International Journal of Software Innovation
Volume 6 • Issue 1 • January-March 2018

A Novel Approach for Ontology-Based Feature Vector Generation for Web Text Document Classification

Mohamed K. Elhadad, Computer Engineering Department, Military Technical College, Cairo, Egypt

Khaled M. Badran, Computer Engineering Department, Military Technical College, Cairo, Egypt

Gouda I. Salama, Computer Engineering Department, Military Technical College, Cairo, Egypt

ABSTRACT

The task of extracting the used feature vector in mining tasks (classification, clustering …etc.) is considered the most important task for enhancing the text processing capabilities. This paper proposes a novel approach to be used in building the feature vector used in web text document classification process; adding semantics in the generated feature vector. This approach is based on utilizing the benefit of the hierarchal structure of the WordNet ontology, to eliminate meaningless words from the generated feature vector that has no semantic relation with any of WordNet lexical categories; this leads to the reduction of the feature vector size without losing information on the text, also enriching the feature vector by concatenating each word with its corresponding WordNet lexical category. For mining tasks, the Vector Space Model (VSM) is used to represent text documents and the Term Frequency Inverse Document Frequency (TFIDF) is used as a term weighting technique. The proposed ontology based approach was evaluated against the Principal component analysis (PCA) approach, and against an ontology based reduction technique without the process of adding semantics to the generated feature vector using several experiments with five different classifiers (SVM, JRIP, J48, Naive-Bayes, and kNN). The experimental results reveal the effectiveness of the authors' proposed approach against other traditional approaches to achieve a better classification accuracy F-measure, precision, and recall.

KEYWORDS

Dimensionality Reduction, Feature Extraction, Feature Vectr Generation, kNN, Natural Language Processing, Ontology, PCA, Semantic Similarity, Semantic Similarity Measures, Term Frequency Inverse Document Frequency, Vector Space Model, Web Text Documents Classification, Wordnet

1. INTRODUCTION

With the increasing availability of web text documents, and the rapid growth of the World Wide Web makes the task of automatic handling and processing the text documents to become an interesting area for research (Khan, 2011). The task of automatic classification of text documents, as one of mining tasks, is considered to be the key method for organizing the information and the knowledge on the web. For this task to be accomplished; web text documents are processed and transformed from the

DOI: 10.4018/IJSI.2018010101

full text version to a document vector by mapping each document into a compact form of its content, which makes the handling them much easier and to reduce their complexity (Elhadad, 2017).

The classification process for web text documents faces some challenges, mainly which are the extremely high dimensionality of text data (Elhadad, 2017), and the ignorance of the semantic information in them. (Çelik, 2013).

This curse of dimensionality leads to a lot of problems and makes the work with it more difficult as the more the size of the feature vector, the more noise appears and more errors occur. As a result, this will increase the running time, and raise the overfitting problem. So, the dimensionality of feature vector size is needed to be reduced in order to increase the efficiency, decrease the computational costs, decrease the storage cost, increase the classification performance, and ease of interpretation / modeling.

Furthermore, the ignorance of any semantics between words accordingly with the text documents themselves, makes the learning algorithms to be restricted to detect patterns in the used terminology only, while semantic meaning in the textual data remain ignored. This leads to lose important information in data and affect badly the performance of the classification system (Elhadad, 2017).

To overcome the problem of cruse dimensionality; dimensionality reduction techniques are being used. This process aims in reducing the high dimensionality of the feature vector into a lower dimensional space by selecting the best subset features from the original feature set (Elhadad, 2017; Kumbhar, 2016; Mwadulo, 2016; Zhang, 2016; Han, 2006).

This paper is organized as follows. A related work is discussed in Section 2. The proposed Ontology based feature vector generation for web text documents classification is introduced, proposing an efficient approach based on utilizing the hierarchy of WordNet ontology to add semantic information to the generated feature vector, in Section 3. Experimental results and performance evaluation are presented in Section 4. Finally, conclusions are given in Section 5.

2. RELATED WORK

In this section, we briefly review some background research including the handling of the process of extracting the feature vector from text documents, some of the previously applied techniques for web text document classification, and some previous attempts to apply semantic knowledge to enhance the classification accuracy.

In References (Rasane, 2016; Uma, 2016; Venkata Sailaja, 2016), a full review of the current trends for text documents classification, and classification algorithms are introduced and the techniques used in extracting feature vectors used in different mining tasks. Also in (Said, 2007), a comparative study between Dimensionality reduction (DR) techniques that allows users to make comprehensive choices among available techniques for enhancing automatic text categorization is conducted.

In reference (Davy, 2007), the PCA has been used as an efficient technique for dimensionality reduction for text document classification, the experimental results shows that using dimensionality reduction techniques significantly increases the performance results when using a KNN classification algorithm over two benchmark corpora (Subset of 20 Newsgroups and a Subset of Reuters-21578).it uses both Document Frequency performed Globally technique and Principal Components Analysis technique for dimensionality reduction. In both sets of experiments PCA technique was found to outperform Document Frequency performed globally technique.

Both Reference (Elberrich, 2008), and Reference (Çelik, 2013), introduces that the Traditional text classification methods do not consider the semantic relationships among words so that cannot accurately represent the meaning of documents. They introduced the use of the semantic information from ontologies, such as WordNet ontology, to improve the accuracy of text mining tasks. Also, (Abdullah Bawakid, 2010), introduces WordNet-based semantic similarity features attaining good results but the system depends on setting a threshold value when deciding about whether the similarity score for a pair of words for obtaining optimum results. And in (Wei, 2015) (Meng, 2013) (Anitha