INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES
DOI: 10.46300/9101.2021.15.18 Volume 15, 2021

Arabic Documents Classification by a Radial Basis Hybridization

Taher Zaki
IRFSIC Laboratory
Ibn Zohr University
Agadir, Morocco
tah_zaki@yahoo.fr

Driss Mammass
IRFSIC Laboratory
Ibn Zohr University
Agadir, Morocco
mammass@uiz.ac.ma

Abdellatif Ennaji
LITIS Laboratory
University of Rouen
Rouen, France
Abdel.Ennaji@univ-rouen.fr

Stéphane Nicolas
LITIS Laboratory
University of Rouen
Rouen, France
stephane.nicolas@univ-rouen.fr

Received: June 8, 2019. Revised: September 22, 2021. Accepted: October 12, 2021. Published: November 23, 2021.

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 Okapi 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-OKAPI 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 language; classification; graph semantic; indexing; kernel function; n-grams; okapi; semantic dictionary; semantic vicinity; similarity.

I. INTRODUCTION

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

Arabic is one of the most used languages in the world, however so far there are only few studies looking for textual information in Arabic. It is considered as a difficult language to deal in the field of processing automatic language, considering its morphological and syntactic properties [4][21].

The information retrieval in Arabic language, object of our study, is a very delicate area, given its power and its wealth. However, this research poses major problems [9][10].

Faced with these failures, we propose a new approach based on the model of n-grams and the Okapi 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.

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

II. 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 for which [14][33] explains the lack of research methods in the field of treatment of Arabic.

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 [38].

A set of statistical models for classification and machine learning techniques have been applied to text classification : the linear regression model LLSF (linear least square fit mapping) [39], the K nearest neighbor [18][1][36], the decision tree [22], the Bayesian model [15], SVM model (Support Vector Machines) [16][5], SVM combined with Chi-2 for feature extraction [23][24][25], neural networks [17], Maximum Entropy[34], the Rocchio algorithm [36], the distances-based classifiers [12][20][13], the knowledge-based classifiers as WordNet [7] and AdaBoost [31][32].

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, it is likely that they did not use the same documents for learning and testing of their classifiers. The last reason is that each author used different evaluation measures: precision, recall and F-measure …

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

Thabtah [37] 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).

Al-Shalabi [3] have applied k-nearest neighbors algorithm and the keywords are extracted based on their TFIDF weighting in the documents, their system has reached a micro-average precision of 0.95.

Zubi [40] 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 [6] compared 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 been observed that KNN classifier was more efficient.

Khreisat [20] 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.

III. ARCHITECTURE OF THE PROPOSED SYSTEM

A. *Process diagram*

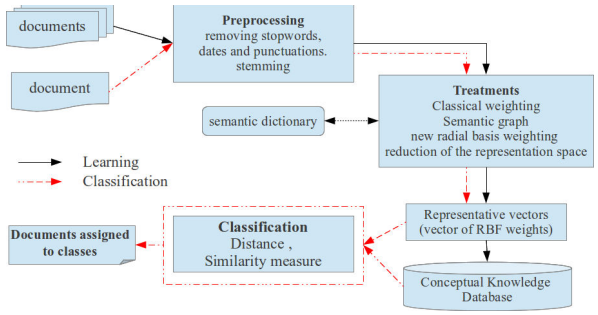

Fig. 1. *learning and clasification procedure*

B. *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. More this database is discriminating and representative 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 sites Aljazeera[1] and Al Arabiya[2].

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

C. *Preprocessing*

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

- Convert text files in UTF-16 encoding.
- Elimination of punctuation marks, diacritics and non-letters and stopwords.
- Standardization of the Arabic text, this step is to transform some characters in standard form as "إ ,أ ,آ" to "ا" and "ى ,ء ,ي" to "ئ" and "ؤ" to "و".

D. *Representation Space*

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 [35] [11][8] 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 extracted terms are classified 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 [19] and 3-grams as the optimal choice.

E. *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 stopwords, 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 frequencies in the document.

1 http://www.aljazeera.net/
2 http://www.alarabiya.net/

F. *Weighting of descriptors by Okapi*

Generally, the best terms deemed relevant and discriminating, being those that appear frequently in documents, but rarely in the rest of the collection. A document can then be represented only by the terms of a frequency sufficiently significant. The incompetence of this model is that it does not take into account the length of the document, some normalization techniques have been proposed (eg. BM25 proposed by Robertson and Walker [27][28][29] for not giving more importance to a long document with regard to a short document.

The underlying idea is that such words helps to discriminate between texts with different themes. Hence, we must get rid of the dependence of occurrence frequency in the documents as a word that appears n times in a document d_j, that does not necessarily mean that it is n times more important than in a document d_k where it appears only once. The second idea is that longer documents typically have rather high weight because they contain more words, so the appearance frequencies tend to be higher. To avoid these problems, they have adopted a new indexing technique known as Okapi formula [30]:

$$Okapi(i,j) = \frac{tf(i,j)\cdot idf(i)}{[(1-b)+b\cdot NDL(d_j)]+f(i,j)} \quad (1)$$

where NDL (dj) is the normalized length of dj, i.e. its length (the number of words that it contains) divided by the average length of documents in the corpus. The constant b is a parameter belonging to the interval [0,1], depending on the collection, which control the effect of the document length , this parameter is used to adjust the impact of document length normalization, a value b = 1 , indicates that the documents are long because they contain terms which are repeated, it fully corresponds to the words weight scaling by the document length , while b = 0 corresponds to no standardization of length, in this case documents are long because they are multitopics (contain separate terms). Experiments have shown that a reasonable value is b = 0.75.

IV. *Semantic Indexing with Kernel Function*

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. We schematize the process diagrams as follows:

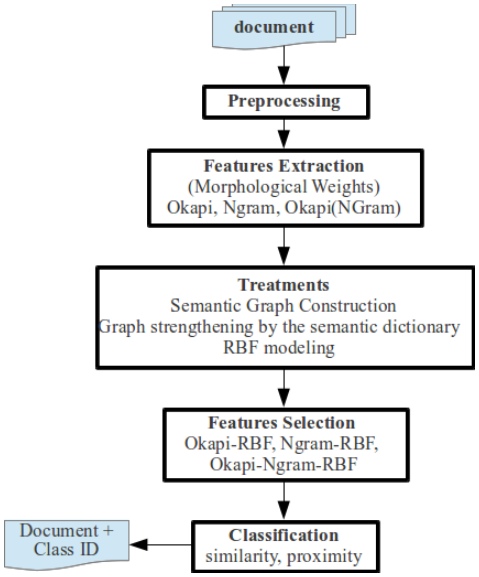

Fig. 2. *Process of the proposed System*

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

A. *Calculation of the new weights*

Unlike existing methods, we do not restricted to 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 $\varphi(d)$ 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. The most obvious way to evaluate semantic similarity in a taxonomy is to compute the distance between the nodes as the shortest path. However, taxonomies are not always at the same level of granularity, some parts may also be denser than others. These problems can be solved by associating weights to links. Thus we have chosen to consider all types of relations supposedly existing between words (conceptual issue) and the distribution of words in documents (structural issue).

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.

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.

1) Semantic resources

a) 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.).

b) 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 Figures 3 and 4 below:

economy, finances, enterprise, industrialism, market, capitalism, socialism, system, brevity, conservation, downsizing, financial status, productive power **…**
finances, budget, account, bill, financing, money, reckoning, score, banking, business, commerce, economic science, economics, political economy, investment **…**
budget, account, bill, calculate, estimate, finance, money, matters, reckon, reckoning, score, assortment, bunch, balanced, cheap,operating budget **…**
……..

Fig. 3. *Semantic dictionary of finance and economics*

رياضة :تدريب :تمارين :تمرين : ألعاب :لعب :تداريب :جري :عدو :السباحة :وثب :سباق :قوى :المراثون :الموتو :رماية :سكوانش.....

تدريب:رياضة:تمرين:ألعاب:لعب:جري:عدو:وثب:سباق:السباحة:قوى :الدوري:الاحترافي: بطولة:كأس:ابطال:كرة:منتخب.....

لعب:رياضة:تداريب:جري:عدو:وثب:سباق:قوى:بطل:نجم:مهاجم:فريق:أولمبي:فيفا:مشي:م ضمار :قفز :مدرب.....

العدو:رياضة:سباق:جري:المسافات:قوى:أولمبي:مضمار:ألعاب:تدريب:تمارين:المراثون .....

Fig. 4. *Example of Arabic semantic dictionary of the sport theme*

The initial 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..

c) Semantic networks

Semantic networks [26] 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.

d) 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. Figures 5 and 6 illustrate this process for an example of the theme " finance and economy".

WASHINGTON (Reuters) – **President** Barack Obama signed a $30 billion small **business** lending **bill** into **law** on Monday, claiming a victory on **economic policy** for his fellow **Democrats** ahead of November **congressional elections**.
The **law** sets up a lending **fund** for **small businesses** and includes an additional $12 billion in **tax breaks** for small **companies**."It was critical that we cut **taxes** and make more **loans** available to **entrepreneurs**," Obama said in remarks at the White House. "So today after a long and tough fight, I am signing a **small business jobs bill** that does exactly that."
Obama is trying to show **voters**, who are unhappy about 9.6 percent **unemployment**, that he and his party are doing everything they can to boost the tepid U.S. **economy**.
Democrats said they backed the **bill** because **small businesses** had trouble getting **loans** after the **financial crisis** that began in December 2007.
They estimate the **incentives** could provide up to $300 billion in new **small business credit** in the coming years and create 500,000 new **jobs**.

Fig. 5. *Raw text*

president , business, bill, law , economic policy, democrats , congressional elections
law, fund , small businesses, tax breaks, companies , taxes, loans, entrepreneurs , small business jobs bill
voters, unemployment , economy
democrats, bill, small businesses, loans, financial crisis
incentives, small business credit, jobs

Fig. 6. *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 a some vicinity (area of

3 http://www.almaany.com

4 http://ar.thefreedictionary.com/

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. Then this graph is enriched through the auxiliary semantic dictionary by adding connections which weight 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 a 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 (Figures 7, 8).

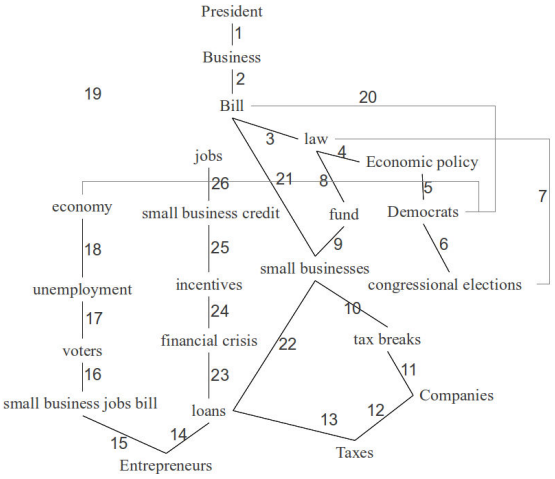

Fig. 7. *Semantic graph extracted from the document*

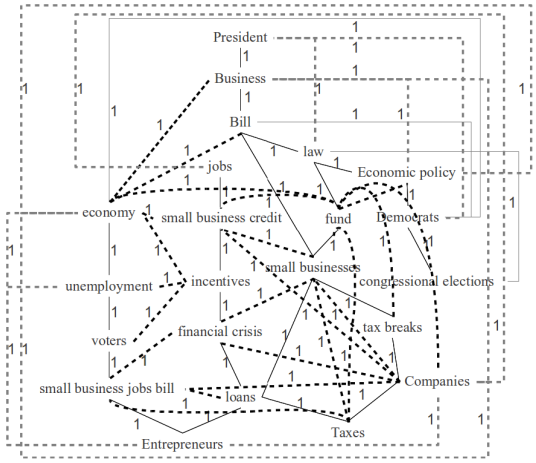

Fig. 8. *Strengthening of the graph by semantic connections extracted from the auxiliary dictionary*

First, the constructed graph represents all the lemmas of the text and synthesizes their mutual relations of: co-occurrence, synonymy, Antonymy, polysemy,...

Secondly, this graph supports the presence of compound words. These words are juxtapositions of two free lexemes to form a third that is a lemma (Word) and whose meaning is not necessarily guessed by one of the two components separately (for example: comic strip, Air Force, vice president , mayor-elect ...).

These terms lose any informational data if they are considered separately or if they have undergone the traditional operations of filtering and preprocessing. To this end, we have proposed a partial solution of the problem by including in the semantic dictionary, compound terms deemed relevant and informational.

2) The new weights with radial basis

The NGRAM-OKAPI with radial basis function (NGRAM-OKAPI-RBF) relies on the determination of support in the representation space. However, unlike the classical NGRAM-OKAPI, the NGRAM-OKAPI-RBF may be a fictional forms that are a combination of classical NGRAM-OKAPIs 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-OKAPI-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:

$$g(x) = w_0 + \sum_{i=1}^{N} w_i \, \varphi(\, d(x, \sup_i)) \qquad (2)$$

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 φ the radial basis function.

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

The NGRAM-OKAPI-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 :

$$\varphi(d) = \frac{1}{1 + d} \qquad (3)$$

we define two new operators:

$$Relw(c) = \frac{degree(c)}{total \quad number \quad of \quad concepts} \qquad (4)$$

Relw(c) is the relational weight of the concept c (n-gram or 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.

$$SemDensity(c_1, c_2) = \frac{MinCost(c_1, c_2)}{minimal\ \ cost\ \ of\ \ the\ \ Spanning\ \ Tree} \quad (5)$$

SemDensity(c_1, c_2) is the semantic density of the link (c_1, c_2). This is the ratio of the minimal semantic distance CoutMin (c_1, c_2) between c_1 and c_2, calculated by Dijkstra's algorithm (Cormen et al., 2001). This distance is calculated from the semantic graph, this latter is built from the document based on the minimal cost of the spanning tree (ie the minimal cost tree by following all minimal paths from c_1 to c_2 through the other vertices of the semantic graph). This reflects the importance of the link (c_1, c_2) compared to all existing minimal paths. Subsequently we calculate the semantic distance (conceptual) as follows:

$$SemDist(c_1, c_2) = Relw(c_1) \cdot Relw(c_1) \cdot SemDensity(c_1, c_2) \quad (6)$$

The proximity measure is a Cauchy function :

$$Proximity\ (c_1, c_2) = \frac{1}{1 + SemDist\ (c_1, c_2)} \quad (7)$$

The contribution of these defined operators is that they give more importance to concepts which have a 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 weight of the terms are calculated according to their distribution in documents following three classical statistical measures, n-grams, Okapi and Okapi-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-ABR, Okapi-ABR and Okapi-Ngrams-ABR.

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 OKAPI separately . However, after the calculation of the NGRAM-OKAPI-ABR weighting these terms were better classified at the top of the rankings.

a) 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 removal of stopwords, 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 to add semantic relevance to this measure taking into account the semantic vicinity of extracted terms by combining N-gram with a kernel function, the formula becomes:

$$NGRAM\ \ -ABR\ (t, T) = NGRAM\ _0(t, T) + \sum_{i=1}^{n} NGRAM\ \ (t_i, T) \cdot\ \ (SemDist\ \ (t_i, t)) \quad (8)$$

Or simply,

$$NGRAM\ -ABR(t, T) = NGRAM\ _0(t, T) + \sum_{i=1}^{n} NGRAM\ (t_i, T) \cdot Proximity\ (t_i, t) \quad (9)$$

With $SemDist\ (t, t_i) < threshold$, $Proximity(t, t_i) < threshold$

$t_i \in T_n$ as T_n the 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).

$NGRAM_0(t,T)$ the initial value of the occurrence frequency of trigrams t in the theme T calculated by classical n-grams.

b) Radial basis OKAPI

We have opted for the model of Okapi proposed by [30] by introducing a semantic extension. To do this, the function $\varphi(SemDist)$ calculates the degree of relevance for each term located in its semantic vicinity (zone of influence). We indicate by $Proximity(t, t_i)$ the set of terms semantically related to t. A similarity threshold is needed to characterize all of its elements. We set a similarity threshold for the value of which corresponds to the degree of similarity between t and the concept representing the theme where it appears (the term is accepted if it is in the influence zone of the central word defined by the radial basis function). The relation becomes:

$$OKAPI - ABR(t, T) = OKAPI_0(t, T) + \sum_{i=1}^{n} OKAPI(t_i, T) \cdot\ \ (SemDist(t_i, t)) \quad (10)$$

Or simply,

$$OKAPI - ABR(t, T) = OKAPI_0(t, T) + \sum_{i=1}^{n} OKAPI(t_i, T) \cdot Proximity(t_i, t) \quad (11)$$

With $SemDist\ (t, t_i) < threshold$, $Proximity(t, t_i) < threshold$

$t_i \in T_n$ as T_n the 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). $OKAPI_0(t,T)$. The initial value of the weight of term t (root) to the theme T calculated by the classical Okapi.

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 that the expression is:

$$\text{Dice}\ (P_i, P_j) = \frac{2\,|P_i \wedge P_j|}{||P_i| + |P_j||} \quad (12)$$

Where,| P_i | is the number of terms in the profile Pi (vector representing the document i) and| $P_i \wedge P_j$ | is the number of terms of intersection between the two profiles P_i and P_j.

V. 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 I. Results of OKAPI and OKAPI- RBF

Method	Corpus	Precision	Recall
Okapi	sport	0.82	0,75
	politic	0.79	0,67
	finance & economics	0.73	0,65
Okapi-RBF	sport	0.91	0,81
	politic	0.81	0,70
	finance & economics	0.76	0,69

TABLE II. Results of NGRAM and NGRAM- RBF

Method	Corpus	Precision	Recall
Okapi	sport	0.78	0,68
	politic	0.65	0,50
	finance & economics	0.66	0,49
Okapi-RBF	sport	0.81	0,67
	politic	0.60	0,53
	finance & economics	0.62	0,51

TABLE III. Results of NGRAM-OKAPI and NGRAM-OKAPI- RBF

Method	Corpus	Precision	Recall
NGram-Okapi	sport	0.89	0,80
	politic	0.87	0,77
	finance & economics	0.57	0,60
NGram-Okapi-RBF	sport	0.92	0,80
	politic	0.83	0,79
	finance & economics	0.78	0,63

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.

VI. Discussion and Conclusion

The preceding tables present the experimental results that we obtained on the indexing and classification of an Arabic corpus. We have chosen to apply statistical measures *OKAPI* 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 modelling.

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 *OKAPI* 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 *OKAPI*, however, they have shown less performance at the level of n-grams, Although this method is widely invested on a number of 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 redo a second filtering after extracting of the n-grams list, but that appears unnecessary since we will lose more semantic information subsequently degrades the precision. However, these measures may also be combined between them as in the case of n-grams- *OKAPI* hybridization which has improved results compared with n-grams or N-grams-ABR 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 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 a track interesting 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.

REFERENCES

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

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

[3] Al-Shalabi R., Kanaan G., and Gharaibeh M., "Arabic Text Categorization Using kNN Algorithm", 6th International Conference on Advanced Information Management and Service (IMS), 2010, Seoul.

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

[5] Alsaleem S., "Automated Arabic Text Categorization Using SVM and NB", International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011.

[6] Bawaneh M. J., Alkoffash M. S., and Al Rabea A. I., "Arabic Text Classification using K-NN and Naive Bayes", Journal of Computer Science 4 (7): 600-605, 2008.

[7] Benkhalifa M. , Mouradi A., and Bouyakhf H., 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.

[8] Blei D. M., Ng A. Y., and Jordan M. I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

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

[10] Daimi K., 2001. Identifying Syntactic Ambiguities in Single-Parse Arabic Sentence. In Computer and humanities 35:333:349.

[11] Deerwester S., Dumais S.T., Furnas G.W., Landauer T. K. and Hrashman R.,1990. Indexing by latent semantic analysis. Journal of th american society for information science, 41(6) :391–407, 1990. 22, 24.

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

[13] Duwairi R.M., Machine Learning for Arabic Text Categorization, Journal of American society for Information Science and Technology, Vol. 57, No. 8, 2006, pp.1005-1010.

[14] El-Halees A.M., Arabic Text Classification Using Maximum Entropy, The Islamic University Journal, Vol. 15, No. 1, 2007, pp 157-167.

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

[16] Gharib T. F., Habib M. B., and Fayed Z. T., (2009) Arabic Text Classification Using Support Vector Machines. International Journal of Computers and Their Applications, 16 (4). pp. 192-199.

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

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

[19] Khoja S. and Garside S., 1999. Stemming Arabic Text. Computing Department, Lancaster University, Lancaster, U.K. http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, September 22, 1999.

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

[21] Larkey L. S., Ballesteros L. and Connell M., 2002. Improving Stemming for Arabic Information Retrieval: Light Stemming and Cooccurrence Analysis, In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, August 2002, pp. 275-282.

[22] Li, Y. H. and Jain, A. K. 1998. Classification of text documents. Comput. J. 41, 8, 537–546.

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

[24] Mesleh A.M., CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System, Journal of Computer Science, Vol. 3, No. 6, 2007, pp. 430-435.

[25] Mesleh A., "Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study", 12th WSEAS Int. Conf. on Applied Mathematics, Cairo, Egypt, December 29-31, 2007.

[26] Quillian M. R., 1968. Semantic memory. Semantic information processing, 1968. 65.

[27] Robertson S. E. and Walker S., 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proceedings of SIGIR 1994, pages 232–241, 1994. 18, 51.

[28] Robertson S. E., Payne A., Beaulieu M.M., Gatford M. and Walker S., 1995. Okapi at trec-4. In NIST Special Publication 500-236 : The Fourth Text REtrieval Conference (TREC-4), pages 73–96, 1995. 18, 21, 51.

[29] Robertson S.E. and Spark J. K., 1997. Simple proven approches to text retrieval. Technical report, City University, Department of Information Science, 1997. 18, 21, 49, 51.

[30] Robertson S., Walker S., and Beaulieu M., 2000. Experimentation as a way of life : Okapi at TREC, Information Processing and Management, vol. 36, no 1,2000,pp. 95-108.

[31] Schapire, R. E. AND Singer, Y. 2000. BoosTexter: a boosting-based system for text categorization. Mach. Learn. 39, 2/3, 135–168.

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

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

[34] Sawaf H., Zaplo J., and Ney H., Statistical Classification Methods for Arabic News Articles, Paper presented at the Arabic Natural Language Processing Workshop (ACL2001), Toulouse, France. (Retrieved from Arabic NLP Workshop at ACL/EACL 2001 website: http://www.elsnet.org/acl2001-arabic.html).

[35] Sebastiani F., Sperduti A. and Valdambrini N., 2000. An improved boosting algorithm and its application to automated text categorization. Technical report, Paris, France.

[36] Syiam M. , Fayed Z., and Habib M., An Intelligent System for Arabic Text Categorization, International Journal of Intelligent Computing and Information Sciences, Vol.6, No.1, 2006, pp. 1-19.

[37] Thabtah F., Hadi W., Al-shammare G. (2008) VSMs with K-Nearest Neighbour to Categorise Arabic Text Data. In The World Congress on Engineering and Computer Science 2008. (pp.778-781), 22-44 October 2008. San Francisco, USA.

[38] Wei G., Gao X., and Wu S., "Study of Text Classification Methods for Data Sets With Huge Features", 2nd International Conference on Industrial and Information Systems, 2010.

[39] Yang Y. and Chute C. G., An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems, 12:252-277, 1994.

[40] Zubi Z. S., "Using Some Web Content Mining Techniques for Arabic Text Classification", RECENT ADVANCES on DATA NETWORKS, COMMUNICATIONS, COMPUTERS, 2009.