Expert Systems with Applications 41 (2014) 3204–3212

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Document clustering method using dimension reduction and support vector clustering to overcome sparseness

Sunghae Jun [a], Sang-Sung Park [b,*], Dong-Sik Jang [c]

[a] Department of Statistics, Cheongju University, 298, Daeseong-ro Sangdang-gu, Cheongju, Chungbuk 360-764, Republic of Korea
[b] Graduate School of Management of Technology, Korea University, 1, 5-Ka, Anam-dong Sungbuk-ku, Seoul 136-701, Republic of Korea
[c] Division of Industrial Management Engineering, Korea University, 1, 5-Ka, Anam-dong Sungbuk-ku, Seoul 136-701, Republic of Korea

ARTICLE INFO

Keywords:
Document clustering
Sparseness problem
Patent clustering
Dimension reduction
K-means clustering based on support vector clustering
Silhouette measure

ABSTRACT

Many studies on developing technologies have been published as articles, papers, or patents. We use and analyze these documents to find scientific and technological trends. In this paper, we consider document clustering as a method of document data analysis. In general, we have trouble analyzing documents directly because document data are not suitable for statistical and machine learning methods of analysis. Therefore, we have to transform document data into structured data for analytical purposes. For this process, we use text mining techniques. The structured data are very sparse, and hence, it is difficult to analyze them. This study proposes a new method to overcome the sparsity problem of document clustering. We build a combined clustering method using dimension reduction and *K*-means clustering based on support vector clustering and Silhouette measure. In particular, we attempt to overcome the sparseness in patent document clustering. To verify the efficacy of our work, we first conduct an experiment using news data from the machine learning repository of the University of California at Irvine. Second, using patent documents retrieved from the United States Patent and Trademark Office, we carry out patent clustering for technology forecasting.

1. Introduction

Technological literature is mainly based on document data such as patents (Han & Kamber, 2005; Roper et al., 2011; Subramanian & Soh, 2010). Therefore, we analyze patent documents to discover future technological trends (Amadi-Echendu, Lephauphau, Maswanganyi, & Mkhize, 2011; Badawy, 2009; Geum, Lee, Kang, & Park, 2011). Document (text) data analysis requires more delicate techniques than numerical data analysis, which uses statistics and machine learning (Andrews & Fox, 2007). Most statistical analysis and machine learning algorithms are focused on numeric data types (Han & Kamber, 2005). These methods are not suitable for document data (Han & Kamber, 2005; Tseng, Lin, & Lin, 2007). Therefore, documents have to be converted into structured data that are suitable for analytical models. In this paper, we transform document data to document-term matrix (DTM) using text mining techniques. The transformed document data is structured data consisting of documents and terms as the rows and columns of the DTM, respectively. The elements of this matrix signify the frequency of occurrence of each term in each document. In general, the size of columns (terms) is larger than that of the rows (documents). In addition, many of the elements in the matrix have a value of zero. Although DTMs are suitable for statistics and machine learning, it is difficult to analyze them because they have a very sparse data structure. Andrews (2007) and Feinerer, Hornik, and Meyer (2008) published research results indicating that they solved this problem. They focused on building the structured data, and used basic methods such as descriptive statistics and visualization to analyze the data. However, these research efforts were limited to the prediction of future trends in technology and the discovery of vacant technology, because technological documents such as patents have complicated data types including text, date, number, picture, and so on (Choi, Park, Kang, Lee, & Kim, 2012; Lee, Lee, & Yoon, 2011; Soo, Lin, Yang, Lin, & Cheng, 2006; Trappey, Hsu, Trappey, & Lin, 2006). So, we need to develop an efficient clustering method for patent or technology documents in this paper.

Our research also proposes a new method to overcome the sparseness in document data clustering. We perform dimension reduction by combining singular value decomposition (SVD) and principal-component analysis (PCA). We call this technique SVD–PCA. Ding and He (2004) showed that PCA is an efficient approach to dimension reduction for clustering such as *K*-means clustering algorithm. Also, SVD was used for efficient document classification (Li & Park, 2009). We present an approach that removes the sparseness of DTM by using SVD–PCA. Our research combines SVD and PCA for solving the sparsity of document clustering. We then apply

[*] Corresponding author. Tel.: +82 2 3290 3900; fax: +82 2 953 4750.
E-mail addresses: shjun@cju.ac.kr (S. Jun), hanyul@korea.ac.kr (S.-S. Park), jang@korea.ac.kr (D.-S. Jang).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.11.018

support vector clustering (SVC) and Silhouette measure to the *K*-means clustering method for effective document clustering, and verify the efficacy of our proposed approach using news data and patent documents. First, we conduct an experiment using Reuters data from the University of California at Irvine's (UCI) machine learning repository (University of California – Irvine, 2011). In the experiment, we demonstrate the basic operation of our proposed method. Next, we conduct experiments using two patent data sets from the United States Patent and Trademark Office (USPTO) (The United States Patent, 2011). One is used to give a more detailed illustration of how our method works, while the other is used to validate the performance of the proposed model.

Next, in the literature review section, we review related work such as the dimension–reduction method, the sparseness problem of document data, and clustering methods such as SVC and Silhouette measure. We present our clustering model for sparse document data and describe how to solve the sparseness problem of document data in Section 3. In Section 4, we examine our experimental results, and show how they confirm the efficacy of this research. Finally, we conclude this paper and describe our future work in Section 5.

2. Literature review

Clustering has been used in many fields and by diverse approaches (Ding & He, 2004; Pan & Zhang, 2011; Zhong & Zhang, 2011). Recently, clustering was applied to document data analysis that was one of big data learning (Aliguliyev, 2009; Isa, Kallimani, & Lee, 2009; Maziere & Hulle, 2011; Saracoglu, Tutuncu, & Allahverdi, 2007; Tseng, 2010). Document clustering is a clustering approach for text-based data (Andrews, 2007). It is used in many applications including web information retrieval, natural language processing, bioinformatics, and technology analysis (Andrews, 2007; Chow, Zhang, & Rahman, 2009; Jun, Park, & Jang, 2012). Document clustering is also used to analyze scientific publications, such as articles, papers, and patents (Jun et al., 2012; Lee et al., 2011; Soo et al., 2006; Trappey et al., 2006). Technological trends and associations can be discovered by analyzing these document clusters. For example, in the analysis of scientific documents, many researchers use the citation information in patent documents (Duplenko & Burchinsky, 1995). Recently, a research method for core documents using co-citation information was introduced (Glänzel & Thijs, 2011). In clustering, text-based document data are not structured for general clustering methods such as those based on statistics and machine learning (Jun et al., 2012; Tseng et al., 2007). As a result, clustering models specific to document data are needed (Chen, Tai, Harrison, & Pan, 2005). Currently, there are some problems that need to be solved in relation to document clustering. First, at present the number of clusters has to be determined before clustering. In general, this number is selected by researchers based on their knowledge of the subject area (Everitt, Landau, & Leese, 2001; Han & Kamber, 2005; Jun & Uhm, 2010). There has also been a lot of research published calling for the objective selection of the number of clusters (Everitt et al., 2001; Jun & Uhm, 2010; Rousseeuw, 1987; Wang, Leckie, Ramamohanarao, & Bezdek, 2009). However, we have not found any general method for document clustering in existing research results. To solve the problem, we consider two approaches: specifically, SVC and Silhouette measure. The second problem is the fact that the structure of text-based document data is not suitable for traditional clustering methods based on statistics and machine learning, and so it must be converted. Tseng et al. (2007) introduced text mining techniques for transforming documents into structured data for statistical analysis or machine learning. In addition, Feinerer et al. (2008) proposed text mining as a preprocessing method for document data. In this paper, we construct a DTM from given documents to settle the second problem. The rows and columns of DTM represent documents and terms that occur, respectively. Moreover, each element of DTM shows the frequency of the terms that occur in each document. At this point, we encounter the third problem associated with document clustering: specifically, the fact that most cells have a value of zero, so the data structure of DTM is sparse. We solve this problem using SVD–PCA, and thereby introduce a method to overcome the three problems associated with document clustering.

SVC considered in our research is an efficient clustering algorithm that is based on statistical learning theory (Ben-Hur, Horn, Siegelmann, & Vapnik, 2001; Karatzoglou, Meyer, & Hornik, 2006; Kees, Marchiori, & Vaart, 2003; Lee & Lee, 2005; Puma-Villanueva, Bezerra, Lima, & Zuben, 2005; Vapnik, 1998). Also, SVC showed good performance in document analysis (Hao, Chiang, & Tu, 2007). In SVC, all data points are mapped from the data space to a feature space. The feature space is larger in dimension than the data space. The general mapping function used by SVC is the Gaussian kernel. In our method, we search for the smallest sphere enclosing data points in the feature space. This sphere is mapped back to the data space to form contours. These contours enclose the data points for cluster boundaries. In data space, the data points enclosed by each contour are considered to be in the same cluster. In our research, we use SVC for patent document data clustering. In addition, we determine the number of clusters present in the SVC result using Silhouette width measure (Rousseeuw, 1987). The number of clusters is needed for clustering methods such as *K*-means clustering and the hierarchical method. For example, to use *K*-means clustering, the number of clusters (*K*) has to be specified. Thus, the result of *K*-means clustering is affected by this number. In our work, we apply the results of SVC and Silhouette to *K*-means clustering for efficient document clustering. In our study, we use SVC to cluster document data.

3. Methodology

Thus far, we have outlined some problems associated with document clustering. One of the problems is sparseness. In this paper, we propose a document clustering method that solves this problem. First, we convert documents to structured data for document clustering. In this process, we employ preprocessing using text mining techniques. In other words, we transform document data into structured data using document parsing and term extraction methods. We then obtain a DTM from document data via the document preprocessing (Tseng et al., 2007). As mentioned in Section 2, DTM form is very sparse. To remove this sparsity from the DTM, we change the DTM to document principal component matrix (DPCM) form. Fig. 1 shows our DTM and DPCM forms.

In Fig. 1(a), the rows and columns of the DTM represent documents and terms, respectively. Further, *freq ij* (an element of the DTM) expresses the frequency of occurrence of *Term j* in *Document i*. In general, most elements of a DTM have a value of zero. In addition, the number of terms (variable) is greater than the number of documents (observation). However, in general data analysis by statistics and machine learning, the observation is greater than the variable. To settle the sparseness problem as well as that of the dimension, we use a dimension–reduction method to construct the equivalent DPCM shown in Fig. 1(b). In our research, we utilize PCA as our dimension–reduction approach. Moreover, *score ij* is the *jth* PC value of *document i*. Most DPCM element values are not zero because they are PCs. Of course, m is much smaller than p ($m \ll p$). For example, let us assume that we have four documents and we consider two PCs. Fig. 2 shows our approach.