Expert Systems with Applications 36 (2009) 10914–10918

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Automatic classification of Tamil documents using vector space model and artificial neural network

K. Rajan [a,*], V. Ramalingam [a], M. Ganesan [b], S. Palanivel [a], B. Palaniappan [a]

[a] Annamalai University, Department of Computer Science and Engineering, Annamalainagar, Chidambaram, India
[b] Centre for Advanced Studies in Linguistics, Annamalai University, Annamalainagar, Chidambaram, India

ARTICLE INFO

Keywords:
Tamil text classification
Vector space model
Artificial neural network model
Corpus building

ABSTRACT

Automatic text classification based on vector space model (VSM), artificial neural networks (ANN), K-nearest neighbor (KNN), Naives Bayes (NB) and support vector machine (SVM) have been applied on English language documents, and gained popularity among text mining and information retrieval (IR) researchers. This paper proposes the application of VSM and ANN for the classification of Tamil language documents. Tamil is morphologically rich Dravidian classical language. The development of internet led to an exponential increase in the amount of electronic documents not only in English but also other regional languages. The automatic classification of Tamil documents has not been explored in detail so far. In this paper, corpus is used to construct and test the VSM and ANN models. Methods of document representation, assigning weights that reflect the importance of each term are discussed. In a traditional word-matching based categorization system, the most popular document representation is VSM. This method needs a high dimensional space to represent the documents. The ANN classifier requires smaller number of features. The experimental results show that ANN model achieves 93.33% which is better than the performance of VSM which yields 90.33% on Tamil document classification.

1. Introduction

Today, a huge amount of information is available in online documents, e-books, journal articles, technical reports and digital libraries. Major part of this content is free form text of natural language mostly in English. The development of the internet led to an exponential increase in the amount of electronic documents not only in English, but also other regional languages. Therefore the need for automatic classification of documents is growing at a fast pace.

Automatic text classification is the task of assigning predefined categories to unclassified text documents. When an unknown document is given to the system it automatically assigns it the category which is most appropriate. The classification of textual data has practical significance in effective document management. In particular, as the amount of available online information increases, managing and retrieving these documents is difficult without proper classification.

There are two main approaches for document classification namely *Supervised* and *Unsupervised* learning. In supervised learning, the classifier is first trained with a set of training data in which documents are labeled with their category, and then the trained system is used for classifying new documents. The unsupervised learning is mainly based on clustering.

Due to the development of information technology, extensive studies have been conducted on document classification. Many statistical and machine learning techniques have been proposed for document classification such as KNN (Chiang & Chen, 2001), NB (Tan, 2005), SVM (Joachims, 1998), Neural network (Lin & Chen, 1996; Miguel & Padmini, 1998), etc.

One of the popular approaches in supervised learning is the VSM. This is based on assigning weights proportional to the document frequencies of a word in the current category as against the rest of the categories. The VSM represents the text documents as vectors where each distinct word is a separate component. It assigns some weight to each component of the vector depending on the importance of that component (Raghavan & Wong, 1986).

The application of SVM is one of the important progresses in document categorization which is very popular and proved to be one of the best algorithms for document categorization (Sebastiani, 2002). Neural network is also a popular classification method, it can handle linear and non-linear problems, for document categorization, both of the linear and non-linear classifiers achieved good results (Cheng Hua & Soon Choel, 2006). For neural network, training documents and test documents are represented as vectors.

* Corresponding author. Tel.: +91 04144 229419; fax: +91 04144 238275.
E-mail address: kaliyaperumalrajan@yahoo.co.in (K. Rajan).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.02.010

Input vectors and the corresponding target vectors are used to train until it can approximate a function, associate input vectors with specific target vectors.

The automatic classification of text plays a major role in the process of corpus building. The documents available online can be added to the corpus by proper classification of those documents. Text categorization can be used in applications where there is a flow of dynamic information that needs to be organized. In this paper, the corpus developed by Central Institute of Indian Languages (CIIL), Mysore, (CIIL Corpus) is used for training and testing the models. These models are used in the process of automatic corpus building process in which new Tamil documents are classified into one of the predefined classes and added in the corpus.

The rest of the paper is organized as follows: In Section 2, the nature of Tamil documents, and the features of Tamil corpus are provided. In Section 3 the vector space model is explained. In Section 4, how the neural network model is trained to classify the documents, is discussed. The experimental results and the performance analysis are carried out in Section 5. Concluding remarks are provided in the Section 6.

2. Tamil language

Tamil is one of the oldest languages and it belongs to the South Dravidian family. Of all Dravidian languages, Tamil has the longest literary tradition. The earliest records are cave inscriptions from the second century B.C. Tamil is a morphologically rich and agglutinative language.

Inflections are marked by suffixes attached to lexical base, which may be augmented by derivational suffixes. When morphemes or words combine, certain morphophonemic changes occur (Annamalai & Steever, 1999). Words in Tamil have a strong postpositional inflectional component. For verbs, these inflections carry information on the person, number and gender of the subject. Further, model and tense information for verb are also collocated in the inflections. For nouns, inflections serve to mark the case of the noun (Lehmann, 1993). The inflectional nature of the Tamil words prevents a simple stemming process like the one which is used for English documents. A complete morphological analysis to find the stem is also cumbersome since it requires a stem dictionary.

2.1. Tamil corpus

Tamil corpus (CIIL corpus) developed at CIIL-Mysore-India, consists of around 3.5 million words of written Tamil. The subject areas of Tamil corpus are *literature, fine arts, social science natural, physical and professional sciences, commerce, official and media languages and translated materials*. Another Tamil corpus is 'Mozhi corpus' which has 150000 sentences from wide ranging contemporary Tamil writings (Rajan, Ramalingam, & Ganesan, 2002a). The number of documents available in the CIIL corpus is shown in the Table 1.

Table 1
Tamil documents in CIIL corpus.

Major categories	Total number of documents
Social science	301
Natural Science	140
Aesthetics	188
Fine arts	36
Official and media language	57
Translated material	18
Spoken Tamil	8
Commerce	6

2.2. Feature extraction

Features for the text documents are words or phrases occurring in the documents. For text representation, in extreme case, we can consider each word as a feature. But this will result in more computation time and memory requirement. It will affect the classification accuracy as well. A careful selection of words is desired instead of all words (Marvin & Scott, 1999). A simple unordered list of words and associated weights are usually sufficient to represent a document. Studies have shown that passage meaning can be extracted without using word order (Landauer, Laham, Render, & Schreiner, 1972).

To build a document representation, a collection of documents is indexed rather than individual documents. The main goal of creating an index is to make it easy to classify documents. The size of an index can be reduced when the stems of words are used instead of all word forms (Salton, Wong, & Yang, 1975). Indexing has two subtasks, namely (i) assignment of tokens for a document (ii) assignment of weight to these tokens.

One such simple method for document indexing is defined by the following steps:

1. Find the unique words in each document in the collection of training documents.
2. Calculate the frequency of occurrence of each of these unique words for each document in the database.
3. Compute the total frequency of occurrence of each word across all documents in the database.
4. Sort the words in ascending order of their frequency.
5. Remove the words with very high and very low frequency of occurrences from the list.
6. Remove the words with invalid characters and words having less than 3 bytes.

2.3. Stop words

Noise is generally defined in IR as the insignificant, irrelevant words or stop words, which are normally present in any natural language text. Stop words have an average distribution in any standard language corpus and do not normally contribute any information to classification tasks. These stop words have high frequencies of occurrences.

2.4. Term weighting

A weight is a numerical value which is directly proportional to the importance of the word in the document. The text of each document is split into tokens and the occurrence of unique tokens in the text is listed. Only content words are considered in the index. We use the absolute count of the word occurrences in the index. This makes it difficult to compare documents of different length. The index of a document is normalized. A normalized frequency for a word is a number between 0 and 1. Each word frequency is divided by the total number of content words in the document.

3. Vector space model

Before any digital text can be processed by a machine learning (ML) classifier, a mapping must be performed on the data that is somehow able to represent the required characteristics or 'features' into a more compact and computationally appropriate form (Rijsbergen, 1979). The most established and well-known method of the document weighting approaches is the vector space model. The VSM is an approach that encodes a so-called "bag-of-words"