International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 208 - 213

Document decomposition for contextual indexation

Mohamed Salim EL BAZZI[1*], Abdelatif ENNAJI[2], Driss MAMMASS[3]

[1,3]*IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco*

[2]*LITIS Laboratory, University of Rouen, Rouen, France*

Abstract

The systems of features extraction from texts use a wide range of different approaches and techniques. On the one hand, this is due to the wide morphosyntactic range that the textual document possesses and the various problems that may arise during the extraction of knowledge. On the other hand, Text Mining and Understanding is very promising, and approaches and methods that can be used as precise tools still gaining momentum. In this paper, we are interested in sentence classification techniques to generate contexts. The investment in sentence grouping approaches provides great precision for indexing large documents. We validate our approach with metrics on the results of document classification.
Keywords: *Indexation, Context, sentence clustering, Classification.*

1. Introduction

The corpora have such a wide variety of styles and genres that no generic NLP tool can homogeneously handle them, that is to say, to handle all kinds of texts autonomously. The bet on text mining is it does not seek to understand the deep meaning of large quantities of texts but to effectively deal with certain precise and well-defined tasks.

Text mining is an interdisciplinary field, which relies not only on standard data mining, machine learning, and statistics but also on linguistics and natural language processing.

Our first goal is to create an extensible and generic platform for text mining tasks, which will serve as a learning aid for classification. The second objective is to develop an algorithm for indexing documents and extracting contextual information from unstructured documents.

The main problem is how one could define Doc1 and Doc2 as being linked. One possibility is to manually compare the terms to deduce when two texts are linked.

This means using supervision in an information extraction process. Supervision is very practical when you have documents labeled by categories beforehand. However, it is not easy or possible to verify similarity manually for large data corpus. For this reason, it is important and even essential to automate this process, which must be unsupervised to be able to take advantage of all the power of artificial intelligence in its axis oriented to the extraction of textual knowledge.

Similarity can be measured by wide mathematical functions, such as the cosine function. Cosine measures the degree of dependence between pairs of documents. A document pair is similar if their cosine similarity is greater than a given threshold.

This paper is ordered as follows. The following part expresses the paper background. Then, we will critic related works. After, we detail our proposed approach. Later, we explain the conducted experiments and results. Finally, we conclude by synthesizing the contributions of this work.

2. Background

The main type of features extracted from text is words. The similarity of a pair of documents depends on how words are grouped in each document. Since words compose phrases, sentences, paragraphs and documents, their strong grouping can contribute to accurate the pairs of documents similarity.

In this paper, we are particularly interested in the similarity between the sentences. The role of the term here is not overlooked because it constitutes the discriminating element for the similarity calculation, which is the product of two normalized vectors. The problem with this method is that it favors the most frequent words even if they do not have a great discriminating power.

In most cases, the optimization of the selection of terms is based on a presupposed similarity function (for example, cosine). This can be applied for the development of unsupervised methods; it proves less efficient for small textual structures like sentences.

Studies applied to the categorization of sentences become a center of interest which facilitates the discovery of similar sentences in the same document, and more generally in the whole corpus. in what follows, we rely on these techniques to form contexts belonging to a document. a context is a cluster of similar sentences.

3. Related Works

Features selection aims to extract relevant data from text to organize it into a meaningful representation of the document. If this processing is carried out correctly with a high degree of precision, all the techniques, which exploit the selection of data, will be remarkably improved. The approach we propose is based on the grouping of sentences that are semantically close to form contexts. In this part, we will focus on the work that studies the classification of sentences as being an essential axis for the rest of our work.

For a varied dataset, it is very important to pay meticulous attention to extract the most important data for classification reasons as in [1]. If the data extracted is very large, this will weaken the computation and resources. Extraction of a number that is both small and precise is highly requested.

K-means is one of the most popular methods for text clustering because of its simplicity in implementation and its relatively low level of calculation. This method has been compared to SOM in [2], which recommends the use of this last method while respecting the proportion number of neurons and number of classes.

The calculation of semantic proximity is very important for classifying texts and has a great role in the efficiency of clustering. To do this, the authors of [3] use the SOM method, arguing that it allows a faster clustering. Keywords are extracted from each document and constitute the input for SOM and the output is a vector of indexes that represents the entire document.

In the work [4], the authors compare several clustering methods. Through this study, they give the advantages and disadvantages of different methods. The remarks disclosed are very interesting. For example, they stressed the importance of using various criteria and the difficulty of finding optimal values for the parameters used, as the k used in K-Means algorithm. However, they report that it is not necessary to use different metrics to compare methods because changing the metric only changes the conclusion.

In [5], the authors propose a clustering method based on non-oriented graphs. The terms of the sentences constitute the vertices of the graphs, and the semantic relationship between the terms constitutes the edges. To proceed with clustering, they divide the clustered graph by applying algorithms that take into account the weighting of vertices and edges. These contain terms that can be from the same context.

The work [6] proposes a SiSPI system. It is composed of two processing phases, the first is the delimitation of sentences and the second is the clustering of sentences. The authors use preprocessing methods like stop words removal and stemming. Then they identify similar sentences. This system can also handle longer passages like paragraphs. To calculate the similarity SiSPI uses the cosine measure applied to the frequency vector of the terms of a sentence.

Several experiments have been carried out by the authors of [7] for documents summarizing. For preprocessing, only the removal of stop words is mentioned without stemming. The major difference in this automatic summary study is the different scheduling of the clusters to arrive at the right combination.

The paper [8] investigates the automatic clustering of sentences that describe an event in a labeled dataset. The annotations of the clusters are done by integers. The value 0 is given to the sentences, which express multiple events.

4. Proposed Approach

4.1. Document decomposition

The problem that arises is which delimiter we should choose to mark the end of each of the sequences. The processing begins with a syntactic analysis of the input, to perform operations like "stop words removal". After the system must distinguish where each sentence ends. We take the notion of sentence in its grammatical sense, which expresses an idea through the subject, the verb, and the object.

This lead to the second step that is word labeling after a syntactical analysis. This process makes it possible to associate with each word a morphosyntactic label whose granularity can go from a simple morphosyntactic category, or part of the speech, to a finer category and enriched by morphological traits. This processing aims to simplify the identification of the sentence by the system and remove any ambiguity that can be caused by a bad syntactic analysis. Some types of morphosyntactic labeling are described on the rules, while others are described in Statistics (another basic technique for processing word sequences).

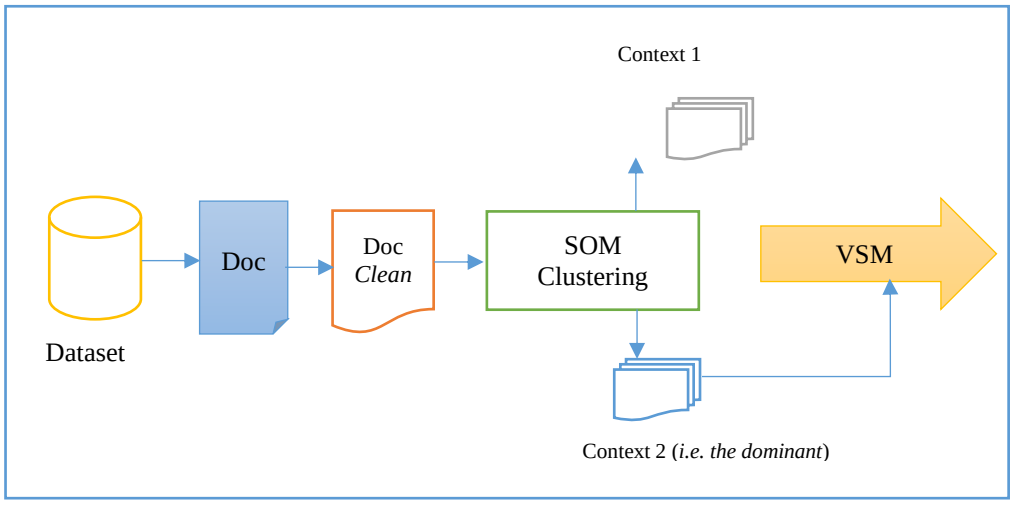

Figure 1. Document decomposition and contextualization process

Finally, terms weighting plays a very important role in the preprocessing of textual documents. The more the selection method is perfected, the more intermediate and final results of the processing will be.

4.2. Contextual indexation

In order to design an efficient context grouping process, word frequencies must be normalized according to their importance in the text and their distribution across the whole corpus. In the literature, the most used method to evaluate the importance of a word is TF-IDF formula. This decreases the importance of common words in the corpus while ensuring that the matching of documents is more impacted by the most discriminating words with low scores in the corpus. In this work, we used the method described in [6], for its clarity, and its satisfactory results.

To model the document, we use the vector model (VSM). Indeed, we start by forming the contexts, and then we calculate the score of each word of the context in order to build a vector of scores. Thus, the dominant context will represent the whole document through the corresponding vector.

$$V(D) = \{Max\ V'i\ (D),\ i \in [1, n]\} \quad (1)$$

Since the shape of the documents has changed, considering the generated contexts, we have introduced a slight modification for the TFIDF formula. Named TF-ICF (Term Frequency – Inverse Contexts Frequency), it is expressed as follows:

$$TF\text{-}ICF_{context(i)}(t) = TF_{context(i)}(t) \times \log(tf(t)_{context(i)} / C(t)) \quad (2)$$

Where t is a term of the context i , $TF_{context(i)}(t)$ is the frequency of t within the context(i) and C(t) is the occurrence of t in all contexts of the corpus.

The obvious advantage of using this method is to calculate the relevance of a term according to all contexts of the corpus. This induces to express the value of the terms judged irrelevant in the conventional TFIDF system, whereas they have a powerful discrimination role in the document [9] [10].

5. Experiments and Results

5.1. Data

During our research, we often face the problem of corpus. To approve the efficiency of an indexing system, it is essential to test it on large masses of data. The best-labeled corpora are often not open access. The corpora that have been created by researchers have two major problems: the size is relatively small for difficult texts and the areas covered are geared towards specific issues [10];

We present by the occasion a new labeled corpus, in Arabic language, of 20,000 texts, with 27,605,263 words after document pretreatment (stemming and sropwords removing), and 16 classes labeled as follows:

Table 1. Corpus details

Class label	Number of Documents
Archeology	258
Literature	1777
History and Geography	3830
Technology	585
Civilization	3306
Mathematics	442
Agriculture	895
Islamic sciences	123

Architecture	1148
Philosophy	928
Law	947
Physics – Chemistry	1530
Art	1159
Education	704
Medicine	1147
Biology	1221
Total	20.000

The corpus is collected from Arabic encyclopedias, and the particularity of having homogeneous themes and other heterogeneous for the testing of precision systems. We make this data freely available to researchers.

5.2. Results

We implemented the contextual indexing system using SOM network, combined to the TF-ICF method and we tested it on larger and large document set. We used SVM as a classifier for comparison purposes.

Table 2. SVM Classification results

Approach	**Results %**		
	precision	**recall**	**F-measure**
STANDARD TF-IDF	53.39	34.81	42.14
CONTEXTUALTF-ICF	**71.08**	**57.68**	**63.68**

We have presented the results of TF-ICF proposed method that uses contextualization based on sentence clustering with SOM algorithm as described in [6]. The results are expressed by Precision, recall and F-measure. The performance of our proposed method is effectively demonstrated according to classification results (table 2).

The results obtained by the classification are quite good, in general. This is explained by several factors, which can influence the classification result, such as the size of the test corpus, the thresholds, etc. Nevertheless, we focus in this work on the contribution of contextual indexing. However, another particular problem is added to this work. It is the language of the corpus, which is Arabic. This language has a particular difficulty that is manifested in its very rich vocabulary, the inflection of words and the great morphosyntactic variation. Furthermore, the existing works carried out on the Arabic language mainly study classification behavior, with little importance for data extraction. Our work contributes more precisely to this last point.

In most works on Arabic texts, and in the absence of a free standard corpus, the authors build their own body of work. They choose the number of categories and themes to use. For

each category, the documents are collected manually and those belonging to several categories are eliminated.

However, in order to test the accuracy of the different methods, they must be applied to the same corpus, even more, so that a method proves its effectiveness, it must be applied to several corpora with different themes.

6. Conclusion

The contextual approaches aim, on the one hand, to remove the ambiguity on the meaning of words. On the other hand, they highlight the semantic relations between these words. Semantic relationships can also be calculated using methods that evaluate the quantity of information shared between two-to-two words.

In this paper, we have introduced an approach that takes advantage of sentence categorization techniques to form contexts from similar sentences in the same document. Then, we have proposed a modeling of each document by its most dominant vector, which corresponds to the cluster that has the largest number of similar sentences. Contextual indexing has provided good results for a large corpus, also proposed in this paper.

Acknowledgments

This work was supported by LITIS Laboratory, University of Rouen, France.

References

[1] Gonçalves, C. A., Iglesias, E. L., Borrajo, L., Camacho, R., Vieira, A. S., & Gonçalves, C. T. (2019, May). Comparative study of feature selection methods for medical full text classification. In *International Work-Conference on Bioinformatics and Biomedical Engineering* (pp. 550-560). Springer, Cham.

[2] Chen, Y., Qin, B., Liu, T., Liu, Y., & Li, S. (2010). The Comparison of SOM and K-means for Text Clustering. *Computer and Information Science*, *3*(2), 268-274.

[3] Liu, Y. C., Liu, M., & Wang, X. L. (2012). *Application of self-organizing maps in text clustering: a review* (Vol. 10). chapter.

[4] Kwale, F. M., Wagacha, P. W., & Mwaura, A. (2016). A Text Clustering Comparison Methodology. *International Journal of Computer Applications*, *975*, 8887.

[5] Kotlerman, L., Dagan, I., Gorodetsky, M., & Daya, E. (2012). Sentence clustering via projection over term clusters. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (pp. 38-43).

[6] Seno, E. R. M., & Nunes, M. D. G. V. (2008, September). Some experiments on clustering similar sentences of texts in portuguese. In *International Conference on Computational Processing of the Portuguese Language* (pp. 133-142). Springer, Berlin, Heidelberg.

[7] Sarkar, K. (2009). Sentence clustering-based summarization of multiple text documents. *TECHNIA–International Journal of Computing Science and Communication Technologies*, *2*(1), 325-335.

[8] Naughton, M., Kushmerick, N., & Carthy, J. (2006, April). Clustering sentences for discovering events in news articles. In *European Conference on Information Retrieval* (pp. 535-538). Springer, Berlin, Heidelberg.

[9] El Bazzi, M. S., Mammass, D., Ennaji, A., & Zaki, T. (2018). Toward a Complex System for Context Discovery to Index Arabic Documents. *JCP*, *13*(8), 955-962.

[10] El Bazzi, M. S., Ennaji, A., & Mammass, D. (2019). ConIText: An Improved Approach for Contextual Indexation of Text Applied to Classification of Large Unstructured Data.