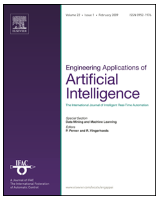
Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



A novel framework for termset selection and weighting in binary text classification

Dima Badawi, Hakan Altınçay *

Department of Computer Engineering, Eastern Mediterranean University, Famagusta, Northern Cyprus, Turkey

ARTICLE INFO

Article history:
Received 1 May 2013
Received in revised form
24 March 2014
Accepted 16 June 2014
Available online 8 July 2014

Keywords:
Co-occurrence features
Termset selection
Termset weighting
Document representation
Text categorization

ABSTRACT

This study presents a new framework for termset selection and weighting. The proposed framework is based on employing the joint occurrence statistics of pairs of terms for termset selection and weighting. More specifically, each termset is evaluated by taking into account the simultaneous or individual occurrences of the terms within the termset. Based on the idea that the occurrence of one term but not the other may also convey valuable information for discrimination, the conventionally used term selection schemes are adapted to be employed for termset selection. Similarly, the weight of a selected termset is computed as a function of the terms that occur in the document under concern where a termset is assigned a nonzero weight if either or both of the terms appear in the document. This weight estimation scheme allows evaluation of the individual occurrences of the terms and their co-occurrences separately so as to compute the document-specific weight of each termset. The proposed termset-based representation is concatenated with the bag-of-words approach to construct the document vectors. Experiments conducted on three widely used datasets have verified the effectiveness of the proposed framework.

1. Introduction

Automatic text classification is one of the key tasks in various problems such as spam filtering in which the main aim is to get rid of unwanted emails, email foldering that aims to group the incoming messages into folders and sentiment classification where the main goal is to recognize whether a document expresses a positive or negative opinion. Because of this, text categorization has become an attractive research area for many researchers in the last two decades. One of the fundamental problems in text categorization is document representation. The conventional approach is the bag-of-words (BOW) (Sebastiani, 2002). In this representation, a subset of the terms that exist in the training collection is firstly selected after sorting them using a term selection measure such as χ^2, Gini index or information gain (IG) (Chen et al., 2009; Liu et al., 2009; Yang et al., 2012). Then, the document vectors are constructed using the frequencies and inverse document frequencies ($tf \times idf$) of the selected terms where the frequency of a term denotes the number of times it occurs in the document under concern. Alternatively, as a more simple method, binary representation is used where the feature value of a term is one if it appears in the document and zero otherwise. Experiments have shown that the feature value of a term, also known as its weight, can be more effectively calculated as the product of two factors, the term frequency and the collection frequency factors where the latter is used to take into account the discriminative abilities of different terms (Debole and Sebastiani, 2003).

In the BOW-based approach, the orders of words and their syntactic relations are not taken into account. As an extension to the BOW-based approach, the use of syntactic phrases and word sequences (n-grams) that are also known as statistical phrases is studied (Mladenic and Grobelnik, 1998; Lewis, 1992b). With the use of syntactic phrases, grammatical relations are also taken into consideration. Alternatively, n-grams which are generally defined as consecutive occurrences of pairs (bigrams) or triples of terms (trigrams) are employed to extract novel features (Caropreso et al., 2001; Tan et al., 2002; Bekkerman and Allan, 2004; Mladenic and Grobelnik, 1998). The main motivation for considering phrases is that a sequence of adjacent terms may be more discriminative than the individual terms in some cases. For instance, when considered individually, the terms “bill” and “gates” in the phrase “bill gates” may not be as informative as the phrase itself about the topic of the document (Bekkerman and Allan, 2004). Taking this into account, features representing phrases are defined where a phrase is said to occur if the corresponding sequence of adjacent terms appears in the document under concern. As another alternative, the use of termsets (or, compound features, itemsets)

* Corresponding author. Tel.: +90 392 6302842; fax: +90 392 3650711.
E-mail addresses: dima.badawi@emu.edu.tr (D. Badawi), hakan.altincay@emu.edu.tr (H. Altınçay).

http://dx.doi.org/10.1016/j.engappai.2014.06.012

defined as the co-occurrences of terms having arbitrary order and position is also studied (Figueiredo et al., 2011; Tesar et al., 2006). In this approach, irrespective of their positions and order, if all terms appear, the corresponding termset is said to occur. Syntactic and statistical phrases are subsets of the set of all termsets. Since the number of termsets increases exponentially with the size of the vocabulary, termsets generally include pairs of terms but not triples. Experiments conducted on various datasets have shown that, when termsets or phrase-based features are concatenated with the BOW-based representation, better scores are generally achieved compared to the cases that exclude BOW and use only the termsets or phrases-based features (Lewis, 1992a; Boulis and Ostendorf, 2005).

As in the BOW-based approach, selection of a good subset of co-occurrence based features is important, and various criteria are utilized for this purpose. In his study on the use of syntactic phrases, Lewis (1992b) has argued that high dimensionality of the feature spaces, rare occurrence of distinct phrases and high redundancy due to synonymy are the major factors for achieving worse results compared to the BOW-based representation. Following his study, extensive work is carried out on selecting a good subset of co-occurring terms (Özgür and Güngör, 2010; Fürnkranz, 1998; Tan et al., 2002; Bekkerman and Allan, 2004). For instance, IG (Tan et al., 2002) and mutual information (MI) (Bekkerman and Allan, 2004) are used for selecting a subset of bigrams. Redundancy of features is a criterion that is considered for computing a discriminative set of features for text categorization (Baker and McCallum, 1998). This criterion is also used for selecting a good subset of bigrams. For instance, Boulis and Ostendorf (2005) argued that bigrams may not help improving the BOW representation when they are correlated with the features in the BOW-based representation, mainly due to the increased complexity especially when the training data is limited. They proposed a new measure to quantify the redundancy of a given bigram by considering the terms included in the bigram and reported improved accuracies on three different datasets. In a recent study, significant improvements compared to the BOW-based representation are achieved by applying pruning on both words and lexical dependencies (Özgür and Güngör, 2010). In fact, a weakness stated by Lewis is avoided by eliminating the rare words and the term dependencies with low occurrences. Figueiredo et al. (2011) underlined the importance of employing the most informative terms in termset generation. As a discrimination criterion, the number of classes in which the termsets appear is considered. Significantly better scores are achieved on four benchmark datasets by employing termsets of pairs of terms which are not restricted to be adjacent. The use of thresholds on the number of documents each phrase or termset appears in the training set is also considered in their selection (Figueiredo et al., 2011; Fürnkranz, 1998).

The studies mentioned above mainly aim at developing more intelligent schemes for selecting the best subset of phrases or termsets to be used together with BOW. However, in the case of BOW-based representation, term weighting is shown to be as important as selection and, various other measures such as relevance frequency and probability based scheme are proposed to replace the *idf* factor (Lan et al., 2009; Liu et al., 2009). Using these weighting schemes, it is also shown that significantly better performance scores can be achieved when compared to using binary or $tf \times idf$ based representation in the case of BOW. On the other hand, the termsets-based features are generally defined as binary where the feature value is computed as one if the corresponding termset appears (Figueiredo et al., 2011) and phrases-based features are defined as either binary or real-valued. In the case of real-valued features, only the frequencies are generally considered for their weighting.

In this study, a novel framework is proposed for selecting and weighting of termsets including non-adjacent pairs of terms. The idea is based on revising the definition of termset-based features. Consider a termset of two different terms. In the conventional representation, a termset is said to occur if both terms exist in the document. The proposed approach is based on utilizing the joint occurrence statistics of the terms for termset selection and weighting. More specifically, selecting and weighting termsets is performed by considering which term(s) occurred. The main motivation for this approach can be better explained by an example. Let us re-consider the “bill gates” example. If either of the terms is missing, the individual terms of the phrase are not as informative as the phrase itself as mentioned above. Hence, only the co-occurrence of these terms is deemed as valuable. However, there are other cases for which this phrase is not representative. For instance, consider the termset “tennis court”. It can be argued that the occurrence of both terms supports the sports topic. But, different from the previous example, the occurrence of the first term without the second term also supports the same topic. Hence, it may be useful to assign large weights to the termset in both of these cases. The occurrence of the second term but not the first may also be statistically valuable. For instance, it may signify a different topic such as law. In other words, the term “court” may not be discriminative on its own since it appears in both sports and law related documents, but it becomes more informative when evaluated together with “tennis”. It can be concluded that co-occurrence is not essential for a termset to represent valuable information. As a matter of fact, instead of focusing on only the co-occurrence of the terms, evaluation of all three possibilities in selecting and weighting termsets is promising. In this study, the joint occurrences of the individual terms within the termsets including two terms are investigated for their selection and weighting. The conventionally used selection and weighting schemes are adapted to employ this information. Experiments conducted on three widely used benchmark datasets have shown that the proposed scheme is remarkably superior to the baseline.

The rest of this paper is organized as follows. In Section 2, a brief review about the related work is presented. In Section 3, the proposed framework is described. The experiments conducted on three different datasets are presented in Section 4. The conclusions drawn and the future work are provided in Section 5.

2. Related work

In co-occurrence based document representation, there are three major steps. These steps are defining the features, selecting the best subset of these features and weighting the selected features. In this section, a literature review about the work carried out on these tasks is presented.

2.1. Definition of co-occurrence based features

The co-occurrence based features can be categorized into three groups, namely syntactic phrases, statistical phrases and termsets.

2.1.1. Syntactic phrases

Syntactic phrases are sequences of words ordered according to grammatical relations. Noun phrases, verb phrases and adjective phrases are typical syntactic phrases. The use of syntactic phrases for text classification was firstly studied by Lewis (1992b). He studied the use of BOW and syntactical phrases-based features separately and reported that syntactic phrases do not provide better scores compared to the BOW-based representation. Dumais et al. (1998) have observed that using syntactic phrases in addition to BOW generally degrades the performance achieved by using BOW alone. Scott and Matwin (1999) also noted that syntactic phrases do not provide a better representation compared to BOW. However, it is shown that voting over the outputs of the classifiers making use of BOW and phrase-based representation can provide better scores than the individual systems. This verifies that the phrases and BOW-based representations may complement each