Information Processing and Management 54 (2018) 105–115

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Beyond vector space model for hierarchical Arabic text classification: A Markov chain approach

Fawaz S. Al-Anzi*, Dia AbuZeina

Department of Computer Engineering, Kuwait University, Kuwait

ARTICLE INFO

Keywords:
Arabic text
Classification
Vector space model
Markov chain
Hierarchy

ABSTRACT

The vector space model (VSM) is a textual representation method that is widely used in documents classification. However, it remains to be a space-challenging problem. One attempt to alleviate the space problem is by using dimensionality reduction techniques, however, such techniques have deficiencies such as losing some important information. In this paper, we propose a novel text classification method that neither uses VSM nor dimensionality reduction techniques. The proposed method is a space efficient method that utilizes the first order Markov model for hierarchical Arabic text classification. For each category and sub-category, a Markov chain model is prepared based on the neighboring characters sequences. The prepared models are then used for scoring documents for classification purposes. For evaluation, we used a hierarchical Arabic text data collection that contains 11,191 documents that belong to eight topics distributed into 3-levels. The experimental results show that the Markov chains based method significantly outperforms the baseline system that employs the latent semantic indexing (LSI) method. That is, the proposed method enhances the F1-measure by 3.47%. The novelty of this work lies on the idea of decomposing words into sequences of characters, which found to be a promising approach in terms of space and accuracy. Based on our best knowledge, this is the first attempt to conduct research for hierarchical Arabic text classification with such relatively large data collection.

1. Introduction

The rapid growth of online textual data raises the need for efficient information retrieval (IR) methods in terms of both time and space complexities. Text classification is the process of finding the category of a document based on the contents. Hierarchical Arabic text classification has recently received noticeable attention that is also called multi-level text classification. However, few studies tackled this domain as most of the researchers focus on flat or one-level text classification. In general, text classification employs the vector space model (VSM) that was proposed by Salton, Wong, and Yang (1975) as a model for documents and queries representations. One of the limitations of VSM is the space problem, as each document has to be represented using the entire words in the dictionary (i.e. vocabulary). Despite the number of dimensionality reduction techniques to reduce the dimensions of the textual feature vectors, however, the research is still open to employ space efficient algorithms for text classification.

In this paper, we propose to overcome the space problem by performing text classification using a space-independent text classification algorithm. That is, a method that relaxes the condition of using all words in the dictionary when creating document feature vectors. The proposed method depends on the (first order) Markov chain theory in which the neighbour characters sequences

* Corresponding author.
E-mail addresses: fawaz.alanzi@ku.edu.kw (F.S. Al-Anzi), dia.abuzeina@ku.edu.kw (D. AbuZeina).

http://dx.doi.org/10.1016/j.ipm.2017.10.003
Received 15 April 2017; Received in revised form 17 August 2017; Accepted 11 October 2017

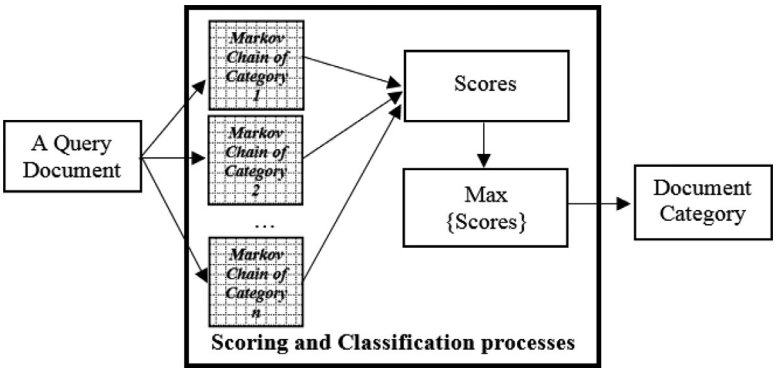

Fig. 1. The framework of the proposed method.

are used to create the probabilities transition matrices. Hence, each document is represented using a sequence of characters co-occurrences in the document. Hence, each category of the corpus is used to create a single probability transition matrix to be used in the classification process. The framework of the proposed work is described in Fig. 1. In this figure, each category represented by a Markov chain (a probability transition matrix) is used to score a query document for classification purposes. Once scored, a comparison process is performed to find the maximum score (i.e. the most likely category) among the collected scores. More details of the proposed method are found in Section 7, the proposed method.

In the next section, we present the literature review. In Section 3, the objectives is presented followed by the Markov chain background in Section 4. Section 5 presents the hierarchical text classification followed by the hierarchical data collection used in this work in Section 6. The proposed method is presented in Section 7 and the experimental results in Section 8. We present the discussion in Section 8 and, finally, we conclude in Section 9.

2. Literature review

This study includes two main topics. The first is the hierarchical Arabic text classification and the second is the Markov chain. For the first topic, the literature shows that the hierarchical Arabic text classification is still a limited research area. According to the authors' best knowledge, no research has been conducted to tackle this area for the Arabic language, however, many studies have been found to employ different classification methods for the flat Arabic text classification. For instance, Al-Anzi and AbuZeina (2017) have a thorough literature of the flat Arabic text classification. Alabbas, Al-Khateeb, and Mansour (2016) demonstrate a comprehensive study of the classifiers and the corpora of Arabic text classification. Uysal et al. (2014) presented a study for efficient use of the preprocessing tasks in text classification. In fact, most of the studies perform performance comparisons between the machine learning tools. Since the proposed method is based on the Markov chain, this literature focuses on the power of this modeling technique especially for linguistic applications. The literature shows that the Markov chain concept is used in many natural language processing (NLP) applications as shown in Table 1.

The Markov chain has also been used in other computing applications, such as software testing, DNA–sequence analysis, network path congestion, image compression, object segmentation, image retrieval, network traffic evaluation, frauds detection, evaluate groundwater quality, estimating phylogenetic trees using DNA, network service recognition, wind power, forecasting, etc.

Table 1
Some of Markov chain based linguistic applications.

Reference	Linguistic field
He, Li, and Chen (2012)	Morphological segmentation
Shen et al. (2014)	Digital document authentication
Haji et al. (2012)	Distributions of words in text lines
Goyal, Jadon, and Pujari (2013)	Documents clustering
Baomao et al. (2009)	Word segmentation
Sampathkumar, Chen, and Luo (2014)	Mining adverse drug reactions
Dowman et al. (2008)	Detecting topical structure
Meng et al. (2009)	Steganography detection
Li, Ding, and Huang (2008)	Recognizing names location
Osiek, Xexéo, and de Carvalho (2010)	Extracting acronyms and their meaning
Cai, Kulkarni, and Verdú (2006)	Text compression
Ahmed et al. (2015)	Authorship attribution
Rodrigues et al. (2013)	Inferring the location of Twitter users
Erkan et al. (2004)	Text summarization
Gao et al. (2011)	Detecting malicious attack