Classification of Arabic Documents by a Model of Fuzzy Proximity with a Radial Basis Function

Taher ZAKI[1], Driss MAMMASS[1], Abdellatif ENNAJI[2], F. NOUBOUD[3]
[1]*IRF-SIC, Faculty of Science, Ibn Zohr University,*
Agadir, Morocco
tah_zaki@yahoo.fr , mammass@univ-ibnzohr.ac.ma
[2]*LITIS - University of Rouen,*
Rouen, France
abdel.ennaji@univ-rouen.fr ,
[3]*Université du Québec à Trois – Rivières ,Québec, Canada*
nouboud@uqtr.ca

Abstract

In this paper we propose a model of classification based on the principle of the fuzzy proximity of the terms within the documents.

Given the heterogeneous nature of the Arabic documents in our possession, we have studied for this purpose the research model based on the semantic proximity of terms and inspired from the classic Boolean model.

Our approach is based on the assumption that more the occurrences of terms in query are close with good connectivity in the extracted semantic graph from the set of document , more this document is relevant to this query.

We propose a measure that provides a contextual and semantic search. 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 graph and therefore the discrimination of documents relevant to the query.

Keywords: *document classification, semantic graph, semantic vicinity, dictionary, kernel function, similarity.*

1. Introduction

The information retrieval is an area of investigation for the semantic similarity. Indeed, the problems of polysemy and synonymy of our languages generates ambiguities in the research and the difficulties of consensus in the choice of terms for indexing and searches are still present. However, it is essential to pass to a semantic level, to avoid syntax problems and comparing term to term.

To achieve this end, we opted for research systems based on fuzzy proximity of terms. The principle of these systems is to assess the density of the query terms in the texts to assess their relevance.

Mercier and al. [1] had shown that more the terms of query are close in a document more this document is relevant. This assumption is very interesting; however it does not take into account the semantics of terms (in the case where terms semantically close to terms of query appear directly close to an element of the base).

We will provide a semantic extension to this model by integrating semantic graphs to highlight the semantic connections between words and we will use a dictionary auxiliary to increase the connectivity of the graph thus constructed. This new model provides a contextual search and semantics by taking into account the explicit information about the text, namely the structure, as well as the implicit information, ie the semantics.

Our research problem has the text as pivot element. We focus on the inclusion of explicit information in the text, namely the structure, as well as implicit information, namely the semantics.

2. Measure of proximity

We focus on the inclusion of explicit information about the text, namely the structure, as well as implicit information, namely the semantics.

2.1. Semantic proximity

The measure between terms should be put in context when dealing with documents. The vector model introduced by Salton [9] excludes all position notion and distance notion between words. Moreover, the model of Salton is better suited to the codification of longer texts than for the codification of shorter texts [2].

Given the heterogeneous nature of the texts in our possession, it is essential to broaden our thinking to the models of representation adapted to the nature of our resource (heterogeneous). We studied for this purpose the model of research based on the proximity of the terms, inspired from the classic Boolean model.

In the following section, we will introduce the fuzzy model of Mercier. We think that this model has an ergonomic design and easily adaptable to our measure since the two approaches are based on fuzzy research.

2.2. The model of fuzzy proximity

Beigbeder and Mercier [11] estimate that more the query terms appear close within a document more this last is relevant compared to the target query. The fuzzy model of proximity of a term A has compared to a term B is formalized by :

$$\mu_{NEAR\ (A,B)}(d) = Max_{\substack{i \in d^{-1}(A) \\ j \in d^{-1}(B)}} \left(Max \left(\frac{k - |j - i|}{k}, 0 \right) \right) \quad (1)$$

Where $d^{-1}(t)$ indicates the whole of the positions taken by the term t, and k one fixed constant selected representing the size of the slipping window of the cooccurences of the terms. The parameter K is a constant which characterizes the degree of influence of an occurrence. (a low value (5) evaluates the proximity within the framework of an expression, a value of K equal to 100 translated the proximity in a context of a paragraph and so on).

After comparing two terms and before comparing a request and a document, it is necessary to compare a term and a document. For this, the function calculates the degree of relevance for each term t of r in the whole of the possible positions x of t in d.

$$\mu_t^d(x) = Max_{i \in d^{-1}(t)}\left(Max\left(\frac{k - |x - i|}{k}, 0\right)\right) \quad (2)$$

2.3. Relevance of a query against a document

The model of direct co-occurence rests on the assumption that the queries obey the classical Boolean model, it is to be said that a query is a series of conjunctions and disjunctions of terms. Consequently the local relevance of the query R follows the same logical diagram between the respective relevances of the terms of the request. The logical operators are the traditional operators (AND, OR). For example, for R = A AND B and R = A OR B, the local relevance of R corresponds to:

$$\begin{aligned} \mu^d_{A\ AND\ B}(x) &= Min\left(\mu^d_A(x), \mu^d_B(x)\right) \\ \mu^d_{A\ OR\ B}(x) &= Max\left(\mu^d_A(x), \mu^d_B(x)\right) \end{aligned} \quad (3)$$

The relevance at a document level is generalized in a natural way by an aggregation of the results obtained in a whole of the possible positions.

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

Thus, the similarity is obtained by normalizing the whole of the scores by the length of the document.

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

We brought an extension to the model of Mercier-Beigbeder by combining it with a radial basis function in order to take into consideration the semantic vicinity of the terms which appears absent in this model.

3. The semantic resources

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 overriding choice 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…).

Taxonomy: We introduce taxonomy as the organization of concepts linked by hierarchical relations [24].

Semantic networks: The semantic networks [13], 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. 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 [17].

5. Radial basis function

5.1. Definition

Radial basis functions are means to approximate multivariable (also called *multivariate*) functions by linear combinations of terms based on a single univariate function (the radial basis function). This is radialised so that it can be used in more than one dimension. They are usually applied to approximate functions or data ([6] and [15]) which are only known at a finite number of points (or too difficult to evaluate otherwise), so that then evaluations of the approximating function can take place often and efficiently [25].

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, so that $\phi(x) = \phi(\|x\|)$; or alternatively on the distance from some other point c, called a *center*, so that $\phi(x,c) = \phi(\|x-c\|)$. Any function φ that satisfies the property $\phi(x) = \phi(\|x\|)$ is a radial function. The norm is usually Euclidean distance, although other distance functions are also possible. For example, by using Lukaszyk-Karmowski metric it is for some radial functions possible [18] to avoid problems with ill conditioning of the matrix solved to determine coefficients w_i (see below), since the $\|x\|$ is always greater than zero.

Sums of radial basis functions are typically used to approximate given functions. This approximation process can also be interpreted as a simple kind of neural network.

5.2. Approximation

Radial basis functions are typically used to build up function approximations of the form :

$$y(x) = \sum_{i=1}^{N} w_i \, \phi(\| x - c_i \|) \tag{6}$$

Where the approximating function $y(\boldsymbol{x})$ is represented as a sum of N radial basis functions, each associated with a different center $\boldsymbol{c}_i$, and weighted by an appropriate coefficient w_i. The weights w_i can be estimated using the matrix methods of linear least squares, because the approximating function is *linear* in the weights ([10] and [3]).

Approximation schemes of this kind have been particularly used in time series prediction and control of nonlinear systems exhibiting sufficiently simple chaotic behaviour, 3D reconstruction in computer graphics (for example, hierarchical RBF).

Indeed, one of the greatest advantages of this method lies in its applicability in almost any dimension (whence its versatility) because there are generally little restrictions on the way the data are prescribed (there are no restrictions on the data except that they need to be at distinct points). This should be contrasted to, e.g., multivariable *polynomial interpolation* (but see [5] for an especially flexible approach) or *splines*. A further advantage is their high accuracy or fast convergence to the approximated target function in many cases when data become dense.

For applications it is indeed desirable that there are few conditions on the geometry or the directions in which the data points have to be placed in space. No triangulations of the data points or the like are required for radial basis function algorithms, whereas for instance finite element ([14] and [16]) or multivariate spline methods ([4] and [12]) normally need triangulations. In fact, the advance structuring of the data that some other approximation schemes depend on can be prohibitively expensive to compute in applications, especially in more than two dimensions. Therefore our approximations here are considered as meshfree approximations, also for instance to be used to facilitate the numerical solution of partial differential equations [8].

6. Construction of the auxiliary dictionary

The Auxiliary Semantic Dictionary It is a hierarchical dictionary with standard vocabulary based on generic terms and on specific terms to a domain**.** For example (finances and economy) :

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 **…**
….

….

Figure 1. Dictionary of finances and economy

This dictionary will be enriched progressively during the search and classification to give more flexibility to our model.

7. Construction of the semantic graph

Let us take for example a document of finances 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**.

Figure 2. Plain text

Business
Bill
Economic

Fund
Businesses
Tax
Companies
Taxes
Entrepreneurs
Business
Jobs
Bill

Unemployment
Economy

Bill
Businesses
loans
financial crisis

incentives
business credit
jobs

Figure 3. Text after preprocessing, filtering

The extraction of keywords is done in order of appearance in the document, the order is important here.

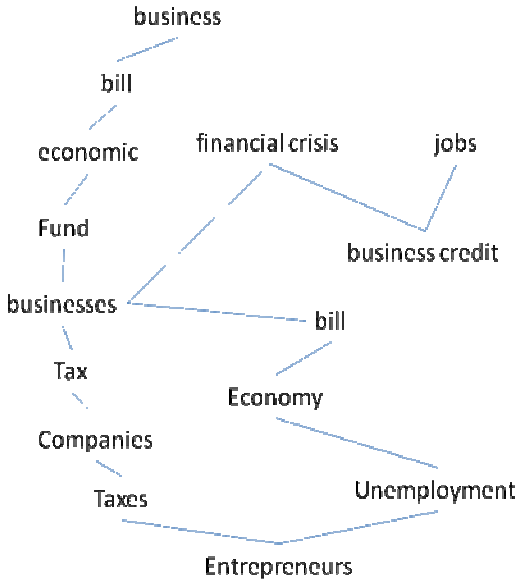

Figure 4. Semantic graph extracted from document

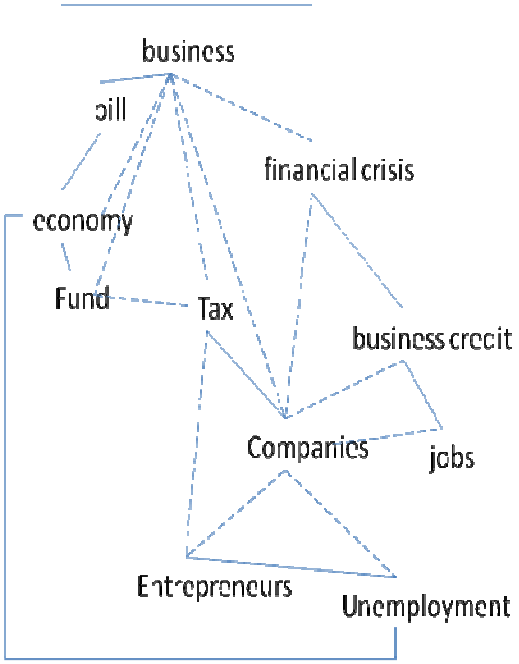

Figure 5. Strengthening of the graph by semantic connections from the auxiliary dictionary

The construction of the semantic graph holds in account the order of extraction and the distribution of terms within the document. Each term is associated with a radial basis function which fixes the proximity at certain vicinity (zone of influence semantics of the term).

The correspondence, query-document, is done by a projection of the terms of the request on the semantic graph. If these terms are in zone of strong semantic influence then this document is relevant at this request.

In the following we will define our radial basis function and we'll see the utilty of the graph to calculate the semantic proximity between the query and the document.

8. Choice of the radial basis function

The radial basis function is of Cauchy type :

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

We have defined two new operators:

- **The relational weight :**

$$\text{WeightRel}(t) = \frac{\text{degree}(t)}{\text{total number of concepts}} \tag{8}$$

- **Semantic density :**

$$SemDensity\ (c_1, c_2) = \frac{Dist\ (c_1, c_2)}{\text{recovering tree with minimal cost}} \quad (9)$$

Thus the semantic distance between two concepts

$$\text{DistSem}(c_1, c_2) = \text{WeightRel}(c_1) * \text{WeightRel}(c_2) * SemDensity\ (c_1, c_2) \quad (10)$$

The proximity measure is a Cauchy function:

$$\text{Pr}\ oximity\ (c_1, c_2) = \frac{1}{1 + DistSem\ (c_1, c_2)} \quad (11)$$

$\deg ree(t)$: 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 [7] and [19], 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.

9. Our new model: The fuzzy model of the semantic proximity with radial basis

9.1 The local semantic relevance

We bring an extension to the existing model by combining it with the measure of conceptual similarity based on radial basis function mentioned before :

$$\mu_t^d(x) = Max_{i \in d^{-1}(zone(t))}\left(Max\ \left(\frac{k - |j - i|}{k} \left(\phi(d) \left| freq(t) - freq(t_i) \right| \right), 0 \right) \right) \quad (12)$$

We indicate with $zone(t)$ the set of terms semantically close to t. A threshold of similarity is necessary to characterize the whole of its elements. We fix a threshold of similarity for the value of proximity which corresponds to the degree of similarity between T and the concept power station (the term is accepted if it is in the zone of influence of core term defined by the radial basis function ϕ).

We added the difference in frequency to circumvent the problem of co-occurence.

9.1 The fuzzy model of the semantic proximity with radial

The measure of ([1] and [11]) is very interesting; however it does not take account of the semantics of the terms (if terms semantically close to the terms used in the query appear directly close within a document of the base). Indeed, this model is limited by the relation of direct co-occurence of the terms which does not capture the semantic proximity between words. The equation presented in the model of Mercier and Beigbeder becomes:

$$\mu_{NEAR(A,B)}(d) = Max_{\substack{i \in d^{-1}(A) \\ j \in d^{-1}(B)}} (Max \ (\frac{k - |j-i|}{k} (\phi(d) \, | freq(A) - freq(B) |), 0) \quad (13)$$

$\phi(d)$ is the radial basis function associated with the zone of semantic influence, it represents the whole of the terms close to A according to the used conceptual measurement of pairing . Our new model brings the knowledge of semantics to the existing model. The results of the degrees of proximity which we obtained at the stage of experimentation on real data and simulated data are are greatly improved by using higher semantics.

10. Results

To validate our prototype we tested our model on two databases of documents written in English and Arab languages.

The database of English documents is a corpus of 300 documents of associated press (the Associated Press (AP)), it is a very rich and varied database of 2246 documents [23]. And for the Arab documents we worked on a corpus of 300 documents of Arab electronic presses ([20], [21] and [22]).

Table 1. Results of experiments

Corpus	Method	Recall	Precision	accuracy (%)
English	TFIDF	0.80	0.83	80.3
	Fuzzy model with RBF	0.92	0.92	92.77
Arabic	TFIDF	0.81	0.81	81.0
	Fuzzy model with RBF	0.98	0.98	98.79

11. Conclusion

The concept of vicinity semantic is a very interesting way for the information retrieval, to exploit it, we took into consideration all types of links between concepts, Based on the proximity measure between concepts.

We put forward a new measure introducing a radial basis function to hold in account the semantic vicinity and the results obtained are encouraging and our system ensured a good performance for classification associated with each unit semantic. But it remains of other way to add to increase its performances.

Acknowledgements

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

References

[1] A. Mercier and M. Beigbeder, «Application de la logique floue à un modèle de recherche d'information basé sur la proximité». Dans les Actes LFA 2004, 231–237, 2005.

[2] A. Singhal, «Pivoted length normalization». In Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR'96), 21–29, 1996.

[3] B. J. C. Baxter 1992 , A dissertation presented in fulfilment of the requirements for the degree of Doctor of Philosophy, Cambridge University ,August 1992.

[4] C. de Boor, Multivariate piecewise polynomials Acta Numerica 2: 65-109, 1993.

[5] C. de Boor and A. Ron, On multivariate polynomial interpolation Constructive Approximation 6: 287-302, 1990.

[6] E. W. Cheney, Introduction to Approximation Theory McGraw-Hill, New York, 1966.

[7] Edsger W. Dijkstra : A short introduction to the art of programming , contenant l'article original décrivant l'algorithme de Dijkstra (pages 67 à 73).

[8] Greg Fasshauer, Meshfree Approximation Methods with Matlab World Scientific, Singapore, 2007.

[9] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24 :513–523, 1988. 19.

[10] H. H. Zhang, G. M. , Genton, and P. Liu, Institute of Statistics Mimeo Series 2570, NCSU, 2004.

[11] M. Beigbeder, and A. Mercier, « Fuzzy set theory applied to a proximity model for information retrieval ». Nantes, France. LFA, 231-237, 2004.

[12] M. J. Lai and L.L. Schumaker, Spline functions on triangulations Cambridge University Press, Cambridge, 2007.

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

[14] P. Ciarlet, The finite element method for elliptic problems North-Holland, Amsterdam, 1978.

[15] P. J. Davis, Interpolation and Approximation Dover, New York, 1975.

[16] S. Brenner and L. Scott , The mathematical theory of finite elements Springer, New York, 1994.

[17] 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.

[18] S. Lukaszyk, A new concept of probability metric and its applications in approximation of scattered data sets. Computational Mechanics, 33, 299-3004, 2004.

[19] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, Introduction à l'algorithmique, (version (en) (ISBN 0-262-03293-7) deuxième édition, 2001, MIT Press and McGraw-Hill, section 24.3, Dijkstra's algorithm, pages 595–601).

[20] Al Jazeera, http://www.aljazeera.net/

[21] Al charq Al awsat , http://www.aawsat.com/

[22] Al ahdat Al maghrebiya , http://www.almaghribia.ma/.

[23] Associated Press, http://www.cs.princeton.edu/~blei/lda-c/ap.tgz.

[24] Wikipedia, http://en.wikipedia.org/.

[25] Scholarpedia, http://www.scholarpedia.org/article/Radial_basis_function.

Authors

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

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

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

Fathallah NOUBOUD received the B.Sc. Degree in mathematics and physics from University of Marrakech, Morocco, in 1983, and the Master's degree in mathematics and the Ph.D. degree in computer science, in 1984 and 1988, respectively, both from University of Caen, France. His doctorate project consisted in the design of a handwritten signature verification system.
In 1989, he joined Laboratoire Scribens at École Polytechnique of Montréal as a Researcher. In 1992, he became Professor at University of Quebec at Trois-Rivières. His research interests include mathematical imaging, document analysis, and environmental applications.