Expert Systems with Applications 41 (2014) 8203–8210

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A three-phase approach to document clustering based on topic significance degree

Yinglong Ma [a,*], Yao Wang [a], Beihong Jin [b]

[a] School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, PR China
[b] Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences, Beijing 100190, PR China

ARTICLE INFO

Article history:
Available online 15 July 2014

Keywords:
Document clustering
Topic model
K-means
K-means++

ABSTRACT

Topic model can project documents into a topic space which facilitates effective document clustering. Selecting a good topic model and improving clustering performance are two highly correlated problems for topic based document clustering. In this paper, we propose a three-phase approach to topic based document clustering. In the first phase, we determine the best topic model and present a formal concept about significance degree of topics and some topic selection criteria, through which we can find the best number of the most suitable topics from the original topic model discovered by LDA. Then, we choose the initial clustering centers by using the k-means++ algorithm. In the third phase, we take the obtained initial clustering centers and use the k-means algorithm for document clustering. Three clustering solutions based on the three phase approach are used for document clustering. The related experiments of the three solutions are made for comparing and illustrating the effectiveness and efficiency of our approach.

1. Introduction

In the last decades, clustering has found numerous applications in the text domain such as document organization, classification, summarization, browsing and retrieval (Aggarwal & Zhai, 2012; Cai, He, & Han, 2011; Lu, Mei, & Zhai, 2011; Ng, Jordan, & Weiss, 2002; Xu & Gong, 2004; Xu, Liu, & Gong, 2003; Larsen & Aone, 1999). Document clustering aims to organize similar documents into clusters, so that documents in the same cluster are similar and documents in different clusters are distinct. The traditional clustering methods usually represent documents by a bag-of-words (BOW) model purely based on raw terms. The latent semantic information residing in document corpus is hard to be captured. Topic models have received more attention in text domain (Xie & Xing, 2013; Blei, 2012; Hofmann, 2001) in the last years. The use of topic model such as LDA (Blei, Ng, & Jordan, 2003), can organize words with similar semantics and further associates them with the same semantic concept called topic. Corpus can be projected into a topic space. As such, topic based semantic information will be sufficiently used when clustering is used to identify document clusters. Over the last decades, clustering analysis with topic modeling has demonstrated its vast success in modeling and analyzing texts.

Document clustering essentially is highly associated with topic models. A good topic model for document clustering can reduce the noise of similarity measure and identify the grouping structure of the corpus more effectively. On one hand, it is crucial for document clustering to determine the best topic model in which every topic should be understandable, meaningful and semantically compact, and can be discriminated from each other. The key problem is to determine the number of the most suitable topics from a topic model. In most of existing approaches such as LDA, etc., statistical models are used to discover a beforehand specified number of topics, which is closely associated with the layer of the topic structure. Unfortunately, the specified number of the discovered topics often makes topic structure hard to understand and discriminate from each other because some of them are possibly trivial and irrelevant to characterize genuine theme and semantic concepts of the domain. Although some methods such as Blei and Lafferty (2006), Li and McCallum (2006), AlSumait, Barbara, Gentle, and Domeniconi (2009) and Wang, Wei, and Yuan (2011) are proposed to model correlations between topics, they failed to resolve the problem of how to determine the number of the most suitable topics.

On the other hand, a good topic model should contribute to improving efficiency and accuracy of clustering. Most of existing clustering approaches are achieved based on finite dimensional vector space, so the problem of efficiency and accuracy of clustering should be taken into account. Some popular clustering methods,

* Corresponding author. Tel.: +86 10 61772643.
E-mail address: yinglongma@gmail.com (Y. Ma).

http://dx.doi.org/10.1016/j.eswa.2014.07.014
0957-4174/

such as the k-means (Lloyd, 1982; Gan, Ma, & Wu, 2007; Hamerly, 2010; Gao & Hitchcock, 2010), etc., have high computational complexity and require more running time based on BOW model in which every document vector often has a very large dimension over word terms. In contrast, every vector in topic model has a dramatically reduced dimension, so the required computational complexity and running time for clustering will be also dramatically reduced. However, existing clustering approaches do not consider to take advantage of the merit of the best topic model that can further effectively reduce dimension over topics and in which each of the topics are semantically compact and significant. We argue that clustering based on the best topic model can further improve efficiency and accuracy of clustering. This paper will address the two related problems.

In this paper, we propose a method for determining the best number of topic model, and achieve the document clustering based on the best topic model. To the best of our knowledge, there is little work currently made for document clustering based on the best topic model. The contributions of this paper are as follows.

1. We propose a three phase document clustering approach based on the best topic model. First, we determine the best number of topic model by presenting a novel concept about significance degree of topics with respect to documents. The most significant topics are selected from original topics discovered by LDA. Second, we use the k-means++ algorithm to choose the initial clustering centers. In the third phase, k-means method is used for topic based document clustering.
2. We present a definition of significance degree of topics with respect to documents for determining the best number of topics.
3. But not the least, the related experiments based on three clustering solutions are made for illustrating and comparing their effectiveness and performance of our approach.

This paper is organized as follows. Section 2 introduces the k-means++ clustering and LDA. In Section 4, we give an overview of our three-phase approach. Section 5 proposes a method to determine the best number of topics by the concept of significance degree of topics and some topic selection criteria. Section 6 is to discuss how to choose initial centers over different topic space. In Section 7, document clustering is made based on different topic models. Section 8 is the related experiments and evaluation based on three clustering solutions. Section 9 is the conclusion and the future work.

2. Related work

This paper aims to achieve document clustering based on the best topic model. There are many clustering approaches proposed for document clustering. K-means clustering (Lloyd, 1982; Gan et al., 2007; Hamerly, 2010) is regarded as the most popular partitional method (Gao & Hitchcock, 2010), and has been widely applied in many fields such as information retrieval (Aggarwal & Zhai, 2012; Kumar & Srinivas, 2010), medicine (Zheng, Yoon, & Lam, 2014), and data management (Wei, Lee, & Chen, 2013), etc. K-means algorithm is highly precarious to initial cluster centers. There are two main classes of ongoing work made to optimize the original k-means algorithm (Lloyd, 1982). The first class focuses on accelerating k-means clustering (Hamerly, 2010; Drake & Greg Hamerly, 2012; Elkan, 2003; Agarwal & Mustafa, 2004; Kanungo et al., 2002) s. Another class of research work focuses on providing an initialization leading to a high-quality solution (Pena, Lozano, & Larranaga, 1999; Agha & Ashour, 2012; Zhang, 2012; Emre Celebi, Kingravi, & Vela, 2013). One of the state of the art approaches is the k-means++ algorithm (Arthur & Vassilvitskii, 2007), which simply extends k-means by seeding the initial cluster centers. It can obtain optimal clustering, and its simplicity and speed is practically appealing. So the k-means++ algorithm is selected and used for document clustering in this paper.

Our three phase document clustering is different from some existing two-phase k-means clustering algorithms such as Jiang (2001), Nguyen, Nguyen, and Pham (2013) and Pham, Dimov, and Nguyen (2004). In Jiang (2001), a two-phase clustering algorithm was proposed for outliers detection. The traditional k-means algorithm was modified by building a minimum spanning tree (MST) (Pham et al., 2004). Nguyen et al. (2013) used a buffering technique and developed a scalable k-means algorithm. Our three phase clustering begins with the selection of the best number of topics, and uses the k-means++ algorithm for document clustering based the best topic model.

Topic models are used to identify and extract the semantic concepts in text documents and uncover the latent semantic structure embedded in document collections. In the last decade, many topic models have been proposed such as PLSI (Hofmann, 1999), LDA (Blei et al., 2003), and Pachinko allocation (Li & McCallum, 2006), etc. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one of the most popular methods currently in use, allowing documents to have a mixture of topics. It is based on the Bayesian model and makes use of latent variables to represent the semantic concepts (i.e.. topics). Topics discovered by LDA are independent from each other because LDA assumes the topic proportions are randomly drawn from a Dirichlet distribution. However, this assumption is not always true because topic correlation is very common in the real world data. There are also much work to explore the correlations between topics, such as CTM (Blei & Lafferty, 2006), Pachinko allocation (Li & McCallum, 2006), HTMM (Gruber, Rosen-Zvi, & Weiss, 2007), ITM (Hu, Boyd-Graber, & Satinoff, 2011), and IFTM (Putthividhya, Attias, & Nagarajan, 2009), etc. Unfortunately, most existing approaches need to resolve the problem of how to determine the number of the most suitable topics.

In this paper, we present an approach to selecting the most significant topics by calculating significance degrees of topics for accelerating clustering and improving clustering accuracy. Cao, Xia, Li, Zhang, and Tang (2009) proposes a method of adaptively selecting the best LDA model based on density calculated by average cosine distance, but the selection of topics is not made in terms of their significance, and no criterion is given for topic ranking. AlSumait et al. (2009) measures the distance between a topic distribution and a junk distribution, and a four-phase Weighted Combination approach is used to rank the significance of topics. Wang et al. (2011) proposes two topic significance re-ranking methods: Topic Coverage (TC) and Topic Similarity (TS). However, they do not discuss the problem of how to determine the best topic model that is crucial to reduce the dimension of vector space, and accelerate the clustering.

To the best of our knowledge, this is the first work for the k-means++ document clustering based on the best topic model by topic significance ranking and topic selection criterion.

3. Preliminaries

3.1. LDA topic model

In LDA, documents are viewed as a distribution over topics while each topic is a distribution over words. It firstly samples a document-specific multinomial distribution over topics from a Dirichlet distribution, and then repeatedly samples the words from these topics. The posterior probability over the latent variables and