ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 9, No. 2, 2014, *pp*.141-153

A Kernel Hybridization NGram-Okapi for Indexing and Classification of Arabic Documents

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

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

[2] Affiliation LITIS Laboratory EA 4108, University of Rouen France

(*Received August 01, 2013, accepted December 14, 2013*)

Abstract. In this paper, we propose a hybrid system for contextual and semantic indexing of Arabic documents, bringing an improvement to classical models based on n-grams and the Okapi model. This new approach takes into account the concept of the semantic vicinity of terms. We proceed in fact by the calculation of similarity between words using an hybridization of NGRAMs-OKAPI statistical measures and a kernel function in order to identify relevant descriptors. Terminological resources such as graphs and semantic dictionaries are integrated into the system to improve the indexing and the classification processes.

Keywords: Arabic documents, classification, indexing, kernel function, n-grams, okapi.

1. Introduction

Arabic is one of the most used languages in the world, however so far there are only few studies looking for textual information in Arabic. It is considered as a difficult language to deal in the field of processing automatic language, considering its morphological and syntactic properties [2][16].

Faced with these failures, we propose a new approach based on the model of n-grams and the Okapi measure offering information extraction techniques based on portions of words. Therefore, this new method seeks to find the words which best describe the content of a document. However, the task is not easier because the management of the ambiguity in the analysis of Arabic texts (inflected language, derivation, vowel ...) is the challenge of all information retrieval systems in Arabic.

2. Related works

Compared to other languages, Arabic has a rich morphological variation and inflectional syntactic characteristics extremely complex, which is one of the main reasons for which [9][22] explains the lack of research methods in the field of treatment of Arabic.

A set of statistical models for classification and machine learning techniques have been applied to text classification : the K nearest neighbor [13][1], the decision tree [17], the Bayesian model [10], SVM model (Support Vector Machines) [11][3], SVM combined with Chi-2 for feature extraction [18][19][20], neural networks [12], Maximum Entropy[23], the distances-based classifiers [12][20][13], the knowledge-based classifiers as WordNet [7].

3. Architecture of the proposed system

3.1. Process diagram

+ Corresponding author. Tel.: +212- 5 28 22 02 67; fax: +212- 5 28 22 01 00.
E-mail address: tah_zaki@yahoo.fr.

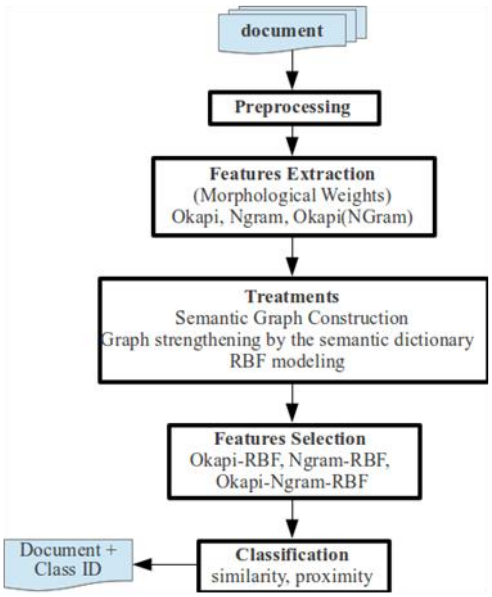

Fig. 1: The stages of the proposed indexing system.

3.2. Used corpus

During the learning phase we used a very reduced database of documents (initial corpus), labeled and representative of classes (sport, politics, economy & finances) sought to discriminate or to learn. The more this database is discriminating and representative the more our method becomes effective and showing better results.

To test our approach we used a corpus of Arabic-language press. This database is a collection of 5000 documents extracted from the Aljazeera[1] and Al Arabiya[2] sites.

Tables (1,2,3) show different results for each used measure. These results are expressed through the two criteria of recall and Precision. They show in particular the relevance of using our approach in comparison with known statistical approaches.

3.3. Preprocessing

The preprocessing phase starts by applying a noise filtering (stopwords elimination , punctuation, date) to the entire text which is followed by a morphological analysis (lemmatization, stemming) and concluded by the filtering of extracted terms. This treatment is necessary due to changes in the way in which the text can be represented in Arabic. The preparation of the text includes the following steps:

- Converting text files in UTF-16 encoding.
- Eliminating punctuation marks, diacritics and non-letters and stopwords.
- Standardizing the Arabic text, this step is to transform some characters in standard form as "آ, أ, إ" to "ا" and "ى, ي, ء" to "ئ" and "ؤ" to "و"

3.4. Space of Representation

This step allows to adopt statistical vector representation using the selected terms to best represent the document. Then, to avoid the combinatorial problems related to the dimension of the space of representation [32][8], we have adopted a frequency thresholding approach (Document Frequency Thresholding) and a principal components analysis to reduce this size.

For the choice of terms, we use a deductive method, which is to extract the vocabulary from the documents to be indexed. Therefore, we bring together a volume of documents believed to be representative of the domain, and we classify the extracted terms according to their weights.

Then we eliminate the terms deemed insignificant and out of considered domain. We distinguish thereafter between "descriptors" and "equivalent terms" (or synonyms). At the end of this phase, there is a glossary including usable descriptors and their equivalent terms for indexing and classification. Two ways for features

1 http://www.aljazeera.net/
2 http://www.alarabiya.net/