A Fuzzy Radial Basis Model for Arabic Documents Classification

Taher Zaki[1(✉)], Mohamed Salim El Bazzi[1], Driss Mammass[1], and Abdellatif Ennaji[2]

[1] IRFSIC Laboratory, Faculty of Science, Ibn Zohr University, PB 8106, Agadir, Morocco
{t.zaki,mammass}@uiz.ac.ma, elbazzi@yahoo.fr
[2] LITIS Laboratory EA 4108, University of Rouen, 76000 Rouen, France
abdel.ennaji@univ-rouen.fr

Abstract. In this paper, we bring an improvement to the classical fuzzy model of classification by implementing a new approach which based on radial basis functions for the Arabic documents classification. This approach takes into account the concept of semantic vicinity by calculating of the similarity degree between terms in relation to the documents. We combine the calculation of the relevance of these terms (using NEAR operator) with a radial basis function to identify the relevant documents to the query. The use of linguistic resources namely semantic graphs and semantic dictionaries (specifically created for the studied domain) significantly improves the process of classification.

Preliminary and promising results are shown on a Arabic press database which show very good performance compared to the literature.

Keywords: Arabic language · Classical fuzzy model
Fuzzy classification · Radial basis function · Relevance
Semantic dictionary · Semantic graph · Semantic vicinity · Similarity

1 Introduction

The Arabic language is considered as a difficult language to master in the field of automatic language processing, given its morphological and syntactic properties [1,9].

The information retrieval in Arabic is a scope by excellence of the similarity concept and semantic proximity. Indeed, the problems of synonymy (antonymy, polysemy, hypernymy, meronymy...) generate a lot of ambiguities in the choice of descriptor words and key words, therefore the process of indexing and research becomes difficult to complete [3,7].

However it is important to spend at a semantic level, in order to avoid the problems of syntax and comparison term-by-term, hence the requirement to find methods that can assign to the words the correct senses from the context [4,5,11,15].

A. Mansouri et al. (Eds.): ICISP 2018, LNCS 10884, pp. 153–162, 2018.
https://doi.org/10.1007/978-3-319-94211-7_18

An art state on local semantic similarity measures and global algorithms for lexical disambiguation based on the knowledges is detailed in [6,14].

The use of a radial based modeling would be a good solution that consists, after obtaining the descriptors according to the relevance calculation combined with kernel functions, to use the thesaurus, semantic graphs and semantic dictionaries to improve the information retrieval process.

The rest of the paper is organized as follows: in Sect. 2 the concept of the classical fuzzy model is presented, Sect. 3 describes the text preprocessing phase and Sect. 4 gives detailed description of the classification procedure. Section 5 presents the classification results.

2 Proximity Functions

2.1 Binary Proximity

The Boolean systems that implement the $NEAR$ operator implicitly use the notion of proximity in their process. The $NEAR$ operator behaves like the AND operator with an additional constraint on the positions of occurrences of the terms concerned specifying a maximum distance between two terms A and B of the query q. For example, if we regard in q, A NEAR 7 B, a system implementing the NEAR operator evaluates this request to the value true if and only if at least one occurrence of the term A is less than 7 words (distance of 7 steps) of at least one occurrence of the word B.

2.2 Fuzzy Proximity

In the models based on fuzzy logic, each term t is associated to an influence function defined on $\mathbb{R}$, bounded support, taking values in $[0, 1]$, symmetric, increasing on $\mathbb{R}-$ and decreasing on $\mathbb{R}+$ denoted reflecting the degree of belonging document corresponding to the fuzzy set of the term t:

$$\begin{aligned} \mu_t \colon D &\longrightarrow [0,1] \\ d &\longmapsto \mu_t(d) \end{aligned} \tag{1}$$

Several types of functions (Gaussian, rectangular functions, features Hanning, triangular, etc. ...) can be used.

The fuzzy approach makes the notion of proximity fuzzy by assigning a fuzzy interpretation of the $NEAR$ operator. Indeed, each document is modelled as a finite sequence whose length is equal to l of the text terms T,$(t_0, t_1, \ldots, t_{l-1}) \in T^l$ i.e., a function whose definition domain is an interval of N starting at 0. $d^{-1}(t)$ refers to the position set taken by t in document d. For example, If we look for A near B, we give a proximity local value to the query $NEAR(A, B)$ in document d by:

$$\mu_{NEAR(A,B)}(d) = max_{\substack{i \in d^{-1}(A) \\ j \in d^{-1}(B)}} (max(\frac{k - |x - i|}{k}, 0)) \tag{2}$$

The parameter k is integer according to the evaluation context. For example, a value of $k = 5$ evaluates the proximity in the expression case while $k = 100$ translated proximity in a paragraph context and so on.

The value we attribute to μ_t is related to the distance between the two closest occurrences of the two terms A and B in the document. The maximum value is reached when the value of $|j - i|$ is minimum, i.e. equal to 1 because A and B may not appear in the same position. Consequently, we necessarily have $j \neq i$. Therefore, the minimum value is reached when there is an instance of A that is near a B instance in the document. For more details see [10].

2.3 Local Relevance of a Term Relative to a Document

To compare a term and a document, the function μ_t^d calculates the degree of relevance for each term t of the query q in all possible positions x in d. The positions x are defined by positive integers as well as by negative ones since the influence of terms extends either side of the their occurrence positions which overflow either before the start of the document or after it has ended.

$$\mu_t^d(x) = max_{i \in d^{-1}(t)}(max(\frac{k - |x - i|}{k}, 0)) \tag{3}$$

2.4 Relevance of a Query in Relation to a Document

The Relevance to the document is generalized in a natural way by an aggregation of the results obtained in all possible positions.

$$score(r, d) = \sum_{x \in [0, N-1]} \mu_r^d(x) \tag{4}$$

Thus, the similarity is obtained by the normalization of all scores by the cardinality of the fuzzy set d^{-1}.

$$Sim(r, d) = \frac{\sum_{x \in [0, N-1]} \mu_r^d(x)}{N} \tag{5}$$

The choice of terms is made simply from a correspondence according to the form of the keywords (lemmas or stems) of the document.

3 System Process

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 (Figs. 2 and 3). 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", "ىء ، يء" to "ئ" and "ؤ" to "و".
- Stemming the remaining terms is performed using the Khoja stemmer [8] for Arabic documents.

Subsequently, we proceed to the step of documents representation. This phase consists of eliminating the terms deemed insignificant and out of the considered fields. Then we distinguish between the terms "descriptors" and "equivalent". At the end of this phase, there is a graduated scale (axis) vector whose points correspond to the positions of descriptors and their equivalent terms that will be used by the fuzzy classifier to assign the corresponding category.

3.1 Weighting of Terms

Unlike classical models which are based on a vector representation whose features are the frequencies of appearance in documents, or any other statistical measures that refer to this modeling. The fuzzy model calculates the degree of belonging of a term or a query to a document. The result is a vector whose characteristics are the local semantic relevance of the terms.

We made an extension to the model of Mercier and Beigbeder [10] using a radial basis modeling to take into account the semantics vicinity of the terms that seems absent in this model, knowing that a term which has a semantically rich vicinity in a document is often relevant to characterize its content.

Starting from this idea, we have proposed a new measure of relevance based on the classical model that holds significantly the close proximity of the terms concept.

3.2 Semantic Resources

3.2.1 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 (semantic proximity, close to, related to, etc.).

The dictionary is initially constructed manually based on the words found in the training set combined with a set of dictionaries available on the web as "Almaany[1]" and "the free dictionary[2]". 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, the built dictionary is shown in Fig. 1 below:

[1] https://www.almaany.com.
[2] http://ar.thefreedictionary.com/.

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. 1. Example of Arabic semantic dictionary of the sport theme.

3.2.2 Semantic Networks

A semantic network [12] 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 [13], etc.

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.

3.2.3 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 "Sport".

كرة قدم

غياب ميسي وإنييستا عن مواجهة سرقسطة

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

المصدر: أ ف ب

Fig. 2. Initial text: theme "Sport"

كرة قدم

غياب ميسي إنييستا مواجهة سرقسطة

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

Fig. 3. 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) terms. 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

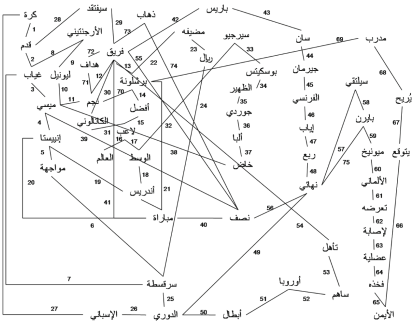

Fig. 4. Semantic graph extracted from the document.

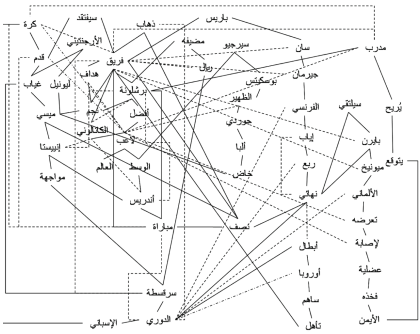

Fig. 5. Strengthening of the graph by semantic connections extracted from the auxiliary dictionary.

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.

The constructed graph (Figs. 4 and 5) 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, etc.).

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.

4 Semantic Classification Based on Radial Basis Function

4.1 The Radial Basis Fuzzy Proximity

The discriminating function g of RBF fuzzy proximity with one output is defined by the distance between the input form of each prototype and the linear combination of the corresponding radial basis functions:

$$g(x) = w_0 + \sum_{i=1}^{N} w_i \Phi(d(x, sup_i)) \tag{6}$$

While $d(x, sup_i)$ is the distance between the input x and the support sup_i, $\{w_0, ..., w_N\}$ are the combination weights and Φ the radial basis function.

The modeling of RBF fuzzy proximity is both discriminating and intrinsic. Indeed the layer of radial basic functions corresponds to an intrinsic description of the training data, then the output combination layer seeks to discriminate different classes. In our system, a Cauchy function is used as a radial basis function:

$$\Phi(d) = \frac{1}{1+d} \tag{7}$$

we define two new operators:

$$Relw(C) = \frac{degree(C)}{total \;\; number \;\; of \;\; concepts} \tag{8}$$

$Relw(C)$ is the relational weight of the concept C (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.

$$SemDensity(C_1, C_2) = \frac{MinCost(C_1, C_2)}{minimal \;\; cost \;\; of \;\; the \;\; Spanning \;\; Tree} \tag{9}$$

$SemDensity(C_1, C_2)$ 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 [2]. This distance is calculated from the semantic graph, this latter is built from the document based on the minimal cost of the spanning tree (i.e. the minimal cost tree by following all minimal paths from 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:

$$SemDist(C_1, C_2) = Relw(C_1) \cdot Relw(C_2) \cdot SemDensity(C_1, C_2) \tag{10}$$

The proximity measure is a Cauchy function:

$$Proximity(C_1, C_2) = \frac{1}{1 + SemDist(C_1, C_2)} \tag{11}$$

The contribution of these defined operators is that they give more importance to concepts which have dense semantic vicinity where they have good connectivity within the graph. This has also been verified during the validation of the prototype. In the classification phase, we will see in the following how the weights of indexing descriptors are generated by the new measure of RBF fuzzy proximity based on the semantic distance as a parameter.

4.2 Our RBF Fuzzy Proximity Model

We start from the idea that where terms semantically close to terms which used in the query, appear directly close in the base document. The Measure of

Mercier and Beigbeder [10] is very important, yet it does not take into account the semantic proximity between terms. Indeed, this model is limited by the direct relationship of terms concurrence that does not capture the semantic proximity between words. The equation presented in the model of Mercier and Beigbeder becomes:

$$\mu_t^d(j) = max_{i \in d^{-1}(zone(t))}(max(\frac{k - |x - i|}{k} \cdot \frac{\Phi(Proximity(t, t_i)}{1 + |freq(t) - freq(t_i)|}, 0)) \quad (12)$$

We indicate by $Proximity(t, t_i)$, the semantic proximity between t and their neighbours t_j at position j, as defined in Eq. 11.

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

The $\frac{\Phi(Proximity(t,t_i)}{1+|freq(t)-freq(t_i)|}$ value does not exceed in any case the value 1, the local relevance of a term t at a position i taken by the terms that are semantically close reached the maximum value of relevance when the position i is occupied by the term t itself. The difference in frequencies is added to circumvent the problem of co-occurrence, thus we multiplied the local relevance by $\frac{\Phi(Proximity(t,t_i)}{1+|freq(t)-freq(t_i)|}$ since the positions i of the terms belonging to the influence $zone(t)$ of the term t, and which are semantically close, are taken into account but their influences should depend on the degree of proximity, which they share with the term t Hence the justification of this multiplication.

5 Results

To validate this new approach, we tested it on a varied corpus of 5000 documents electronic press extracted from sites (AL JAZEERA[3], AL ARABIYA[4]). Table 1 show different results for each measure. These results are expressed through the recall and precision criteria. In particular, they show the relevance of using radial basic functions which greatly improves the measures performance with which they are combined.

From Table 1, 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.

[3] http://www.aljazeera.net.
[4] http://www.alarabiya.net.

Table 1. Standard and RBF fuzzy proximity results

Measure	Corpus	Precision	Recall
RBF fuzzy proximity	Economy	0.86	0.66
	Politic	0.78	0.68
	Sport	0.94	0.77
Standard fuzzy proximity	Economy	0.63	0.61
	Politic	0.68	0.60
	Sport	0.74	0.70

6 Discussion and Conclusion

The semantic proximity between words must be highlighted when we deal with complex documents such as texts in Arabic. For this purpose, it is essential to broaden our reflection to the adapted representation models to the nature of our resources. For this, we studied the research model based on the proximity of terms based on the classic fuzzy model. This approach is based on the assumption that most terms occurrences of a query are close in a database document, more this document is relevant to this query, This can partially solve the problems caused by the complex or compound words which may also be an interesting track, since long concepts are often less ambiguous. However, this model does not consider the notion of terms semantics, since it is limited by the presence of co-occurrence relations of the terms, also does not take into account the semantic links which may exist between the query terms and those of the document. The integration of a semantic measure between terms in this model is needed. For this reason, we have introduced our radial basis contribution to formalize the adaptation of the model based on the semantic fuzzy proximity concept to the needs of the semantic pairing. The advantage of this model is that it does not need a preliminary glossary to identify terms in order to assign them a weight, since the identification of terms is made simply from a query-document matching according the shape of document key words (lemma or radicals). The integration of the semantic vicinity concept and radial basis functions improves significantly the performance of the classical measures, especially for the Arabic language, which remains our goal.

References

1. Aljlayl, M., Frieder, O.: On Arabic search: improving the retrieval effectiveness via a light stemming approach. In: 11th International Conference on Information and Knowledge Management (CIKM), pp. 340–347 (2002)
2. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction à l'algorithmique, deuxième édition edn. MIT Press, McGraw-Hill, Cambridge, New York City (2001)

3. Daimi, K.: Identifying syntactic ambiguities in single-parse Arabic sentence. Comput. Humanit. **35**, 333–349 (2001)
4. El Bazzi, M.S., Mammass, D., Ennaji, A., Zaki, T.: Features based approach for indexation and representation of unstructured Arabic documents. Adv. Sci. Technol. Eng. Syst. J. **2**(3), 900–905 (2017)
5. El BazzI, M.S., Mammass, D., Zaki, T., Ennaji, A.: A graph-based ranking model for automatic keyphrases extraction from Arabic documents. Advances in Data Mining. Applications and Theoretical Aspects. LNCS (LNAI), vol. 10357, pp. 313–322. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-62701-4_25
6. El Bazzi, M.S., Zaki, T., Mammass, D., Ennaji, A.: Indexation automatique des textes arabes: état de l'art. E-Ti: Electron. J. Inf. Technol. **41**, 48–64 (2016)
7. El Bazzi, M.S., Zaki, T., Mammass, D., Ennaji, A.: Stemming versus multi-words indexing for Arabic documents classification. In: 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), pp. 1–5 (2016)
8. Khoja, S., Garside, S.: Stemming Arabic Text (1999). http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps
9. Larkey, M., Ballesteros, S.L., Connell, L.: Improving stemming for arabic information retrieval: light stemming and cooccurrence analysis. In: Proceedings of the 25st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2002, pp. 275–282. ACM, New York (2002)
10. Mercier, A., Beigbeder, M.: Application de la logique floue à un modèle de recherche d'information basé sur la proximité. In: Dans les Actes LFA 2004, pp. 231–237 (2005)
11. Navigli, R.: Word sense disambiguation: a survey. ACM Comput. Surv. **41**(2), 10:1–10:69 (2009)
12. Quillian, R.M.: Semantic memory. In: Semantic Information Processing, pp. 216–270. MIT Press, Cambridge (1968)
13. Rada, R., Mili, H., Bicknell, E.: Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. **19**, 17–30 (1989)
14. Tchechmedjiev, A.: État de l'art sur les mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances. In: Actes de la conférence conjointe JEP-TALN-RECITAL. RECITAL 2012, vol. 3, pp. 295–308 (2012)
15. Wagner, C.: Breaking the knowledge acquisition bottleneck through conversational knowledge management. Inf. Resour. Manag. J. **19**(1), 70–83 (2008)