ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 9, No. 2, 2014, *pp*.141-153

A Kernel Hybridization NGram-Okapi for Indexing and Classification of Arabic Documents

Taher Zaki [1,2], Driss Mammass [1] , Abdellatif Ennaji [2] and Stéphane Nicolas [2 +]

[1] IRF-SIC Laboratory, Ibn Zohr University Agadir Morocco

[2] Affiliation LITIS Laboratory EA 4108, University of Rouen France

(*Received August 01, 2013, accepted December 14, 2013*)

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 documents, classification, indexing, kernel function, n-grams, okapi.

1. Introduction

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 [2][16].

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. However, the task is not 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.

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 for which [9][22] explains the lack of research methods in the field of treatment of Arabic.

A set of statistical models for classification and machine learning techniques have been applied to text classification : the K nearest neighbor [13][1], the decision tree [17], the Bayesian model [10], SVM model (Support Vector Machines) [11][3], SVM combined with Chi-2 for feature extraction [18][19][20], neural networks [12], Maximum Entropy[23], the distances-based classifiers [12][20][13], the knowledge-based classifiers as WordNet [7].

3. Architecture of the proposed system

3.1. Process diagram

+ Corresponding author. Tel.: +212- 5 28 22 02 67; fax: +212- 5 28 22 01 00.
E-mail address: tah_zaki@yahoo.fr.

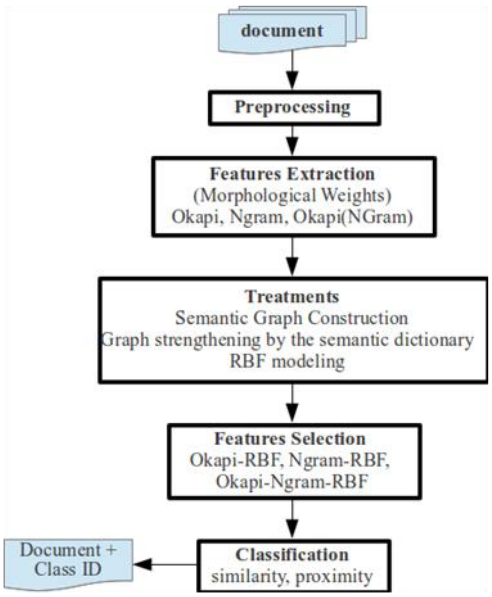

Fig. 1: The stages of the proposed indexing system.

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 "آ, أ, إ" to "ا" and "ى, ي, ء" to "ئ" and "ؤ" to "و"

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 [32][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 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

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

DOI https://doi.org/2024-JICS-22591 | Generated on 2025-04-15 09:07:07

extraction have been used. The Stemming of the terms is operated using the Khoja stemmer [19] 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. Weighting of descriptors by Okapi

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 dj, that does not necessarily mean it is n times more important than in a document dk 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 [21]:

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

where $tf(t,d_j)$ is the term frequency t in document dj and idf (t) which is the number of documents having at least one occurrence of the term t, NDL (d_j) is the normalized length of d_j, 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]. Experiments have shown that a reasonable value is b = 0.75.

4. Semantic indexing with kernel function

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.

4.1. Calculation of the new weights

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. 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. 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.
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 Figures 2 and 3 below:

رياضة:تدريب:تمارين:تمرين: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:الاحترافي:بطولة:كأس:ابطال:كرة:منتخب .....

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

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:.....

Fig. 2: Example of Arabic semantic dictionary of the sport theme

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

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.

4.2.3 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.

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. Figures 4 and 5 illustrate this process for an example of the theme " finance and economy".

3 http://www.almaany.com

4 http://ar.thefreedictionary.com/

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. 4: 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. 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 a 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.

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 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 6, 7).

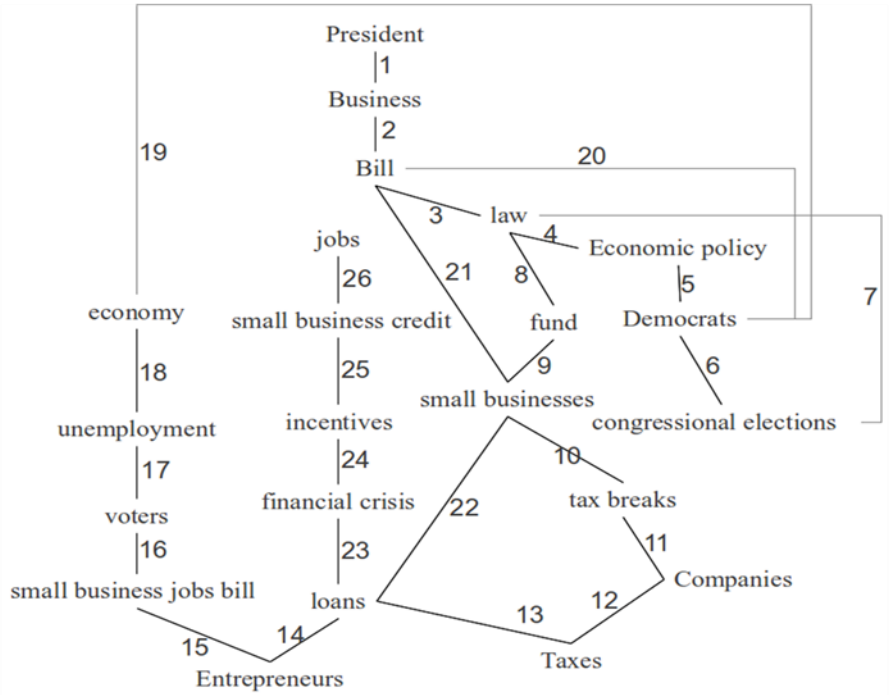

Fig. 6: Semantic graph extracted from the document

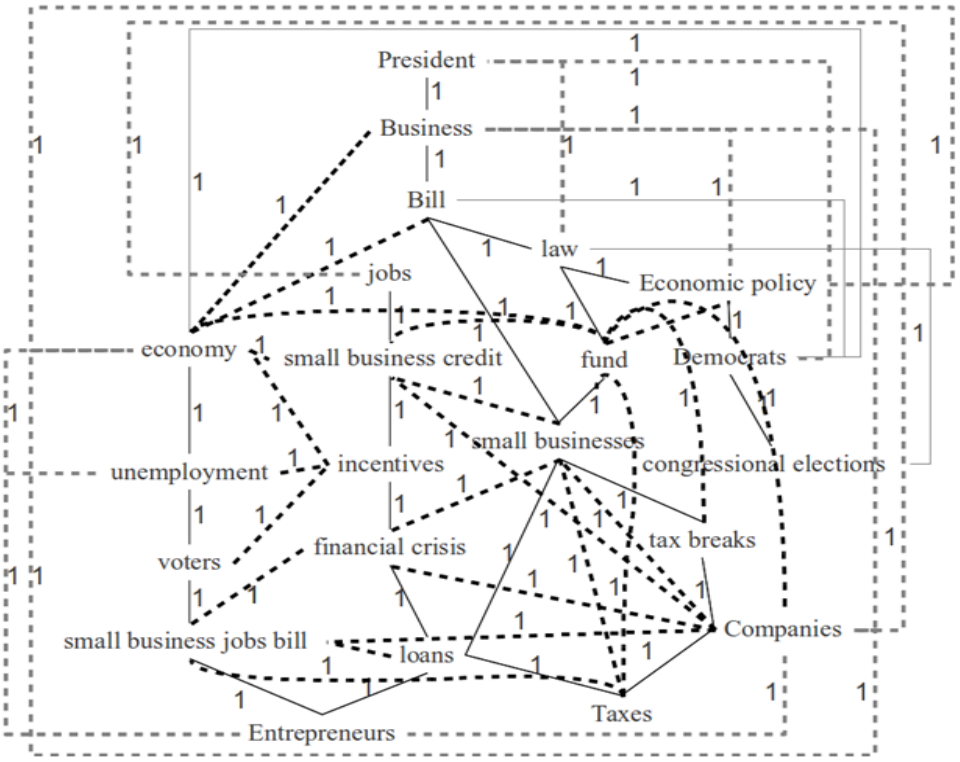

Fig. 7: Strengthening of the graph by semantic connections extracted from the auxiliary dictionary

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

Second, this graph supports the presence of compound words. These words are juxtapositions of two free lexemes to a third form 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 to the problem by including the compound terms deemed relevant and informational in the semantic dictionary.

5. The new weights with radial basis

The NGRAM-OKAPI with radial basis function (NGRAM-OKAPI-RBF) relies on 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:

$$g(x) = w_0 + \sum_{i}^{n} w_i \, \varphi(d(x, \sup_i)) \; . \tag{2}$$

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 :

$$\varphi(x) = \frac{1}{1+x} \; . \tag{3}$$

we define two new operators:

$$\mathrm{Re}\,lw(c) = \frac{\deg ree\,(c)}{total\ \ number\ of\ \ concepts} \; . \tag{4}$$

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.

$$SemDensity\,(c_1, c_2) = \frac{MinCost\ \ (c_1, c_2)}{Minimal\ \ Cost\ of\ \ the\ Spanning\ Tree} \; . \tag{5}$$

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 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 (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:

$$SemDist\,(c_1, c_2) = \mathrm{Re}\,lw\,(c_1) \times \mathrm{Re}\,lw\,(c_2) \times SemDensity\,(c_1, c_2) \; . \tag{6}$$

The proximity measure is a Cauchy function :

$$\Pr oximity\,(c_1, c_2) = \frac{1}{1 + SemDist\,(c_1, c_2)}\ . \tag{7}$$

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-RBF, OKAPI-ABR and OKAPI-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 OKAPI separately . However, after the calculation of the NGRAM-OKAPI-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:

$$NGRAM - RBF\,(t,T) = NGRAM_0\ (t,T) + \sum_i^n NGRAM\,(t_i, T) \cdot \varphi(SemDist\,(t_i, T)) \cdot \tag{8}$$

Or simply

$$NGRAM - RBF\,(t,T) = NGRAM_0\ (t,T) + \sum_i^n NGRAM\,(t_i, T) \cdot \Pr oximity\,(t_i, T) \cdot \tag{9}$$

With $Proximity(t, t_i) < threshold$

$t_i \in T_n$ as T_n 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).

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

5.2. Radial basis OKAPI

We proceed to calculate the terms OKAPI 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 OKAPI and accepting only the terms within the zone of influence. Noted that weight OKAPI -RBF (t) is calculated as follows:

$$\text{OKAPI} - RBF\,(t,T) = \text{OKAPI}_0\ (t,T) + \sum_i^n \text{OKAPI}(t_i, T) \cdot \varphi(SemDist\,(t_i, T)) \cdot \tag{10}$$

Or simply,

$$\text{OKAPI}-RBF(t,T)=\text{OKAPI}_0(t,T)+\sum_i^n \text{OKAPI}(t_i,T)\cdot \text{Pr}oximity(t_i,T)\ . \tag{11}$$

With $Proximity(t,t_i) < threshold$

$t_i \in T_n$ as T_n 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).

$OKAPI_0(t,T)$ The initial value of the weight of term t (root) to the theme T calculated by the classical OKAPI.

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:

$$Dice(P_i,P_j)=\frac{2\left|P_i\wedge P_j\right|}{\left|\left|P_i\right|+\left|P_j\right|\right|}\ . \tag{12}$$

Where,

$|P_i|$ is the number of terms in the profile Pi (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. OKAPI-RBF

Inputs:

$D=\{D_{sp},D_{ec},D_{po}\}$ a textual database (whether pre-sorted or not) of documents representing three themes, sport , economy and politics.

$Dic=\{Dic_{sp},Dic_{ec},Dic_{po}\}$ semantic dictionaries of themes, sport , economy and politics.

Output:

$D'=\{D'_{sp},D'_{ec},D'_{po}\}$ 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.

$\forall d \in D$ do

2. Remove punctuation, stopwords, dates, and vowels

3. For each word of d, calculate the OKAPI 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 OKAPI and N-grams measures).

7. Construction of the semantic graph

$G=\{V,E\}$ where ,

$V=\{t_i,\ \ t_i\in d\}$

$E=\{(t_i,t_j),\ \ SemProx(t_i,t_j)\in\{\infty,1\}\}$

$\forall(t_i,t_j)\in E$, (t_i,t_j) is weighted by the semantic proximity $SemProx(t_i,t_j)$

$\forall t_i \in V$, build V_{t_i} the neighbors set of t_i

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

$\forall t_j \in \{ t_i \cup V_{t_i} \}$ calculate $SemDist(t_i, t_j)$ (see equations (9, 11))

with $SemDist(t, t_i) < threshold$

$$weight - RBF(t,T) = weight_0(t,T) + \sum_{i=1}^{n} weight(t_i,T) \cdot (SemDist(t_i,t))$$

with $weight \in \{NGRAM, \text{OKAPI}\}$

9. generation of new representation space d'

10. reduction of space d'

11. Add d' to D'

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 OKAPI measure. the extracts N-grams are also used for the graph construction as explained by the following algorithm:

Algorithm

inputs:

$D = \{D_{sp}, D_{ec}, D_{po}\}$ textual database of documents (pre-ordered or not)

$Dic = \{Dic_{sp}, Dic_{ec}, Dic_{po}\}$ semantic dictionaries

output:

$D' = \{D'_{sp}, D'_{ec}, D'_{po}\}$ indexed documents, weighted and classified

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

$\forall d \in D$ 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 $okapi(ngram_i, d)$ 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 $okapi(ngram)$).

7. Construction of the semantic graph

$G = \{V, E\}$ where ,

$V = \{t_i, \quad t_i \in d\}$, all words in d.

$\forall t_i \in V$ extracting the set Ng of corresponding n-grams

$$Ng = \{ n_k , n_k \quad subword\ of\ length\ n\ extracted\ \ from\ t_i \}$$

8. Passage from $G = \{V, E\}$ to $G' = \{V', E'\}$

$$V' = \{ x, x \in NG_d \quad set\ of\ n-grams\ extracted\ \ from\ d \}$$

$$E' = \{ \quad (x,y) \quad \in V' \times V' , \quad SemProx(x,y) \in \{\infty, 1\} \quad \}$$

$\forall (x,y) \in E'$, (x,y) is weighted by the semantic proximity $SemProx(x,y)$

$\forall x \in V'$, built V_x , the neighbors set of x

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

$\forall y \in \{x \cup V_x\}$ Calculate $Proximity(x,y)$ (see equations (8 , 10))

DOI https://doi.org/2024-JICS-22591 | Generated on 2025-04-15 09:07:07

with $Proximity(x,y) < threshold$

$$weight - RBF(n,T) = weight_0(n,T) + \sum_{i=1}^{m} weight(n_i,T) \cdot Proximity(n_i,n)$$

with $weight \in \{NGRAM - OKAPI\}$

10. generation of new representation space d'

11. reduction of space d'

12. Add d' to D'

7. 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 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 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- 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.

8. 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 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 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 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 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- OKAPI 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.

9. 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, Vol. 4, Amman, Jordan, April 5-7, 2006.

[2] M. Aljlayl, and O. Frieder, *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.

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

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

[5] 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.

[6] S. Deerwester, 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, 41(6) :391–407, 1990. 22, 24.

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

[8] 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.

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

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

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

[12] 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), 2011.

[13] 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, Vol. 4, Amman, Jordan, April 5-7, 2006.

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

[15] 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.

[16] L. S. Larkey, L. Ballesteros and M. Connell, *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.

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

[18] 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), Vol. 1, Barcelona, Spain, July, 22—25, 2007, pp. 235-240.

[19] 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.

[20] 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, December 29-31, 2007.

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

[22] 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, December, Cairo, Egypt, 2005, pp. 13-15.

[23] H. Sawaf , J. Zaplo, and H. Ney, *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).