A Semantic Proximity Based System of Arabic Text Indexation

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

[1] Ibn Zohr University, Agadir, Morrocco
tah_zaki@yahoo.fr,
mammass@univ-ibnzohr.ac.ma
[2] LITIS EA 4108, University of Rouen, France
Abdel.ennaji@univ-rouen.fr

Abstract. In this paper, we extended the vectorial model of Salton [9], [11], [12] and [14], by adapting the TF-IDF parameter by its combination with the Okapi formula for index terms extraction and evaluation of the in order to identify the relevant concepts which represent a document.Indeed, we have proposed a new measure TFIDF-ABR which takes in consideration the concept of semantic vicinity using a measure of similarity between terms by combining the calculation of TF-IDF-Okapi with a kernel approach (Radial Basis function).

This indexation approach allows a contextual and semantic research. In order to have a robust descriptor index, we used not only a semantic graph to highlight the semantic connections between terms, but also an auxiliary dictionary to increase the connectivity of the constructed graph and therefore the semantic weight of indexation words.

Keywords: Document indexation, semantic graph, semantic vicinity, dictionary, kernel function, okapi formula, similarity, TF-IDF, vectorial model.

1 Introduction

The explosive growth of textual information, in particular published and easily accessible on the world requires the implementation of useful techniques for information extraction from forms that appear in large corpus of texts.

However, it has become essential to design methods making it possible to exploit at the same time the structure and the textual contents of this text corpus.

The goal of indexing is to identify, to locate and retrieve easily some information in a set of documents. Generally, we use indexation systems for information research, but, also for comparing and classifying documents, for keywords extraction, for an automatic synthesis of documents, for calculating co-occurrences of words and so on.

Any document indexation process lead to a loss of some part of the initial information.

Some recent works are interested with redefining the task of information retrieval and indexation for semi-structured documents.

A. Elmoataz et al. (Eds.): ICISP 2010, LNCS 6134, pp. 419–427, 2010.

In this paper, we define a new formalism for the statistical processing of Arabic textual documents and we show how this formalism could be used for the processing of various problems including indexation and classification.

Our work is to be positioned in the context of information retrieval called statistical learning that allows the development of generic methods used easily on different corpus. In section 2, briefly introduces indexing phase. Then, in Section 3 we describe some semantic resources used by our approach. In sections 4 and 5, our framework is presented. Next, we present experimental results in Section 6. Finally, we conclude and discuss some perspectives of this work in section 7.

2 Indexing Phase

One way for tackling the problem of document indexation and classification may be to take into account some explicit informations around the text like the structure and the distribution of words, as well as implicit semantic informations.

2.1 Weighting of Index Units

The goal is to find the terms that best characterize the contents of a document. The easiest way to calculate the weight of a term is to calculate its occurence frequency because a term which often appears in a document can characterize best its contents. Several weighting functions of terms have been proposed [15]. We are interested by the TF-IDF (term frequency - inverse document frequency) used in vectorial model which is used with some adaptations in this work.

The dimensions retained to calculate the weight of a term are:

- A local weight that determines the importance of a given term in the document. It is usually represented by its frequency (TF).
- An overall weight that determines the distribution of the term in the document database. It is generally represented by the inverse of the frequency of documents containing the term (idf).

$$a_{ij} = tf(i,j).idf(i) = tf(i,j)\log\left(\frac{N}{N_i}\right). \quad (1)$$

Where tf(i,j) is the frequency or the number of times the term t_i appears in document d_j and idf(i) is the inverse document frequency or the logarithm of the ratio between the number N of documents in the corpus and N_i the number of documents that contain the term t_i. This indexing scheme gives more important weight to words that appear with high frequency in few documents. The underlying idea is that these words help to discriminate between texts with different subject.

The TFIDF has two fundamental limits. First, the dependence of the term frequency is too important. If a word appears twice in a document d_j, that does not necessarily mean that it has twice more important value that in a document where it appears only once. Second, the weights of long documents are higher because they

contain more words and the term frequencies tend to increase. To address these problems, a new indexing technique has been proposed (the Okapi Formula) [8]:

$$a_{ij} = \frac{tf(i,j).idf(i)}{[(1-b)+b \cdot \mathrm{NDL}(d_j)]+f(i,j)}. \quad (2)$$

Where $\mathrm{NDL}(d_j)$ is the standard length of d_j (number of words that it contains) divided by the average length of the documents in the corpus.

3 The Semantic Resources

3.1 The Auxiliary Semantic Dictionary

It is a hierarchical dictionary with standard vocabulary based on generic terms and on specific terms to a domain. Consequently, It provides the definitions, the relations between terms and their choice overriding the significances. The relations commonly expressed in such a dictionary are:

- The taxonomic relations (hierarchy).
- The equivalence relations (synonymy).
- The associate relations (relations of semantic proximity, close-in, connected to, etc…).

3.2 Taxonomy

We introduce a taxonomy as the organization of concepts linked by hierarchical relations [21].

3.3 Semantic Networks

The semantic networks [5], were initially introduced as a model of the human memory. A semantic network is an oriented and labelled graph (or, more precisely, multigraph). An arc connects a starting node (at least) to a node of arrival (at least). The relations go from the relations of proximity semantic to the relations part-of, cause-effect, parent-child…

The concepts are represented as nodes and their relations as arcs. The heritage of the properties by the connexions is materialized by an arc (kind-of) between the nodes. Different types of connexions can be mixed as well as concepts and instances. So, various developments have emerged and led to the definition of formal languages.

4 Semantic Indexing Based on Radial Basis Functions

Several studies have adapted the vectorial model by indexing directly concepts instead of terms. These approaches treat primarily synonymy by replacing the terms by their concepts. We consider the rich links between the terms by taking into account all the semantic types of relations (ontology of the field). This can solve the problem of synonymy and avoids also the complications caused by the other relations of specialization and generalization for example.

4.1 The Indexation and Classification Proposed System

Unlike existing methods, we do not use concepts only. Indeed, the terms are enriched if they are linked to the concepts or if they provide good semantic connectivity. It is important to note that during the search, we also find terms which are not linked to ontology.

We calculate the proximity between terms. Thus, we define a radial basis function (RBF) which associates to each term a region of influence characterized by the degree of semantic similarity and the relationship between the kernel term and its neighbors.

Rada and al. [6], were the first to suggest that the similarity in a semantic network can be calculated based on the taxonomic links “is-a”.

One of the most obvious ways to evaluate semantic similarity in taxonomy is to calculate the distance between nodes by the shortest path.

In [7], the idea was to calculate the paths as those linking each concept to its closest ancestor to the top of the ontology.

We are aware that the calculation of the similarity measure by restriction on the "is-a" is not always suitable, because the taxonomies are not always at the same level of granularity, some parts may be have a more important density than others. These problems can be solved by associating weights to the links, so we have chosen to take into consideration all types of relationships (conceptual problematic) and the distribution of words in documents (structural problematic).

However, the automatic calculation of degree of semantic relation is too difficult and many previous works has focused on the similarity measures, often based on known hierarchies (eg WordNet [1] and Wikipedia [3]).

We have adapted our system to support any kind of semantic relation such as synonymy, meronymy, hyponimy, taxonomy, antonomy, etc ... and we assign initially a unit weight for semantic links.

A semantic network is constructed in each phase to model the semantic relationships between terms. To avoid connectivity problems, we choose to build an auxiliary dictionary allowing to have a strong connection in the builded network and to grow the weight of the semantic descriptor terms thereafter.

In the next section, we define our TFIDF measure with radial basis function and we show how the weights of the indexing terms are enriched from the outputs of this measure.

4.2 Text Preprocessing

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

- Convert text files to UTF-16 encoding.
- Remove punctuation marks, diacritics, non-letters, stop words.
- Stemming by using the Khoja stemmer [4].

4.3 KNN Text Classification

In many application frameworks, this approach has shown interesting results. Indeed, the probability of error with a knn converges to the Bayes risk when the data learning quantities increases, regardless of k. However, it has several drawbacks. Its robustness depends directly on the relevance and quality of the learning set. Another drawback concern the access to the learning set which generally requires a large memory space and time consuming computing.

In this paper, we use the KNN classifier as reference in the experiments. In fact the KNN can be used with few amount of data what is a very interesting property in our context.

Instead of the euclidean distance, we chose the measure of relevance combined with the measure of Dice after reducing the learning set.

The text preprocessing phase is firstly applied to each document to be classified. Then the RBF Okapi-TFIDF profile is generated (RBF for Radial Basis Function). The RBF Okapi-TFIDF profile of each text document (document profile) is compared against to the profiles of all the documents in the learning classes (class profile) in terms of similarity.

The second measure used is the Dice measure of similarity:

$$Dice \quad (P_i, P_j) \quad = \quad \frac{2\left|P_i \wedge P_j\right|}{\left|\left|P_i\right| + \left|P_j\right|\right|}. \tag{3}$$

where $\left|P_i\right|$ is the number of elements in profile P_i.

$\left|P_i \wedge P_j\right|$ is the number of elements that are found in both P_i and P_j

4.4 The TF-IDF with Radial Basis Functions

The RBF TFIDF with radial basis functions is based on the determination of supports in the space of representation E. However, unlike the traditional TFIDF, those can correspond to fictitious forms which are a combination of the values of traditional TFIDF. We will call them prototype. They are associated with an area of influence defined by a distance (Euclidean, Mahalanobis ...) and a radial basis function (Gaussian, bell ...). The output discriminant function g of the RBF-TFIDF is defined from the distance form the shape in entry to each of the prototype and of the linear combination of the corresponding radial basis functions:

$$g(X) = w_0 + \sum_{i=1}^{N} w_i \, \phi(d(X, \sup_i)) \; . \tag{4}$$

Where d(X, sup_i) is the distance between entry X and the support sup_i, {w_0, ...,w_N} are the weights of the combination and ϕ the radial basis function.

The Learning of this type of model can be done in one or two steps. In the first case, a gradient type method is used to adjust all the parameters by minimizing an objective function based on a criterion such as least squares. In the second case, a first step is to determine the parameters associated with radial basis functions (position of prototypes and areas of influence). To determine the centers, methods of unsupervised

classification are often used. In a second step, the weights of output layer can be learned by different methods such as inverse or pseudo-gradient descent.

The RBF-TFIDF has several advantages in the case of learning in two stages. For example, the separated learning of the radial basis functions and their combination makes learning faster, simple and avoids the problems of local minima (local and global relevance), the prototypes of RBF-TFIDF represent the repartition of the examples in the space of representation E (terms).

Moreover, management of multi-classes problems is easier with RBF-TFIDF. We will see in the following section that RBF-TFIDF is very similar under certain conditions to the Systems of Fuzzy Inference.

Modeling RBF-TFIDF is both discriminant and intrinsic. In fact, the layer of radial basis functions corresponds to an intrinsic description of the learning data and the combination layer output seeks to discriminate the different classes.

In this paper, we use RBF-TFIDF with learning in two stages. A first step is to determine the parameters associated with radial basis functions (position of prototypes and areas of influence), usually unsupervised classification methods are often used. In a second step, the weights of output layer can be learned by different methods such as inverse or pseudo-gradient descent. The radial basis function is of Cauchy type:

$$\phi(d) = \frac{1}{1+d}. \tag{5}$$

We have defined two new operators:

a) the relational weight :

$$\text{WeightRel}(t) = \frac{\text{degree}(t)}{\text{total number of concepts}}. \tag{6}$$

b) Semantic density:

$$SemDensity(c_1, c_2) = \frac{Dist(c_1, c_2)}{\text{recovering tree with minimal cost}}. \tag{7}$$

Thus the semantic distance between two concepts

$$\text{DistSem}(c_1, c_2) = \text{WeightRel}(c_1) * \text{WeightRel}(c_2) * SemDensity(c_1, c_2). \tag{8}$$

The proximity measure is a Cauchy function:

$$\text{Proximity}(c_1, c_2) = \frac{1}{1 + DistSem(c_1, c_2)}. \tag{9}$$

degree(t): The number of incoming and outgoing edges of node t.

Dist(c_1,c_2): The minimum distance between c_1 and c_2, calculated by Dijkstra's algorithm [2] and [16], applied to the semantic network thus built starting from document.

For the indexing phase, we will see later how the weight of the index descriptors are generated by the radial basis measures admitting like parameter a semantic distance.

5 New Weights of the Descriptors Indexes

The documents are represented by sets of vectors of terms. The weights of terms are calculated according to their distribution in the documents. The weight of a term is enriched by the conceptual similarities of words co-occurring in the same topic.

We proceed by calculating the TFIDF terms for the set of themes of the learning basis in order to deduce their overall relevance and then we calculate their local relevance by our radial basis function combined with the traditional TFIDF and accepting only the terms located in the zone of influence. This weight noted RBF-TFIDF (t) is calculated by the formula:

$$\text{RBF-TFIDF}(t,\text{theme}) = \text{TFIDF}(t,\text{theme}) + \sum_{i=1}^{n} \text{TFIDF}(t_i,\text{theme}) * \phi(\text{Proximity}(t,t_i)). \quad (10)$$

With $\phi(\text{Proximity}(t,t_i)) < \text{threshold}$

$t_i \in$ belongs to the set of n terms in the topic.

The threshold: is a value which fixes the proximity at a certain vicinity (zone of semantic influence of the term t), we initially fix it in the proximity between the concept of t and the concept context (concept which represents the topic).

5.1 Extension of the Okapi-Formula

To avoid the inconvenients of the TFIDF measure, and in order to make it more robust, we opted for the Okapi model proposed by [8], with adding a semantic extension.

For this, the function $\phi(d)$ calculates the degree of relevance for each term on its semantic proximity level (zone of influence). The new formula is:

$$a_{ij} = \frac{tf(i,j).idf(i)}{[(1-b)+b \cdot \text{NDL}(d_j)]+f(i,j)} \cdot \phi(d_j). \quad (11)$$

Or more simply,

$$a_{ij} = \frac{TFIDFABR(i,j)}{[(1-b)+b \cdot \text{NDL}(d_j)]+f(i,j)}. \quad (12)$$

$\phi(d_j)$ is the set of the terms d_j close to t_i semantically. A similarity threshold is necessary to characterize all of these elements. We set a threshold of similarity for the value of Proximity (t_i, t) which corresponds to the degree of similarity between t and the concept of the theme where it appears (the term is accepted if it is in the influence zone of term kernel defined by the radial basis function f).

6 Results

For the learning phase, we worked on a very reduced database (initial corpus) of labeled documents representing the classes (sport, policy, economy and finance) that we seeks to discriminate or to learn and this is the strong point of our measurement.

More this base is discriminating and representative, more our method is efficient and more the results are better.

For the phase of test we worked on a corpus of 200 documents of press from the Associated Press (AP) witch is a very rich and varied base of 2246 English documents [20].

And for the Arab documents we worked on a corpus of 300 documents of Arab electronic presses ([17], [18] and [19]).

Table 1. Table of the results of the experimentation

Corpus	Method	Classifier	Recall	Precision	accuracy (%)
English	TFIDF-ABR	kppv	0.89	0.92	90.5
Arabic	TFIDF-Okapi-ABR	kppv	0.98	0.98	98.79

7 Conclusion

After validation of the prototype, our method has shown robustness and adaptability for both Arabic and English corpus. In addition, the results of the indexing contain exactly the required keywords sorted by their relevance. We set a threshold for semantic enrichment which leads to retain a few intruders words away from those sought.

Many elements remain to be evaluated, in particular the addition of an algorithm of clarification and disambiguation.

The presence of the complex concepts can also prove an interesting track by the fact that the more long concepts are often less ambiguous.

Acknowledgment

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. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of semantic distance. Computational Linguistics 32(1), 13–47 (2006)
2. Dijkstra, E.W.: A short introduction to the art of programming, contenant l’article original décrivant l’algorithme de Dijkstra, pp. 67–73
3. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness singWikipedia-based Explicit Semantic Analysis. In: Proc. IJCAI’07, pp. 1606–1611 (2007)
4. Khoja, S., Garside, S.: Stemming Arabic Text. Computing Department. Lancaster University, Lancaster, September 22 (1999), `http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps`
5. Quillian, M.R.: Semantic memory. In: Semantic Information Processing (1968)

6. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics 19(1), 17–30 (1989)
7. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
8. Robertson, S., Walker, S., Beaulieu, M.: Experimentation as a way of life: Okapi at TREC. Information Processing and Management 36(1), 95–108 (2000)
9. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523 (1988)
10. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
11. Salton, G., Yang, C.S., Yu, C.T.: A theory of term importance in automatic text analysis. Journal of the American Society for Information Science and Technology 26(1), 33–44 (1975)
12. Salton, G., Fox, E.A., Wu, H.: Extended boolean information retrieval. Communications of the ACM, 1022–1036 (1983)
13. Salton, G., Singhal, A., Buckley, C., Mitra, M.: Automatic text decomposition using text segments and text themes. In: UK Conference on Hypertext, pp. 53–65 (1996)
14. Salton, G.: The SMART retrieval system: experiments in automatic document processing. Prentice-Hall, Englewood Cliffs (1971)
15. Seydoux, F., Rajman, M., Chappelier, J.C.: Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire. Ph.D. thesis (2006)
16. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction à l'algorithmique, version (en). section 24.3, Dijkstra's algorithm, deuxième edn., pp. 595–601. MIT Press, McGraw-Hill (2001) ISBN 0-262-03293-7
17. Al Jazeera: `http://www.aljazeera.net/`
18. Al charq Al awsat, `http://www.aawsat.com/`
19. Al ahdat Al maghrebiya, `http://www.almaghribia.ma/`
20. Associated Press, `http://www.cs.princeton.edu/~blei/lda-c/ap.tgz`
21. Wikipedia, `http://en.wikipedia.org/`