Graph-Based Text Modeling: Considering Mathematical Semantic Linking to Improve the Indexation of Arabic Documents

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

[1] IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco
elbazzi.mohamedsalim@edu.uiz.ac.ma, {mammass,t.zaki}@uiz.ac.ma
[2] LITIS Laboratory, University of Rouen, Rouen, France
abdel.ennaji@univ-rouen.fr

Abstract. Indexing unstructured documents aims to build a list of words, or concepts, which will simplify the exploration of their exploration later on. The most used model for text modeling is the Vector Space Model. In spite of the simplicity of this model in its implementation and its wide use in different researches in the field of text mining and information retrieval, it has an important limit, which is ignoring the semantic relation between the different textual units, by considering them as independent. However, there is a more suitable technique in Data Mining to highlight the semantic linkage between text units, which is the graph-based representation. A graph can easily be adapted to the textual data by representing words as a vertex and the relation between them as edges. In this work, we have introduced the graph based modeling of textual document. Thus, we conducted a study about the impact of the choice of the semantic relation between the text units on the indexation of documents. We have validated our results through classification results.

Keywords: Text mining · Semantic graphs · Semantic measures · Arabic documents · Indexation · Classification

1 Introduction

The amount of documents available on the Internet and the digitization of textual documents are constantly increasing. This revolutionary change presents major challenges for researchers to explore the hidden information by introducing different approaches.

Documents indexing is the main step in a conventional document classification or information retrieval framework. This study aims to highlight the influence of semantic metrics on the efficiency of a classification system. Empirical results are applied to an Arabic dataset. Precision, recall and F-measure are the metrics adopted to compare the efficiency of the proposed indexing system.

Document indexing involves extracting keywords that best represent a document. Despite the crucial role of this phase in the subsequent processes of mining and analyzing texts, few are the works identified at this level [1]. This paper presents a reading of the different methods of extraction of the descriptors, as well as their applications and

A. Mansouri et al. (Eds.): ICISP 2018, LNCS 10884, pp. 138–144, 2018.
https://doi.org/10.1007/978-3-319-94211-7_16

compatibility with the Arabic language. Nevertheless, we propose a new indexing system based on graph modeling. The aim of this paper is to assess the impact of semantic relation between terms and its impact on the classification step, as matter of fact.

The rest of this article is organized as following. The second part introduces related works. The third part is dedicated to the presentation of the semantic approaches and our proposed indexing method. We will present the experimental results in the fourth section and conclude by discussing our contribution.

2 Related Works

Indexing a document is to elect its most representative descriptors in order to generate the list of indexing terms. It is a way of retrieving all the significant terms characterizing a document. Document indexing is a critical step in the text mining process as it determines how the knowledge contained in the documents is represented.

In [2], a study of different variants of the Vector Space Model (VSM) using the K-Nearest Neighbor (KNN) algorithm is introduced. These variants are the cosine measure, the Dice measure and the Jaccard measure using different methods of terms weighting. The results obtained on an Arabic corpus showed that the performance obtained by Dice-TFIDF and Jaccard-TFIDF outperforms those obtained by Cosine-based TFIDF.

Mohamed and Watada [3] used latent semantic analysis (LSA) to evaluate each term in a document, then they used an evidential reasoning (ER) to assign new document to a category according to the training data. Experiments were performed by combining ER with LSA and ER with TFIDF. ER-LSA gives better results than ER-TFIDF.

In [4], the vector space model was extended by combining TFIDF with the Okapi formula for extracting relevant concepts that better represent a document. The authors propose a new measure that takes into consideration the notion of semantic proximity using a measure of similarity between words, and combining TFIDF-Okapi with a radial basis function. Experimental results confirm the performance of their contribution.

Al-Salemi [5] used characteristics selection techniques such as mutual information (MI), statistics $\chi 2$ (CHI), information gain (GI), ESG coefficient And Odds Ratio (OR) to reduce the size of feature space by eliminating items that are considered irrelevant for a category being studied.

Jamoussi [6] proposes a method of extracting keywords based on the semantic relationship between words. The author introduces two methods based on semantic distances, the Kullback–Leibler divergence and the average mutual information to calculate the quantity of information between two words or two classes of words.

In [7], the authors propose a hybrid system for contextual and semantic indexing of Arabic documents, providing an improvement to conventional models based on n-grams and the Okapi model. Their method consists on calculating the similarity between words using hybridization of the statistical measures N-Grams, okapi and a kernel function. In order to have a strong descriptor, the authors used a semantic graph to model the semantic connections between terms using an auxiliary dictionary to increase the connectivity of the graph. The weights of the words are then calculated using a radial basis function. This method has improved the performance of the indexation system.

Mesleh and Kanaan [8] applied an ant colony optimization (ACO) as a feature space reduction mechanism with $\chi 2$ as a score method and then they classified the Arabic documents using the SVM classifier.

Other models are used in the literature [9, 10]. For example, LSI (Latent Semantic Indexing) is a method that attempts to take into account the semantics of terms for the representation of documents. In this model, the documents are represented in a reduced space of indexing term. Hofmann [11] proposes a probabilistic model of Latent Semantic Indexing (PLSI). It considers the assumption that documents are associated with a certain number of meanings, and that the terms correspond to the expression of these senses.

3 Semantic Document Indexing

A set of statistical classification models and automatic learning techniques have been applied to the classification of texts, including linear regression models such as LLSF [12], K nearest neighbors [13], the decision tree [14], SVM (Support Vector Machines) maximum entropy [15], the distance-based classifier [16, 17], the WordNet knowledge-based classifiers [18]. Hence, the advances in classification methods seems to be satisfying, when the evolution of indexation methods still ignored especially for Arabic documents.

3.1 Linguistic Approaches

In [19], the authors propose a contextual exploration method to remove the ambiguity of a sequence of words. This method is based essentially on the morphological and syntactical analysis, and the exploitation of the grammatical rules for the recognition of the words adjacent to the sequence in question. Using rule-based methods provide good results in particular cases, but it is complicated to use it for a huge set of unstructured data.

However, methods using external semantic resources (dictionary, ontology or other) offer a better semantic coverage of the document. The main problem with this technique is that the recognition of the semantic units is limited to the domain described by the used resource.

3.2 Mathematical Approaches

Statistical Methods

In the field of Natural Language Processing (NLP), the data is not numerical. In order to process them in an automatic way, it is essential to find them an appropriate numerical representation that preserves their semantic and syntactic properties. Then, we need to find effective measures to compare and compute semantic distances between words. This issue is widely discussed in the literature. In fact, a semantic measure must be able to express quantitatively a similarity between two terms from a semantic point of view.

As matter of fact, words co-occurrence statistics describes how words occur together and captures the relationships between them. Words co-occurrence statistics is computed by counting how two (or more) words occur together in a given document.

However, it exists more sophisticated methods to represent the semantic or terms. For example, the mutual information between two words or between two classes of words provided by these two entities, by considering them as two random variables X and Y. The formula of the mutual information between two words $\mathbf{w_i}$ and $\mathbf{w_j}$ is given by:

$$MI_{wi,wj} = log(\mathrm{P}(wi, wj)/P(wi)P(wj)) \tag{1}$$

Where P(wi, wj) is the probability of finding the two words $\mathbf{w_i}$ and $\mathbf{w_j}$ in the same context, P(wi) and P(wj) respectively represent the probability of meeting independently $\mathbf{w_i}$ or $\mathbf{w_j}$. Mutual information between words is thus a representative measure of the co-occurrence of two words in the same sentence or in the same paragraph. It shows that this co-occurrence is not due to chance. If it is important, it is because these two words are often found together and therefore the existence of one depends on the other. By calculating the mutual information between the different words of the text to be analyzed, we can extract the list of its triggers. A trigger is a pair of semantically correlated words, that is to say that the appearance of one of the two words reinforces the probability of occurrence of the other.

Graph Based Method

In this work, we have adapted TextRank Algorithm in order to extract keywords from Arabic documents. TextRank is proposed by Mihalcea and Tarau [20] that represent text as graph. In general, vertices represent words and edges represent the relation between words (semantic, structure…).

TextRank is a basic adaptation of PageRank [21] for automatic keywords extraction. For each node of the graph, a score is calculated by an iterative process to simulate the concept of recommendation of a term by its adjacent vertices. The score at each vertex grants a ranking degree to the word considering its importance in the processed document. Then, the sorted list of words can be used to extract keywords.

The score of the vertex **v**, denoted $\mathbf{S(v_i)}$, is initialized by a default value, and calculated iteratively until convergence using the following formula:

$$S(\mathrm{v_i}) = (1-d) + d \times \sum_{\mathrm{v_j} \in \mathrm{Adj(v_i)}} \frac{\mathrm{w_{ji}}}{\sum_{\mathrm{v_k} \in Adj(\mathrm{v_j})} \mathrm{w_{jk}}} S(\mathrm{v_j}) \tag{2}$$

Where Adj (v_j) represents the neighbors of v, and d is a damping factor set at 0.85 by [21]. A weight w_{ji} is associated to each edge connecting two vertices v_1 and v_2, and represents the frequency of co-occurrence of two words within a window of 2 to10 words.

4 Experiments and Results

We have performed an experimental study of the semantic indexing based on graph representation using semantic metrics. Our approach aims to assess the indexation accuracy using this model. We have used the graph representation as defined in [20] where each document is represented by a graph.

We have implemented the **mutual information** method to weight edges and co-occurrence method as well. According to [20], the best score is achieved when non-oriented edges are used to connect words that co-occur within a window of 2 words. Furthermore, in TextRank the initial score of a edge is initialized by 1. Consequently, our implementation of TextRank follows these indications. Moreover, we conduct a test consisting of initializing edges by information mutual and co-occurrence.

To validate the proposed system, we have tested the methods on a corpus of 1084 documents extracted from Arabic websites. This corpus is classified into three categories: Economics, Politics and Sport.

For experimental aim, we adopted KNN to classify documents considering its significant performance. We have implemented the whole classification process as described in [22]. To evaluate the classification performance, three metrics are used: precision, recall and F-measure.

In this section, we evaluate the performance of the proposed system for structural indexing over standard statistical system (**TFIDF**), on one hand. On the other hand, we conducted serial of tests using semantic methods to weight edges. Table 1 shows different results obtained after classification.

Table 1. Classification results comparing TFIDF to TextRank

	Precision	Recall	F-measure
TFIDF	0.322	0.369	0.344
TextRank (co-occurrence = 2)	**0.460**	**0.406**	**0.431**
TextRank (co-occurrence = 5)	0.248	0.369	0.321
TextRank (co-occurrence = 10)	0.365	0.372	0.368

The obtained results show that graph modeling improves significantly the process of indexation. In fact, using graph-based representation of a document can be used to extract keywords. A graph reflects not only the frequencies of terms and their adjacencies, but also the contextual information of the documents.

We have noted that the effectiveness of the classification system decreases when using edge weighting for co-occurrence method. We interpret this decrease by the basic graph implementation. In TextRank [20] edges are weighted by one (there is no special difference between terms) (Table 2).

Table 2. Classification results using weighted edges.

	Precision	Recall	F-measure
TextRank (co-occurrence = 2)	0.460	0.406	0.431
TextRank (Mutual Information)	0.534	0.503	0.526

However, when we used semantic metrics we affected to edges new values, the thing that made the deference. These new values evaluated the semantic proximity between words of the document. In other words, the proposed approach has highlighted the semantic relationships that exist between different words of the document. The application of this new parameterization has strengthened the semantic indexing, which has led to better classification results.

5 Conclusion

In this paper, we have conducted an evaluation of graph-based modeling of textual data for keywords extraction. We have tested the impact of weighting edges with semantical methods. Thus, the obtained results showing that using weighted edges do not bring any successful contribution.

The graph model still most suitable representation of text document. It represents better the relationship between words by preserving the structural representation of the context. This will definitely lead to a better result.

Nonetheless, graph based modeling is highly portable to other languages and does not required any detailed linguistic knowledge. Consequently, the proposed approach can be used to perform the indexation of unstructured documents in different languages.

References

1. Zaki, T.: Indexation par le contenu et archivage de fonds documentaires arabes. Thesis. Ibn Zohr University, Agadir, Morocco (2013)
2. Thabtah F., Hadi, W., Al-shammare, G.: VSMs with K-nearestneighbour to categorise Arabic text data. In: Proceedings of The World Congress on Engineering and Computer Science, WCECS 2008, pp. 778–781 (2008)
3. Mohamed, R., Watada, J.: An evidential reasoning basedlsa approach to document classification for knowledge acquisition. In: Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2010, pp. 1092–1096. Institute of Electrical and Electronics Engineers (IEEE) (2010)
4. Zaki, T., Mammass, D., Ennaji, A.: A semantic proximity based system of arabic text indexation. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D., Meunier, J. (eds.) ICISP 2010. LNCS, vol. 6134, pp. 419–427. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13681-8_49
5. Al-Shalabi, R., Obeidat, R.: Improving KNN arabic text classification with n-grams based document indexing. In: Proceedings of the Sixth International Conference on Informatics and Systems, INFOS q 2008, pp. 108–112 (2008)
6. Jamoussi, S.: Une nouvelle représentation vectorielle pour la classification sémantique. TAL 2009, vol. 50 (2009)
7. Zaki, T., Mammass, D., Ennaji, A., Nicolas, S.: A kernel hybridization NGram-Okapi for indexing and classification of Arabic documents. J. Inf. Comput. Sci. **9**(2), 141–153 (2014). ISSN 1746-7659, England, UK
8. Mesleh, A.M., Kanaan, G.: Support vector machine text classification system: using ant colony optimization based feature subset selection. In: Proceeding of the International Conference on Computer Engineering & Systems, ICCES 2008, pp. 143–148 (2008)

9. Hasan, K.S., Ng, V.: Conundrums in unsupervised keyphrases extraction: making sense of the state of the art. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Poster Volume (2010)
10. Mesleh, A.: Support vector machines based Arabic language text classification system : feature selection comparative study. In: Proceedings of the 12th WSEAS International Conference on Applied Mathematics, MATHq 2007, pp. 11–16. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA (2007)
11. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR 1999 Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57 (1999)
12. Yang, Y., Chute, G.C.: An example-based mapping method for text categorization and retrieval. ACM Trans. Inf. Syst. **12**(3), 252–277 (1994)
13. Kanaan, G., Al-Shalabi, R., AL-Akhras, A.: KNN Arabic text categorization using IG feature selection. In: Proceedings of The 4th International Multiconference on Computer Science and Information Technology, CSIT 2006, vol. 4 (2006)
14. Li, H.Y., Jain, K.A.: Classification of text documents. Comput. J. **41**(8), 537–546 (1998)
15. El-Halees, A.M.: Arabic text classification using maximum entropy. Islam. Univ. J. (Ser. Nat. Stud. Eng.) **15**(1), 157–167 (2007)
16. Duwairi, R.M.: A distance-based classifier for Arabic text categorization. In: Proceedings of The 2005 International Conference on Data Mining, DMIN 2005, pp. 187–192. CSREA Press (2005)
17. Khreisat, L.: Arabic text classification using N-gram frequency statistics a comparative study. In: Proceedings of The 2006 International Conference on Data Mining, DMIN 2006, pp. 78–82. CSREA Press (2006)
18. Benkhalifa, M.A., Mouradi, A., Bouyakhf, H.: Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization. Int. J. Intell. Syst. **16**(8), 929–947 (2001)
19. Motasem, A., Joseph, D.: « Levée d'ambigüité par la méthode d'exploration contextuelle: la séquence'alif-nûn (ﻥ) en arabe » , In: Ghenima, M., Ouksel, A., Sidhom, S. (eds.) Systèmes d'Information et Intelligence Economique, 2ème Conférence Internationale (SIIE 2009), organisée par l'université de Nancy, France et l'université de la Manouba, École supérieure de commerce électronique (ESCE), Tunis, Tunis, Hammamet, 12–14 février 2009, IHE éditions, pp. 573–585 (2009)
20. Mihalcea, R., Tarau, P.: Textrank: bringing order into texts. In: Proceedings of EMNLP, pp. 404–411 (2004)
21. Page, L., Brin, L., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
22. Al-Shalabi, R., Kanaan, G., Gharaibeh, M.: Arabic text categorization using kNN algorithm. In: Proceedings of the 6th International Conference on Advanced Information Management and Service, IMS 2010. Institute of Electrical and Electronics Engineers (IEEE) (2010)