Expert Systems With Applications 57 (2016) 232–247

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Ensemble of keyword extraction methods and classifiers in text classification

Aytuğ Onan[a,*], Serdar Korukoğlu[b], Hasan Bulut[b]

[a] Celal Bayar University, Department of Computer Engineering, 45140 Muradiye, Manisa, Turkey
[b] Ege University, Department of Computer Engineering, 35100 Bornova, Izmir, Turkey

ARTICLE INFO

Article history:
Received 4 January 2016
Revised 22 March 2016
Accepted 26 March 2016
Available online 29 March 2016

Keywords:
Keyword extraction
Text classification
Ensemble learning
Scientific text classification

ABSTRACT

Automatic keyword extraction is an important research direction in text mining, natural language processing and information retrieval. Keyword extraction enables us to represent text documents in a condensed way. The compact representation of documents can be helpful in several applications, such as automatic indexing, automatic summarization, automatic classification, clustering and filtering. For instance, text classification is a domain with high dimensional feature space challenge. Hence, extracting the most important/relevant words about the content of the document and using these keywords as the features can be extremely useful. In this regard, this study examines the predictive performance of five statistical keyword extraction methods (most frequent measure based keyword extraction, term frequency-inverse sentence frequency based keyword extraction, co-occurrence statistical information based keyword extraction, eccentricity-based keyword extraction and TextRank algorithm) on classification algorithms and ensemble methods for scientific text document classification (categorization). In the study, a comprehensive study of comparing base learning algorithms (Naïve Bayes, support vector machines, logistic regression and Random Forest) with five widely utilized ensemble methods (AdaBoost, Bagging, Dagging, Random Subspace and Majority Voting) is conducted. To the best of our knowledge, this is the first empirical analysis, which evaluates the effectiveness of statistical keyword extraction methods in conjunction with ensemble learning algorithms. The classification schemes are compared in terms of classification accuracy, *F*-measure and area under curve values. To validate the empirical analysis, two-way ANOVA test is employed. The experimental analysis indicates that Bagging ensemble of Random Forest with the most-frequent based keyword extraction method yields promising results for text classification. For ACM document collection, the highest average predictive performance (93.80%) is obtained with the utilization of the most frequent based keyword extraction method with Bagging ensemble of Random Forest algorithm. In general, Bagging and Random Subspace ensembles of Random Forest yield promising results. The empirical analysis indicates that the utilization of keyword-based representation of text documents in conjunction with ensemble learning can enhance the predictive performance and scalability of text classification schemes, which is of practical importance in the application fields of text classification.

1. Introduction

Automatic keyword extraction is the process of identifying key terms, key phrases, key segments or keywords from a document that can appropriately represent the subject of the document (Beliga, Mestrovic, & Martincic-Ipsic, 2015). The Web is a very rich source of information which is progressively expanding. Hence, the number of digital documents available has been progressively expanding and the manual keyword extraction can be an infeasible task. Keyword extraction is an important research direction in text mining, natural language processing and information retrieval. Since keyword extraction provides a compact representation of the document, many applications, such as automatic indexing, automatic summarization, automatic classification, automatic clustering, and automatic filtering can benefit from the keyword extraction process (Zhang et al., 2008).

Automatic keyword generation process can be broadly divided into two categories as keyword assignment and keyword extraction (Siddiqi & Sharan, 2015). In keyword assignment, a set of possible keywords is selected from a controlled vocabulary of words, whereas keyword extraction identifies the most relevant words

* Corresponding author. Tel.: +90 232 3887221, +90 544 810 70 80; fax: +90 232 3399405.
E-mail addresses: aytug.onan@cbu.edu.tr, aytugonan@hotmail.com (A. Onan), serdar.korukoglu@ege.edu.tr (S. Korukoğlu), hasan.bulut@ege.edu.tr (H. Bulut).

http://dx.doi.org/10.1016/j.eswa.2016.03.045
0957-4174/

available in the examined document (Beliga et al., 2015). Keyword extraction methods can be broadly grouped into four categories as statistical approaches, linguistic approaches, machine learning approaches and other approaches (Han & Kamber, 2006).

Text classification is an important subfield of text mining which assigns a text document into one or more predefined classes or categories. Several forms of text collections, such as news articles, digital libraries and Web pages are important sources of information (Han & Kamber, 2006). Hence, text classification is an important research direction in library science, information science and computer science (Jain, Raghuvanshi, & Shrivastava, 2012). Many applications of text mining can be modelled as a text classification problem. These applications include news filtering, organization, document organization, retrieval, opinion mining (sentiment analysis), and spam filtering (Aggarwal & Zhai, 2012).

High dimensional feature space is a typical challenge of text classification applications (Joachims, 2002). When all the words of the training documents are used as the features, text classification process becomes computationally intensive task (Onan & Korukoğlu, 2015). Hence, keywords of a text collection, which are the most important/relevant words about the content of the documents, can be good candidates to select as features in classification model construction (Liu & Wang, 2007; Rossi, Maracini, & Rezende, 2014). Machine learning algorithms, such as Naïve Bayes, k-nearest neighbour algorithm, support vector machines and artificial neural networks, have been successfully applied in classifying text documents (Sebastiani, 2002). Ensemble methods are a set of learning algorithms, which combine the decisions of these algorithms so that a more robust classification model can be built with higher predictive performance (Dietterich, 2000).

Considering these issues, this paper examines the effectiveness of statistical keyword extraction methods, base learning algorithms and ensemble methods in scientific text document classification. To the best of our knowledge, this is the first attempt, which empirically evaluates the effectiveness of statistical keyword extraction methods in conjunction with ensemble learning algorithms. In comparative evaluation, five popular ensemble methods (Boosting, Bagging, Dagging, Random Subspace and Voting) are utilized. Naïve Bayes algorithm, support vector machines, logistic regression and Random Forest algorithm are utilized as the base learning algorithms. In the experimental analysis, the domain independent statistical keyword extraction framework proposed in (Rossi et al., 2014) is utilized. In summary, the experimental study aims to answer the following research questions:

(1) Which configuration of statistical keyword extraction, classification and ensemble learning algorithms yield the highest performance in scientific text document classification?
(2) Is there an optimal number of keywords to represent the text documents and which number of keywords obtains promising results?

To the best of our knowledge, this is the first extensive empirical analysis which examines the predictive performance of statistical keyword extraction methods in conjunction with ensemble learning algorithms. The presented classification scheme, which integrates Bagging ensemble of Random Forest with the most-frequent based keyword extraction method, yields very promising results on scientific text classification. The rest of this paper is organized as follows. Section 2 briefly reviews the literature on keyword extraction and ensemble methods. Section 3 presents the statistical keyword extraction methods utilized in the experimental evaluations. Section 4 briefly describes the classification algorithms and Section 5 describes the ensemble learning methods. Section 6 presents the experimental results, discussion and statistical analysis of empirical results on ACM document collection. Section 7 presents the results of ensemble classification schemes on a larger text document collection. Finally, Section 8 presents the concluding remarks.

2. Literature review

This section briefly reviews the literature on keyword extraction methods and the ensemble methods.

2.1. Related work on keyword extraction

In statistical keyword extraction methods, statistical measures, such as n-gram statistics, word frequency and TF-IDF measure are utilized to identify keywords. The statistical keyword extraction methods can be domain-independent and do not require training data (Beliga et al., 2015). Matsuo and Ishizuka (2003) presented a statistical keyword extraction method from a single document. Initially, frequent terms are extracted. Then, co-occurrence between each term and the frequent terms are evaluated. Based on the co-occurrence distributions, the significance of a term in the document is determined. The method does not require a training corpus and can yield comparable results to TF-IDF measure. Turney (2003) presented an improved key phrase extraction algorithm, which uses statistical association among the key phrases to improve the coherence of the obtained keywords. In order to measure the association between key phrases, web mining is utilized. In another statistical keyword extraction method, text document is represented as an undirected graph (Palshikar, 2007). The vertices of the graph contains words of the document, whereas the edges are assigned values based on a statistical measure of dissimilarity between the two words.

In linguistic approaches, linguistic features of the document are utilized to identify keywords. These include lexical, syntactic, semantic and discourse analysis (Zhang et al., 2008). The linguistic keyword extraction methods are domain-dependent (Siddiqi & Sharan, 2015). Hulth (2003) examined the incorporation of linguistic knowledge, such as syntactic features to the keyword extraction process. The experimental results indicated that linguistic features can obtain improvements over the use of only statistical measures, such as term frequency or n-grams. HaCohen-Kerner (2003) presented a keyword extraction model from abstracts and titles. In the model, text representation schemes, such as unigrams, bigrams and trigrams are utilized. Nguyen and Kan (2007) presented a key phrase extraction algorithm from scientific publications. In this method, linguistic features, such as the positions of phrases in the text documents, salient morphological phenomena are taken into account. Krapivin, Autayeu, Marchese, Blanzieri, and Segata (2010) incorporated natural language processing methods to automatic key phrase extraction from scientific papers to enhance the performance of machine learning algorithms, such as support vector machines and Random Forests. The experimental results are obtained on ACM dataset. The evaluations are done with expert-assigned key phrases and key phrase extraction algorithm (KEA).

In machine learning approaches, a learning algorithm, such as support vector machines, Naïve Bayes, decision tree, is used to construct a classification model. In model construction, a training set of documents with tags are used and the model is validated via a test set of documents. The drawback of the machine learning based feature extraction models is the need to obtain a tagged set of documents. Witten, Paynter, Frank, Gutwin, and Nevill-Manning (1999) presented a simple and efficient key phrase extraction algorithm (KEA) which utilizes Naïve Bayes algorithm for domain-based key phrase extraction. In this method, possible key phrases are determined by lexical methods and good key phrases are obtained by the machine learning algorithm. HaCohen-Kerner, Gross, and Masa (2005) examined the effectiveness of several automatic