Latent Topic Model for Indexing Arabic Documents

Rami Ayadi, LaTice Lab, University of Sfax, Sfax, Tunisia
Mohsen Maraoui, University of Monastir, Monastir, Tunisia
Mounir Zrigui, LaTice Lab, University of Monastir, Monastir, Tunisia

ABSTRACT

In this paper, the authors present latent topic model to index and represent the Arabic text documents reflecting more semantics. Text representation in a language with high inflectional morphology such as Arabic is not a trivial task and requires some special treatments. The authors describe their approach for analyzing and preprocessing Arabic text then they describe the stemming process. Finally, the latent model (LDA) is adapted to extract Arabic latent topics, the authors extracted significant topics of all texts, each theme is described by a particular distribution of descriptors then each text is represented on the vectors of these topics. The experiment of classification is conducted on in house corpus; latent topics are learned with LDA for different topic numbers K (25, 50, 75, and 100) then they compare this result with classification in the full words space. The results show that performances, in terms of precision, recall and f-measure, of classification in the reduced topics space outperform classification in full words space and when using LSI reduction.

Keywords: Arabic Text Classification, Latent Topic Model, LDA, LSI, Preprocessing Data, Stemming, SVM, Text Representation

1. INTRODUCTION

The growing importance of electronic media for storing and disseminating text documents has created a burning need for tools and techniques that assist users in finding and extracting relevant information from large data repositories. Information management of well organized and maintained structured databases has been a focus of the Data Mining research for quite sometimes now. However, with the emergence of the World Wide Web, there is a need for extending this focus to mining information from unstructured and semi-structured information sources such as on-line news feeds, corporate archives, research papers, financial reports, medical records, e-mail messages, etc.

The collection and organization of various types of information needs automatically become new challenges and new opportunities for the field of information. When documents are classified by an automated system, people could find information and knowledge required faster. Therefore, the construction of an efficient text classification system is very necessary.
The process of text classification discusses essentially three aspects of problems: text representation, the method of classification and evaluation of the effectiveness.

The textual representation is a prerequisite for text classification. As data of TC are documents in natural language, the most important aspect in the classification is the representation or coding of texts. Generally, the text representation includes text indexing method, the method of weighting term characteristic dimension reduction method and so on.

The text indexing and weighting term solve the problem of feature text, and how to quantify them. In previous research, “word” is the term most commonly used for encoding text. However, the vocabulary of natural language is rich, which leads to a large term space in a whole text set and serious sparseness in a single document. Consequently, the research to reduce the size of feature dimensionality calls special attention. The selection of features is to select a preferred part of the set of terms to constitute a subset as a set of features for the classification task.

This paper aims to develop a method of representation and indexing text reflecting more semantics.

2. TEXT REPRESENTATION

Feature selection algorithm seeks to retain certain characteristics to optimize classification performance by removing the noise and redundancy. The feature extraction is to represent original features into another space through some sort of transformation. This transformation is a kind of representation from high-dimensional vector space to low-dimensional vector space.

Vector space model (VSM) (Salton, 1975) is still the most popular method for text representation, which reduces each document in the corpus to a vector of real numbers. Related research focuses on what are the most appropriate terms for document representation and how to calculate the weight of these terms. Much research adopt “word” or “n - gram” as terms and tf*idf as weight.

Although the reduction of tf*idf has some attractive features including the identification of words that are discriminatory for all documents in the collection, the approach also provides a relatively small amount of reduction in description length and reveals little in the way of inter or intra of the structure statistical document. To overcome these shortcomings, researchers have proposed several methods for dimensionality reduction, including latent semantic indexing (LSI) (Deerwester, 1990).

LSI uses a singular value decomposition of the matrix X to identify a subspace in the space of $tf*idf$ features that capture most of the variance in the collection. This approach can achieve significant compression in large collections. In fact, according to Deerwester et al, derived characteristics of LSI, which are linear combinations of the original $tf*idf$ features can capture some aspects of basics linguistic notions such as synonymy and polysemy.

To justify the claims about LSI, and to study its relative strengths and weaknesses, it is useful to develop a generative probabilistic model of text corpus. An important step in this regard was made by Hofmann (1999), who introduced the Probabilistic Latent Semantic Indexing (PLSi) model. The PLSi models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be regarded as representations of “topics” Thus each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and there by reduced to a probability distribution on a fixed set of topics. This distribution is the “reduced description” associated with the document. Although the work of Hofmann is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the documents. In PLSi, each document is represented as a list of numbers (the mixing proportions for topics),
RDF Storage and Querying: A Literature Review
www.igi-global.com/chapter/rdf-storage-and-querying/198561?camid=4v1a