Expert Systems With Applications 84 (2017) 24–36

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering

Laith Mohammad Abualigah[a,*], Ahamad Tajudin Khader[a], Mohammed Azmi Al-Betar[b], Osama Ahmad Alomari[a]

[a] School of Computer Sciences, Universiti Sains Malaysia, 11800 Pinang, Malaysia
[b] Department of Information Technology, Al-Huson University College, Al-Balqa Applied University, P.O. Box 50, Al-Huson, Irbid, Jordan

ARTICLE INFO

Article history:
Received 18 December 2016
Revised 2 May 2017
Accepted 3 May 2017
Available online 3 May 2017

Keywords:
Feature selection
Dynamic dimension reduction
Text document clustering
Weight score
Metaheuristics

ABSTRACT

This paper proposes three feature selection algorithms with feature weight scheme and dynamic dimension reduction for the text document clustering problem. Text document clustering is a new trend in text mining; in this process, text documents are separated into several coherent clusters according to carefully selected informative features by using proper evaluation function, which usually depends on term frequency. Informative features in each document are selected using feature selection methods. Genetic algorithm (GA), harmony search (HS) algorithm, and particle swarm optimization (PSO) algorithm are the most successful feature selection methods established using a novel weighting scheme, namely, length feature weight (LFW), which depends on term frequency and appearance of features in other documents. A new dynamic dimension reduction (DDR) method is also provided to reduce the number of features used in clustering and thus improve the performance of the algorithms. Finally, k-mean, which is a popular clustering method, is used to cluster the set of text documents based on the terms (or features) obtained by dynamic reduction. Seven text mining benchmark text datasets of different sizes and complexities are evaluated. Analysis with k-mean shows that particle swarm optimization with length feature weight and dynamic reduction produces the optimal outcomes for almost all datasets tested. This paper provides new alternatives for text mining community to cluster text documents by using cohesive and informative features.

1. Introduction

With the rapid increase in the amount of electronic information on internet web pages and modern applications, text analysis in the domain of text mining requires complex techniques to deal with numerous text documents. Text clustering (TC) is one of the most efficient techniques used in text mining domain, machine learning, and pattern recognition. Text clustering is used to categorize a set of large text documents into a subset of consonant and colorant clusters (Abualigah, Khader, & Al-Betar, 2016a).

Vector space model (VSM) is used by the majority of text document paradigms in text clustering to represent each document; in this model, each term present in the text documents is featured for document representation (Abualigah & Khader, 2017). The text documents are represented by a multi-dimensional space, in which the position value of each dimension corresponds to a weight value. The text features generated from different text terms, even small documents in a collection, would be represented in hundreds and thousands of text features. However, text clustering algorithms do not perform any feature selection. Moreover, dimension reduction fails because of a vast number of text features and uninformative text features (Bharti & Singh, 2016b).

Text documents contain high dimensional informative and uninformative features, and the latter is irrelevant, redundant, unevenly distributed, and has noisy features (Zheng, Diao, & Shen, 2015). Feature selection mainly aims to determine the most informative features in text documents. However, the high dimensionality of the text document space remains an ultimate challenge. Issues relevant to removing uninformative and non-useful features and reducing the dimensionality of text documents are addressed (Pinheiro, Cavalcanti, & Ren, 2015; Shang et al., 2007). Text clustering becomes sophisticated because the document collection contains hundreds and thousands of text features. Text clustering performance is affected by the dimensionality of the documents. However, increasing the text dimensionality decreases the accuracy of

[*] Corresponding author.
E-mail addresses: lmqa15_com072@student.usm.my (L.M. Abualigah), tajudin@cs.usm.my (A.T. Khader), mohbetar@bau.edu.jo (M.A. Al-Betar), oasa14_com004@student.usm.my (O.A. Alomari).

http://dx.doi.org/10.1016/j.eswa.2017.05.002
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

the resulting clusters. Thus, eliminating uninformative text features through feature selection can improve text clustering (Liu, Kang, Yu, & Wang, 2005; Sorzano, Vargas, & Montano, 2014).

Feature selection (FS) techniques are nondeterministic polynomial-time-hard optimization methods used to determine the optimal subset of informative text features and improve the performance of the text clustering method while maintaining the necessary text information (Lin, Zhang, Huang, Hung, & Yen, 2016). Typically, these techniques are performed even without any foreknowledge of the class label of the document. Conventionally, these techniques divide the documents into three main types: feature selection based on document frequency (DF), feature selection based on the term frequency (TF), and hybrid feature technique based on document frequency and term frequency (DF-TF) (Gong, Zeng, & Zhang, 2011; Wang, Liu, Feng, & Zhu, 2015). Several text-based studies rely on feature selection methods, such as text clustering, text classification, text categorization, and text retrieval (Liu, Liu, Chen, & Ma, 2003).

The application of feature selection techniques produces a new subset with numerous informative text features. However, the dimensionality is high because of feature dispersion. This subset needs to reduce the dimension space to facilitate the process with this subset during text clustering (Uğuz, 2011). High-dimensional feature space has become a significant challenge to the domain of text mining because it increases the computational time while reducing the efficiency of text analysis techniques. Thus, a dimension reduction technique for feature space is established produce a new low-dimensional subset of useful features (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2015). This method reduces the computational time and improves the performance of the underlying text clustering algorithm. Technically, efficient dimension reduction should eliminate non-useful text features; eliminate unnecessary, redundant, and noisy text features; preserve the intrinsic information; and significantly reduce the dimension of text feature space (Bharti & Singh, 2014).

Many researchers who studied text feature selection techniques applied meta-heuristic optimization algorithms for text feature selection problems (Bolaji, Al-Betar, Awadallah, Khader, & Abualigah, 2016; Kabir, Shahjahan, & Murase, 2012). However, these methods focus on selecting a new subset of text features based on existing feature weight; as such, their performance is affected by dimension space (Aghdam, Ghasem-Aghaee, & Basiri, 2009). When the feature weight contains weaknesses in evaluating feature subsets, the method will yield unsatisfactory results (i.e., an inaccurate subset of text feature containing many uninformative features).

This paper mainly aims to propose a new feature weight scheme and a dynamic dimension space for feature selection problems and enhance the performance of feature selection methods in obtaining satisfactory results. The specific sub-objectives are as follows:

- To propose a feature weight scheme that can distinguish document features to provide high weight for features that can reflect the intrinsic document contents. To evaluate the efficiency of the proposed feature weight scheme, which utilizes three metaheuristic algorithms, namely, GA, HS, and PSO. The result of any algorithm is a subset of informative features called D^1.
- To propose a dynamic dimension reduction technique for pruning non-useful text features from the subset D^1, which leads to a new subset of reduced dimension space. The result of this stage is a new redact subset called D^2. This dynamic technique considers changing the value of the document frequency of each term, such as, a number of features in each document, maximum number of features in each document and the overall document, and number of documents that contain the same feature. Updating the term frequency to reach a new document frequency for each document.
- To utilize the redact subset D^2 in K-mean text document clustering algorithm and determine the optimal document clusters that are similar and dense.

A large group of experiments on real text datasets are commonly used in the domain of text mining to demonstrate the benefits and advantages of the proposed methods and its modification in application to text feature selection problems. The proposed methods are evaluated using eight text datasets from the Laboratory of Computational Intelligence (LABIC)[1]. The results produced by the proposed methods are compared with the results generated by same methods but without the proposed feature weight schemes and dynamic dimension reduction for non-biased comparative evaluation. PSO with LFW and DDR produced the optimal results. The proposed feature weight schemes and the effective dimension reduction are impressive additions in the text mining domain for improving the performance of both feature selection and text clustering methods.

The rest of the paper is arranged accordingly. Section 2 presents several related works in the domain of text feature selection for text clustering using meta-heuristic algorithms and dimension reduction. Section 3 illustrates the proposed methodology, defines text document and its preprocessing, explains the text feature selection problem, and describes the proposed dynamic dimension reduction technique. Section 4 shows the experiment results and discussion. Section 5 provides the conclusion and future work.

2. Related works

Feature selection techniques in text mining are used to enhance text clustering algorithms and obtain the most informative text features. Considering the search strategy to determine an informative subset of text features, existing feature selection techniques are categorized into filter, wrapper and hybrid (Alelyani, Tang, & Liu, 2013; Lin et al., 2016).

Filter methods are used for statistical analysis of text documents collected by assessing the relevance score of text features for selecting a discriminative subset of the text features without considering an interaction with the learning algorithm. Wrapper methods employ search strategy to determine a new subset of informative features and evaluate the obtained subset of features by using learning mechanisms. A new category that complements the advantages of both filter and wrapper approaches is a hybrid method for finding accurate informative subset of text features (Bharti & Singh, 2014).

Dimension reduction is an important preprocessing step in text analysis for reducing the dimension of features space by removing uninformative features; this method reduces the computational time and increases the performance of the underlying algorithm (Nebu & Joseph, 2016; Yao, Coquery, & Lê Cao, 2012). DF can be defined as follows. For a text document collection D in matrix notation, D_{nt}, where t is the number of unique text features and n is the number of all text documents. The DF value of feature t (column), DF_t, is determined as the number of text documents in which t appears at least once among the n documents. To reduce the dimensionality space of A from t to m ($m < t$) in accordance with predetermine threshold value, we select the m dimensions (features) with the top df values. Although it is simple, it is illustrated to be an efficient and effective technique in improving the performance of the text clustering algorithm and saves the

[1] http://sites.labic.icmc.usp.br/text_collections/.