Expert Systems with Applications 41 (2014) 735–743

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Language independent semantic kernels for short-text classification

Kwanho Kim [a], Beom-suk Chung [b], Yerim Choi [b], Seungjun Lee [b], Jae-Yoon Jung [a,*], Jonghun Park [b]

[a] *Department of Industrial and Management Systems Engineering, Kyung Hee University, Yongin, Gyeonggi 446-701, Republic of Korea*
[b] *Department of Industrial Engineering, Seoul National University, Seoul 151-744, Republic of Korea*

ARTICLE INFO

Keywords:
Short-text document classification
Kernel method
Similarity measure
Language independent semantic kernel

ABSTRACT

Short-text classification is increasingly used in a wide range of applications. However, it still remains a challenging problem due to the insufficient nature of word occurrences in short-text documents, although some recently developed methods which exploit syntactic or semantic information have enhanced performance in short-text classification. The language-dependency problem, however, caused by the heavy use of grammatical tags and lexical databases, is considered the major drawback of the previous methods when they are applied to applications in diverse languages. In this article, we propose a novel kernel, called language independent semantic (LIS) kernel, which is able to effectively compute the similarity between short-text documents without using grammatical tags and lexical databases. From the experiment results on English and Korean datasets, it is shown that the LIS kernel has better performance than several existing kernels.

1. Introduction

In the past decade, short-text documents have been widely used in various applications in diverse languages. Many recent resources on the Web regardless of the languages exist in the form of short-text documents, including Web-site summaries, document snippets, image captions, and news comments. In social networks and micro-blogging services, users usually write short-texts to describe their ideas, feelings, and opinions within a few sentences such as tweets on Twitter and status updates on Facebook (Musiał & Kazienko, 2011). In addition, the spreading of short-text documents has been limited not only to Web-based services but also to mobile applications.

Therefore, an effective method for classifying short-text documents becomes increasingly important in various research areas such as information retrieval, recommendation, and social network analysis (Leong, Lee, & Mak, 2012; Liu, Rujia, & Liufu, 2012; Tomas & Vicedo, 2012). Short-text document classification is still considered a challenging problem mainly because of the following natures of short-text documents. First, the small number of words in a short-text document is not so enough to effectively classify the documents compared to that of a lengthy text document (Taksa, Zelikovitz, & Spink, 2007). Second, the low occurrence rate of a word across documents causes the small number of words in common among documents (Sheth et al., 2005).

To address the issues, a few classification methods for short-text documents were recently proposed that attempt to calculate the similarity between documents by utilizing learning-based techniques (Faguo, Fan, Bingru, & Xingang, 2010; Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010). It is said that these methods are language dependent in that they mainly exploit grammatical tags and lexical databases to reflect syntactic or semantic features of documents.

The previous methods have limitations in extracting syntactic and semantic features and calculating the similarity between documents due to the heavy use of the grammatical tags and lexical databases which are often unavailable in many languages. As a matter of fact, only a few number of lexical databases such as WordNet (Fellbaum, 2010) and FameNet (Baker, Fillmore, & Lowe, 1998) have currently been developed, and most of them are dedicated to English and Chinese. In addition, there is no natural language processor that produces grammatical tags such as part of speech (POS) tags and predicate argument structure (PAS) tags based on syntactic analysis cannot be used for documents in most languages. Moreover, the previous methods separately address the syntactic and semantic features of documents, which might not result in the satisfactory results.

Motivated by the above remarks, we propose a kernel, called language independent semantic (LIS) kernel, which aim to effectively calculate the similarity between short-text documents by utilizing both the syntactic and semantic features of documents without relying on grammatical tags of words and ready-made lexical databases. Unlike previous kernels, the LIS kernel accommodates the two features, syntactic and semantic, in a single kernel by extracting syntactic patterns and annotating semantic information on words appearing in a document. Specifically, LIS kernel extracts the syntax-based patterns from a document. To address

* Corresponding author.
E-mail address: jyjung@khu.ac.kr (J.-Y. Jung).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.097

the small number and low occurrence rate of words problem, we considers the three levels of semantic annotations, word, document, and category, on each of syntactically extracted pattern.

The remainder of this paper is organized as follows. First, previous kernels for short-text document classification are described in Section 2. The proposed syntax-based pattern extraction and annotation methods are presented in Section 3. Experiment results are presented to show the effectiveness of the proposed kernels in Section 4. Finally, we conclude this article in Section 5.

2. Related work

Existing kernels developed for text classification can be divided into: word occurrences, syntax features, semantic features, and both syntactic and semantic features. Table 1 summarizes kernels in terms of their considered features and language constraints. The most widely used kernels for text classification is Bag-of-word (BOW) kernel, which calculates the similarity between documents based on the number of word occurrences (Joachims, 1998), which implies that it does not use any syntactic or semantic features.

Many studies have been conducted focusing on expanding feature spaces by incorporating syntactic and semantic features. Firstly, String kernel includes syntactic features by using substrings of a document for representing the document (Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002). Specifically, the main idea of string kernel is that the more substrings two documents have in common, the more similar the documents are. Next, syntactic parse tree (ST) kernel uses a syntactic parse tree of a document as its syntactic feature (Collins & Duffy, 2001). This kernel computes the similarity between documents by comparing the production of all possible pairs of nodes and counting the number of common sub-trees.

The limitation of the kernels, BOW kernel, String, and ST kernels, is that they compute the similarities based on the number of common features such as words, substrings, and sub-trees, respectively. It means that when there are few words in common among documents, they cannot show satisfactory performance.

To resolve the problem, there have been some efforts to design kernels that incorporate semantic features by using a priori semantic knowledge such as semantic smoothing (SS) and latent semantic (LS) kernels. SS kernel utilizes a lexical database called WordNet to obtain semantic features of a document (Siolas & d'Alché Buc, 2000). In LS kernel, words in a document are annotated with semantically related words which are extracted from a semantic space where the document is implicitly mapped (Cristianini, Shawe-Taylor, & Lodhi, 2002). More recently, the semantic and syntactic kernel uses predicate-argument structures to consider the lexical dependencies between words (Moschitti, 2009).

Table 1
Kernels for text classification in previous work (G and L represent grammatical tags and lexical databases, respectively).

Considered feature	Kernel	Language constraints	Author
Word occurrence	Bag-of-word (BOW) kernel	None	Joachims (1998)
Word sequence	String kernel	None	Lodhi et al. (2002)
Syntactic structure	Syntactic parse tree (ST) kernel	G	Collins and Duffy (2001)
Word similarity	Semantic smoothing (SS) kernel	L	Siolas and d'Alché Buc (2000)
Word similarity	Latent semantic (LS) kernel	None	Cristianini et al. (2002)
Syntactic structure and word similarity	Syntactic semantic tree (SST) kernel	G and L	Bloehdorn and Moschitti (2007)

Especially for short-text classification, syntactic semantic tree (SST) kernel combines both syntactic and semantic features (Bloehdorn & Moschitti, 2007). The concept of SST kernel is based on ST kernel, which represents a document as a syntactic parse tree by using grammatical tags, and SS kernel, which uses a lexical database to incorporate semantic features. In terms of language independency, BOW, String, and LS kernels are language independently applicable, whereas other kernels, ST, SS, and SST, are not applicable in some languages due to their dependency to grammatical tags and lexical databases. Accordingly, we attempt to suggest a kernel that utilizes both the syntactic and semantic features of documents without relying on grammatical tags and lexical databases.

3. Language independent semantic (LIS) kernel

In this section, we introduce a language independent semantic (LIS) kernel which is developed for short-text classification. LIS kernel composed of three parts as shown in Fig. 1: pattern extraction, semantic annotation, and similarity computation. First, it extracts patterns from a document by considering its syntactic information. Second, each extracted pattern is annotated in terms of three annotation levels, word, document, and category. Finally, LIS kernel computes the similarity between documents by using their extracted patterns according to the three annotation levels. For classification tasks, LIS kernel is then applied in a kernel machine which aims to classify new documents into one of predefined categories based on the similarities between a new document and an existing documents in the category (David Sánchez A, 2003). It provides the similarity between documents in a dataset to the kernel machine during its training and testing stages.

3.1. Syntax-based pattern extraction from short-text documents

To consider the syntactic features of documents when calculating similarity between documents, a syntax-based pattern extraction method is introduced. Here, a syntax-based pattern extracted from a document refers to a set of words appearing the document based on the syntax of a specific language. There exist some methods to extract patterns from a document based on different features of text documents. While a pattern is considered as a word

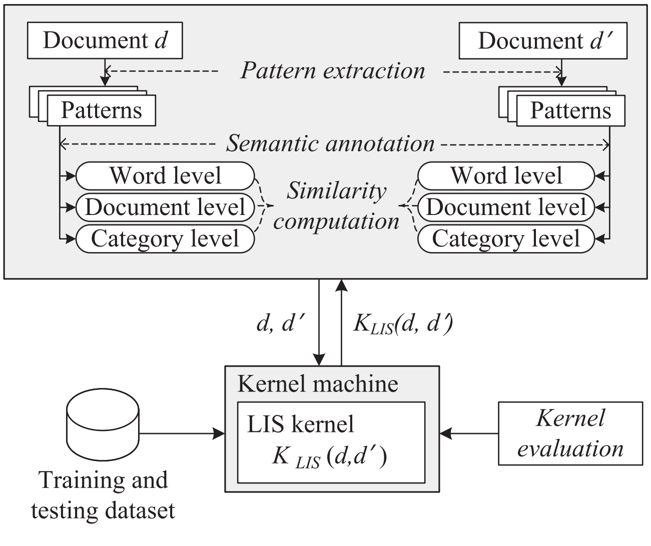

Fig. 1. An overview of LIS kernel.