vol. 4, **(2006)** April 5-7.

[2] R Al-Shalabi and R. Obeidat, “Improving KNN Arabic Text Classification with N-Grams Based Document Indexing”, INFOS2008, Cairo-Egypt, **(2008)** March 27-29.

[3] M. Aljlayl and O. Frieder, “On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach”, 11th International Conference on Information and Knowledge Management (CIKM), Virginia (USA), **(2002)** November, pp. 340-347.

[4] S. Alsaleem, “Automated Arabic Text Categorization Using SVM and NB”, International Arab Journal of e-Technology, vol. 2, no. 2, **(2011)** June.

[5] M. J. Bawaneh, M. S. Alkoffash and A. I. Al Rabea, “Arabic Text Classification using K-NN and Naive Bayes”, Journal of Computer Science, vol. 4, no. 7, **(2008)**, pp. 600-605.

[6] M. Benkhalifa, A. Mouradi and H. Bouyakhf, “Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization”, International Journal of Intelligent Systems, vol. 16, no. 8, **(2001)**, pp. 929-947.

[7] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet allocation”, Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 993-1022.

[8] A. Chalabi, “MT-Based Transparent Arabization of the Internet TARJIM.COM”, White, J.S. (Ed) AMTA Springer: Verlag Berlin Heidelberg, **(2000)**, pp. 189-191.

[9] K. Daimi, “Identifying Syntactic Ambiguities in Single-Parse Arabic Sentence”, Computer and humanities, vol. 35, **(2001)**, pp. 333-349.

[10] S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Hrashman, “Indexing by latent semantic analysis”, Journal of th american society for information science, vol. 41, no. 6, **(1990)**, pp. 391-407.

[11] R. M. Duwairi, “A Distance-based Classifier for Arabic Text Categorization”, Proceedings of the International Conference on Data Mining, Las Vegas, USA, **(2005)**, pp. 187-192.

[12] R. M. Duwairi, “Machine Learning for Arabic Text Categorization”, Journal of American society for Information Science and Technology, vol. 57, no. 8, **(2006)**, pp. 1005-1010.

[13] A. M. El-Halees, “Arabic Text Classification Using Maximum Entropy”, The Islamic University Journal, vol. 15, no. 1, **(2007)**, pp. 157-167.

[14] M. Elkourdi, A. Bensaid and T. Rachidi, “Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm”, Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Scriptbased Languages”, Geneva, **(2004)** August 23-27, pp. 51-58.

[15] T. F. Gharib, M. B. Habib and Z. T. Fayed, “Arabic Text Classification Using Support Vector Machines”, International Journal of Computers and Their Applications, vol. 16, no. 4, **(2009)**, pp. 192-199.

[16] F. Harrag, E. El-Qawasmah and A. Al-Salman, “Stemming as a Feature Reduction Technique for Arabic Text Categorization”, 10th International Symposium on Programming and Systems (ISPS), Algeria, **(2011)**.

[17] G. Kanaan, R. Al-Shalabi and A. AL-Akhras, “KNN Arabic Text Categorization Using IG Feature Selection”, Proceedings of The 4th International Multiconference on Computer Science and Information Technology, Amman, Jordan, vol. 4, **(2006)** April 5-7.

[18] S. Khoja and S. Garside, “Stemming Arabic Text”, Computing Department, Lancaster University, Lancaster, U.K., **(1999)** September 22.

[19] L. Khreisat, “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.

[20] L. S. Larkey, L. Ballesteros and M. Connell, “Improving Stemming for Arabic Information Retrieval: Light Stemming and Cooccurrence Analysis”, Proceedings of the 25th Annual International Conference on Research and Deve lopment in Information Retrieval (SIGIR'02), Tampere, Finland, **(2002)** August, pp. 275282.

[21] Y. H. Li and A. K. Jain, “Classification of text documents”, Comput. J., vol. 41, no. 8, **(1998)**, pp. 537-546.

[22] A. M. Mesleh, “CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System”, Proceedings of the 2nd International Conference on Software and Data Technologies, (Knowledge Engineering), Barcelona, Spain, vol. 1, 22-25, **(2007)** July, pp. 235-240.

[23] A. M. Mesleh, “CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System”, Journal of Computer Science, vol. 3, no. 6, **(2007)**, pp. 430-435.

[24] A. M. Mesleh, “Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study”, 12th WSEAS Int. Conf. on Applied Mathematics, Cairo, Egypt, **(2007)** December 29-31.

[25] M. R. Quillian, “Semantic memory”, Semantic information processing, **(1968)**.

[26] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval”, Information Processing and Management, vol. 24, **(1988)**, pp. 513-523.

[27] R. E. Schapire and Y. Singer, “BoosTexter: a boosting-based system for text categorization”, Mach. Learn, vol. 39, no. 2-3, **(2000)**, pp. 135-168.

[28] R. E. Schapire, Y. Singer and A. Singhal, “Boosting and Rocchio applied to text filtering”, Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne, Australia, **(1998)**, pp. 215-223.

[29] A. M. Samir, W. Ata and N. Darwish, “A New Technique for Automatic Text categorization for Arabic Documents”, Proceedings of the 5th Conference of the Internet and Information Technology in Modern Organizations, Cairo, Egypt, **(2005)** December, pp. 13-15.

[30] H. Sawaf, J. Zaplo and H. Ney, “Statistical Classification Methods for Arabic News Articles”, Paper presented at the Arabic Natural Language Processing Workshop (ACL2001), Retrieved from Arabic NLP Workshop at ACL/EACL, Toulouse, France, **(2001)**.

[31] F. Sebastiani, A. Sperduti and N. Valdambrini, “An improved boosting algorithm and its application to automated text categorization”, Technical report, Paris, France, **(2000)**.

[32] F. Seydoux, M. Rajman and J.-C. Chappelier, “Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire”, Ph.D. Thesis, **(2006)**.

[33] P. Soucy and G. W. Mineau, “Beyond tf-idf weighting for text categorization in the vector space model”, L. P. Kaelbling and A. Saffiotti, editors, IJCAI, Professional Book Center, **(2005)**, pp. 1130-1135.

[34] M. Syiam, Z. Fayed and M. Habib, “An Intelligent System for Arabic Text Categorization”, International Journal of Intelligent Computing and Information Sciences, vol. 6, no. 1, **(2006)**, pp. 1-19.

[35] F. Thabtah, W. Hadi and G. Al-shammare, “VSMs with K-Nearest Neighbour to Categorise Arabic Text Data”, The World Congress on Engineering and Computer Science, San Francisco, USA, **(2008)** October, pp. 778-781.

Authors

Taher ZAKI received the DESA degree in Computer Science from Ibn Zohr University, and now is a PhD student at the same University, Faculty of Sciences, in the " images pattern recognition systems intelligent and communicating " Laboratory, under the supervision of Prof. Driss Mammass. His research interests systems of information retrieval, text indexing and archive of documents.

Youssef ES-SAADY is currently research professor at the Faculty polydisciplinary of Taroudant, University Ibn Zohr, Morocco. He received the PhD degree in computer science from Faculty of Sciences, University Ibn Zohr, in 2012, on automatic recognition of printed and handwritten Amazigh characters, texts and documents. His current research interests include pattern recognition, image analysis, indexing and archiving of documents and automatic processing of natural languages.

Driss MAMMASS is professor of Higher Education at the Faculty of Sciences, University Ibn Zohr, Agadir Morocco. He received a Doctorat in Mathematics in 1988 from Paul Sabatier University (Toulouse - France) and doctorat d’Etat-es-Sciences degrees in Mathematics and Image Processing from Faculty of Sciences, University Ibn Zohr Agadir Morocco, in 1999. He supervises several Ph.D theses in the various research themes of mathematics and computer science such as remote sensing and GIS, digital image processing and pattern recognition, the geographic databases, knowledge management, semantic web, etc. He is currently the head of the Graduate School of Agadir technology *(*ESTA*)* and the IRF-SIC Laboratory (Image Reconnaissances des Formes, Systèmes Intelligents et Communicants) and an unit of formation and research in doctorat on mathematics and informatics.

Abdellatif ENNAJI has been an associate professor at the University of Rouen since 1993. He received his Ph.D. from the University of Rouen in 1993 in the fields of machine learning and pattern recognition. His major scientific interest include incremental techniques for statistical and hybrid machine learning, data analysis and clustering. The main applications of these activities concern pattern recognition problems and Arabic text mining and recognition. Dr Ennaji has coauthored over 80 publications.

Stéphane NICOLAS (18/04/1979) received the PhD degree in computer science from the University of Rouen, France, in 2006, on image segmentation using conditional random fields for document image indexing. He is currently on assistant professor at the University of Rouen since September 2007, and a researcher of the LITIS laboratory where he integrates the “Document and Learning” group. He is a member of the French association for pattern recognition (AFRIF) and a member of the French research group on handwriting recognition GRCE. His main research interests include computer vision, image analysis, pattern recognition, machine learning, and statistical tools for signals modeling and classification, mainly applied to handwritten document layout analysis and information extraction from handwritten documents.