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

model parameters is calculated to extract the latent semantic structures in documents. By the Gibbs sampling (Bolstad, 2010), we sample the posterior distribution, and obtain the two matrices θ and φ. Assume that K is number of topics, V is the number of words in the vocabulary, M is the number of documents, and N_j is the number of words in document j. The total probability of the model is described as follows: $P(\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{\varphi};\alpha,\beta)=\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{w=1}^{N_j}P(Z_{j,w}|\theta_j)P(W_{j,w}|\varphi_{Z_{j,w}})$.

Where θ_j is the distribution of topics in document j and is a K-dimension vector with $\sum_{i=1}^{K}\theta_{j,i}=1$; $Z_{j,w}$ is the identity of topic of word w in document j; φ_i is the distribution of words in topic i; $W_{j,w}$ is the identity of word w in document j. $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_K)$, and $\beta=(\beta_1,\beta_2,\ldots,\beta_V)$, where α_i and β_w $(1\leqslant i\leqslant K, 1\leqslant w\leqslant V)$ are the prior weights of topic i in a document and word w in a topic, respectively. Both α_i and β_w are positive real numbers less than 1. $\boldsymbol{W}$ and $\boldsymbol{Z}$ are respectively the matrices of variables $W_{j,w}$ and $Z_{j,w}$. θ is the Document-Topic matrix and φ is the Topic-Word matrix.

3.2. K-means++ clustering

K-means++ algorithm provides a way to choose initial centers for the k-means algorithm. Let $\mathcal{D}$ be a set of data points, and K be the number of specified centers. let $d(x)$ be the shortest distance from data point x to the closest center. K-means algorithm is described as follows.

1. Randomly take one data point $c_1\in\mathcal{D}$ as the first centroid,
2. Take a data point $c\in\mathcal{D}$ with the highest probability $\frac{d(x)}{\sum_{x\in\mathcal{D}}d(x)}$ as a new centroid,
3. Repeat Step 2 until all the K centers are taken,
4. Taking the K centroids, we proceed with the standard k-means algorithm for clustering.

The standard k-means algorithm is described as follows.

1. Take the K initial centroids obtained in the previous.
2. Assign each point $c\in\mathcal{D}$ to the cluster that has the closest centroid.
3. Calculate the new multivariate mean (or centroid) for each cluster;
4. Repeat Step 2, until the algorithm stops when the means of the clusters is constant from one iteration to the next.

4. Overview of approach

A sketchy description of our three phase approach to document clustering based on the best topic model is illustrated in Fig. 1. Before starting the three phase clustering approach, the corpus is first preprocessed by a series of text preprocessing, and we can obtain the matrix of text word vectors. Then, the number N of topics is beforehand specified. We used LDA model and Gibbs sampling to analyze the given corpus, and obtain the Document-Topic matrix and the Topic-Word matrix. The N topics are finally discovered by LDA. In the following, our three phase clustering approach will start.

From top to bottom in Fig. 1, document clustering will be sequently made by the following three phases represented by dashed boxes: determining the best topic model, seeding the initial centers and the k-means clustering. In Phase 1, we generate the Document-Topic significance matrix by calculating significance degree of topics with respect to a corpus, taking the Document-Topic matrix obtained by LDA as input. All the topics will be ranked according to their significance degrees. Then, the M topics with the higher significance degrees will be selected to build the best topic model by some criterion. M is just the best number of topics discovered by LDA. In Phase 2, we beforehand specify K clusters for a given corpus, and then seed the initial centers using k-means++. Phase 3 is to perform the k-means clustering by taking the chosen initial centers.

From the left side to the right in Fig. 1, three document clustering solutions based on our three phase approach will be achieved. Solution 1 is to determine the best number M of topics, seed the initial centers and perform the k-means clustering over the M-dimensional vectors. In contrast, Solution 2 is to seed the initial centers over the M-dimensional vectors and perform the k-means clustering over the N-dimensional space. In Solution 3, the best topic model is not needed, and the k-means clustering is directly performed for the N-dimensional document vectors over topics.

In this paper, we prefer to use Solution 1 for document clustering. The other two solutions are used as complementary for comparing effectiveness and efficiency of document clustering.

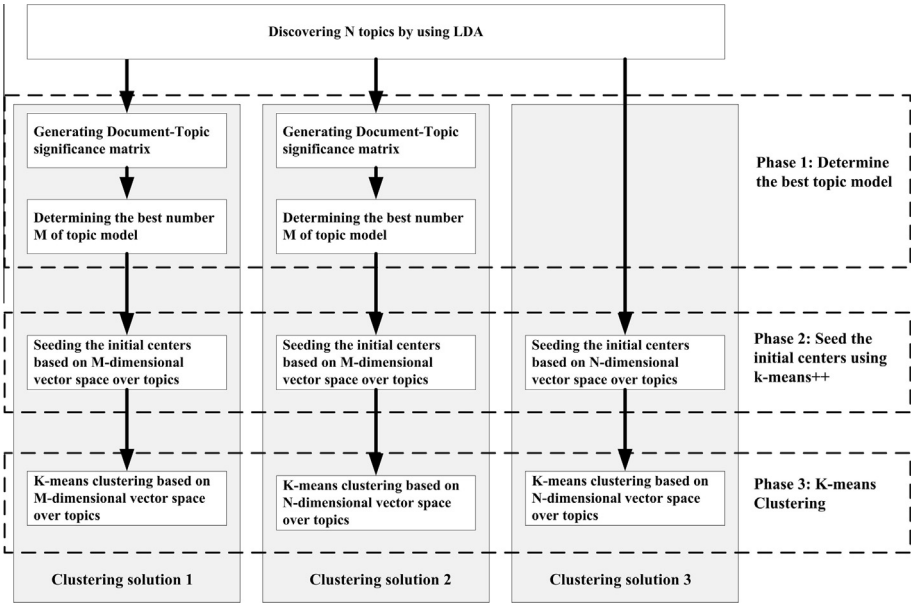

Fig. 1. Overview of topic based k-means clustering.

5. Determining the best topic model

In the following, we introduce the use of notions used in the remaining of this paper. Let $\mathcal{C}$ and $\mathcal{D}$ respectively be the sets of clusters and documents for the given corpus. We used LDA model and Gibbs sampling for topic modeling. Let $\mathcal{T}$ be the set of topics discovered by LDA, and $\mathcal{T}'$ be the set of the topics selected from $\mathcal{T}$. The obtained Document-Topic matrix and the Topic-Word matrix are denoted as ***DT*** and ***TW***, respectively. We denote that $|\mathcal{C}| = K, |\mathcal{D}| = D, |\mathcal{T}| = N$ and $|\mathcal{T}'| = M$, where the notion $|\bullet|$ is to represent the cardinality of a set. We use italic upper case letters for constants, and italic lower case letters for variables.

In order to generate the Document-Topic significance matrix ***M***, we take the Document-Topic matrix ***DT*** as input. Matrix ***DT*** is a $D \times N$ matrix, where each raw corresponding to a document represents a distribution over N topics discovered by LDA.

5.1. Defining topic significance degree

Definition 1. (*Significance Degree of Topic t_i with respect to Document d_j*, $Sig(d_j, t_i)$) Let $t_i \in \mathcal{T}$ be a topic, and $d_j \in \mathcal{D}$ be a document, where $1 \leqslant i \leqslant N, 1 \leqslant j \leqslant D$, and $d_j = (t_1^j, t_2^j, \ldots, t_N^j)$ is a probabilistic distribution over $\mathcal{T}$. The significance degree of topic t_i with respect to document d_j is denoted as $Sig(d_j, t_i) = T(t_i, d_j) * I(t_i, d_j)$, where

$T(t_i, d_j)$ is the conditional probability, i.e.. $T(t_i, d_j) = P(t_i|d_j) = \frac{t_i^j}{\Sigma_{k=1}^N t_k^j}$.

$I(t_i, d_j)$ is defined as $I(t_i, d_j) = \log \frac{1}{P(d_j|t_i)} = \log \frac{\Sigma_{k=1}^D t_i^k}{t_i^j}$.

In Definition 1, $T(t_i, d_j)$ is to examinate the probability that Topic t_i is significant to Document d_j over the N topics. The larger $T(t_i, d_j)$ is, the more significant t_i is to d_j. $I(t_i, d_j)$ is to indicate whether Topic t_i is common or rare to Document d_j across all the D documents. The larger $I(t_i, d_j)$ is, the less common t_i is to d_j. According to Definition 1, we can generate a Document-Topic matrix which reflects the significance degree of each topic with respect to each document.

Definition 2. (*Document-Topic Significance Matrix*, ***M***) The Document-Topic significance matrix ***M*** is a $D \times N$ matrix, where each column corresponding to a document represents a distribution over N topics. Each element $\boldsymbol{M}(i,j)$ in matrix ***M*** is defined $\boldsymbol{M}(i,j) = Sig(d_i, t_j)$.

Definition 3. (*Significance Degree of Topic t_i*, $Sig(t_i)$) The significance degree of topic t_i with respect to the set of all the documents is defined as $Sig(t_i)$, and can be defined as $Sig(t_i) = \Sigma_{k=1}^D Sig(d_k, t_i)$.

By the above definitions, we can easily construct the Document-Topic significance matrix ***M***. According to Definition 3, we can further determine whether a topic is significant to all the document in the corpus by evaluating its significance degree. The larger significance degree a topic has, the more important it is to characterize features of documents.

5.2. Determining the best M

For a given corpus, we can easily rank all the N topics according to their significance degrees. In the following, we give the criteria for determining the best number M from the N topics discovered by LDA.

Definition 4. (*Topic Selection Criterion*) Let $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$ be the set of original topics. For each $1 \leqslant i \leqslant N-1$, we have $Sig(t_i) \leqslant Sig(t_{i+1})$. There exists the best number M such that $\frac{\Sigma_{i=1}^M Sig(t)}{\Sigma_{i=1}^N Sig(t)} \geqslant \Psi$, where Ψ is a threshold value.

In the definition, Ψ is a threshold of the ratio of the total significance degrees of the topics in the best model to the original topics. Here, Ψ is set as 0.5. It is desirable for the best topic model to have a larger ratio of total significance degree.

The computational complexity of calculating significance degree of topics and finding the best number of topics depends on the computational complexity of matrix operations. The computational complexity for determining the best number of topics is $\Theta(dn(d+n))$ at the worst case, where d is the number of documents, and n is the number of topics discovered by LDA.

In the following, we rank significance degree and use the topic selection criterion for determining the best number M. Fig. 2 illustrates the two sub-figures. Fig. 2(a) and (b) respectively represent the ranked significance degrees for Corpus 20_Newsgroups (Corpus 20_Newsgroups) and Fudan Chinese Corpus (Fudan Chinese Corpus). In each of sub-figures, the X-axis is the index of topic ranked by significance degree, and the Y-axis represents the significance degree of each topic. Using the topic selection criterion, we can easily obtain that $M = 12$ for the Fig. 2(a), and $M = 10$ for the Fig. 2(b).

6. Choosing the initial centers

We choose the initial centers of clusters by using the Jensen–Shannon metric (Fuglede & Topsoe, 2004) for calculating the distance between two vectors. For any two vectors $p = (p_1, p_2, \ldots, p_M)$ and $q = (q_1, q_2, \ldots, q_M)$, the Jensen–Shannon distance $D_{js}(p, q)$ between p and q is defined by Eq. (1).

$$D_{js}(p,q) = \frac{1}{2}\left(\sum_{k=1}^{M} p_k \ln \frac{2p_k}{p_k + q_k} + \sum_{k=1}^{M} q_k \ln \frac{2q_k}{p_k + q_k}\right) \quad (1)$$

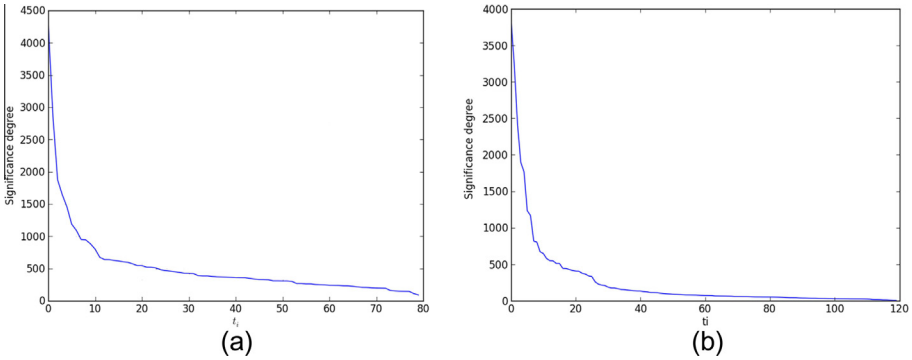

Fig. 2. The ranked topics by significance degree for the given N = 60, 80, 100, 200.

Recalling the Section 4, we will use three clustering solutions for document clustering in this paper. The difference between them is as follows. Initial centers in Solutions 1 and 2 will be chosen based on the M-dimensional vectors over $\mathcal{T}'$. Solution 3 will choose initial centers based on the N-dimensional vector space over $\mathcal{T}$ without the best topic model.

The initialization in k-means++ exploits the fact that a good clustering is relatively spread out. When a new cluster center is selected, the initialization gives preference to those further away from the previously selected centers. So the k-means++ initialization is made in an inherently sequential nature. The total computational complexity of the k-means++ initialization is $\Theta(nkd)$, which is the same as that of a single iteration in the k-means clustering, where n, k and d are the numbers of documents, clusters and dimensions, respectively. The initialization algorithm actually lowers the computational time of document clustering if M is far smaller then N.

7. K-means document clustering

The third phase is to perform the k-means document clustering by using the K initialized cluster centers. The objective function for the k-means clustering is defined by the Eq. (2).

$$\min Dis = \sum_{j=1}^{D}\sum_{i=1}^{K} D_{js}(d_j, c_i) \tag{2}$$

Definition 5. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$ be the set of original topics. For each $1 \leqslant i \leqslant N-1$, we have $Sig(t_i) \leqslant Sig(t_{i+1})$. Let $\mathcal{T}' \subseteq \mathcal{T}$ be the set of topics in the best topic model, and $\mathcal{T}' = \{t_1, t_2, \ldots, t_M\}$, where $M \leqslant N$. For the same document d_j, let $d_j^{\mathcal{T}} = (t_1^j, t_2^j, \ldots, t_N^j)$ represent a probabilistic distribution over $\mathcal{T}$, and let $d_j^{\mathcal{T}'} = (t_1^j, t_2^j, \ldots, t_M^j)$ represent a probabilistic distribution over $\mathcal{T}'$. We say $d_j^{\mathcal{T}'} = reduce(d_j^{\mathcal{T}})$.

All the three clustering solutions use the k-means algorithm for document clustering. In a special case, Solutions 1 and 2 have the same set $\mathcal{C}$ of K initial centers, where $\mathcal{C} = \{d_1, d_2, \ldots, d_K\}$. Each initial center d_i $(1 \leqslant i \leqslant K)$ is a vector representing a document. The difference between them is that Solution 1 represents each d_i as $d_i^{\mathcal{T}'}$, while Solution 2 represents d_i as $d_i^{\mathcal{T}}$. The reason why we are concerned about this case is that we want to examinate the difference between the two clustering solutions when we use the same initial centers and perform k-means clustering over different dimensional vectors. In contrast to both the solutions, Solution 3 possibly has the different initial centers in N topic space, and performs the k-means clustering over the N-dimensional vector space.

For the same set $\mathcal{C}$ of the initial centers, $\mathcal{C} = \{d_1, d_2, \ldots, d_K\}$, its sets for the k-means clustering over M-dimensional topic space and N-dimensional space are denoted as $\mathcal{C}^{\mathcal{T}'} = \{d_1^{\mathcal{T}'}, d_2^{\mathcal{T}'}, \ldots, d_K^{\mathcal{T}'}\}$ and $\mathcal{C}^{\mathcal{T}} = \{d_1^{\mathcal{T}}, d_2^{\mathcal{T}}, \ldots, d_K^{\mathcal{T}}\}$, respectively. For both $\mathcal{C}^{\mathcal{T}'}$ and $\mathcal{C}^{\mathcal{T}}$, the minimum of the objective function Dis can be found if we take both them as initial centers.

Theorem 1. *There exists a minimum of the objective function Dis when documents $d_1, d_2, \ldots$ and d_K are used as the initial centers for the k-means clustering over an N-dimensional or M-dimensional space.*

Proof. Let $\bar{c}_1, \bar{c}_2, \ldots,$ and $\bar{c}_K$ be the cluster centers. For each $1 \leqslant i \leqslant K$, we have

$$\bar{c}_i = \begin{cases} d_i^{\mathcal{T}}, & \text{if Clustering is made over the N topic space;} \\ d_i^{\mathcal{T}'}, & \text{if Clustering is made over the M topic space.} \end{cases}$$

For each $1 \leqslant j \leqslant D$, we have

$$d_j = \begin{cases} d_j^{\mathcal{T}}, & \text{if Clustering is made over the N topic space;} \\ d_j^{\mathcal{T}'}, & \text{if Clustering is made over the M topic space.} \end{cases}$$

Let X be the number of dimensions, we have

$$X = \begin{cases} N, & \text{if Clustering is made over the N topic space;} \\ M, & \text{if Clustering is made over the M topic space.} \end{cases}$$

For the objective function $Dis = \sum_{j=1}^{D}\sum_{i=1}^{K} D_{js}(d_j, \bar{c}_i)$, all the $\bar{c}_i$ can be regarded as variables, and all the d_j are constants. The derivative Dis' of the objective function is as follows.

$$\begin{aligned} Dis' &= \left(\sum_{j=1}^{D}\sum_{i=1}^{K} D_{js}(d_j, \bar{c}_i)\right)' \\ &= \left(\sum_{j=1}^{D}\sum_{i=1}^{K}\frac{1}{2}\left(\sum_{k=1}^{X} t_k^j \ln\frac{2t_k^j}{t_k^j + \bar{t}_k^i} + \sum_{k=1}^{X}\bar{t}_k^i \ln\frac{2\bar{t}_k^i}{t_k^j + \bar{t}_k^i}\right)\right)' \\ &= \left(\frac{1}{2}\sum_{j=1}^{D}\sum_{i=1}^{K}\left(\sum_{k=1}^{X} t_k^j \ln\frac{2t_k^j}{t_k^j + \bar{t}_k^i} + \sum_{k=1}^{X}\bar{t}_k^i \ln\frac{2\bar{t}_k^i}{t_k^j + \bar{t}_k^i}\right)\right)' \\ &= \frac{1}{2}\sum_{j=1}^{D}\sum_{i=1}^{K}\left(\sum_{k=1}^{X} t_k^j \ln\frac{2t_k^j}{t_k^j + \bar{t}_k^i} + \sum_{k=1}^{X}\bar{t}_k^i \ln\frac{2\bar{t}_k^i}{t_k^j + \bar{t}_k^i}\right)' \\ &= \frac{1}{2}\sum_{j=1}^{D}\sum_{i=1}^{K}\sum_{k=1}^{X}\ln\frac{2\bar{t}_k^i}{t_k^j + \bar{t}_k^i} \end{aligned}$$

In the formula mentioned above, all the $\bar{t}_k^i$ are variables, and all the t_k^j are constants. For $1 \leqslant i \leqslant K$ and $0 \leqslant j \leqslant D$, we have $0 \leqslant \bar{t}_k^i \leqslant 1$ and $0 \leqslant t_k^j \leqslant 1$. Furthermore, we find the following facts.

(1) $\ln\frac{2\bar{t}_k^i}{t_k^j + \bar{t}_k^i} \leqslant \ln 1 = 0$, if $\bar{t}_k^i < t_k^j$. Therefore, $Dis' < 0$. We can conclude that the objective function Dis is a decreasing function when all $\bar{t}_k^i \leqslant t_k^j$ hold.

(2) $\ln\frac{2\bar{t}_k^i}{t_k^j + \bar{t}_k^i} \geqslant \ln 1 = 0$, if $\bar{t}_k^i > t_k^j$. Therefore, $Dis' > 0$. We also can conclude that the objective function Dis is a increasing function when all $\bar{t}_k^i \geqslant t_k^j$ hold.

According to the above facts, we can conclude that there must be $K \times N$ variables $\bar{t}_k^i$ such that $Dis' = \frac{1}{2}\sum_{j=1}^{D}\sum_{i=1}^{K}\sum_{k=1}^{N}\ln\frac{2\bar{t}_k^i}{t_k^j + \bar{t}_k^i} = 0$. The objective function Dis has the maximum value when $Dis' = 0$. In the case, the objective function Dis is minimal. So the conclusion holds. □

This theorem is useful and significant for guaranteeing that the function Dis always can find a minimum when the same initial centers are respectively used for the k-means clustering over different dimensional vector space. This makes meaningful the comparison between Solution 1 and Solution 2 when both them choose the same initial centers.

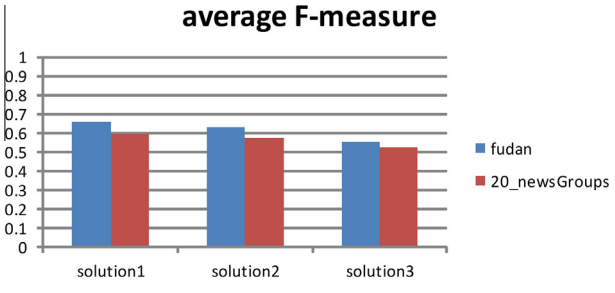

Fig. 3. Clustering effectiveness comparison among three solutions.

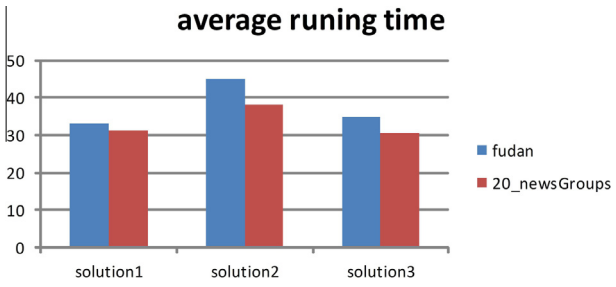

Fig. 4. Clustering efficiency comparison among three solutions.

8. System evaluation, experiments and analysis

8.1. System evaluation

We designed and implemented our three-phase k-means++ clustering algorithm by using both Python and Java. The whole system based on the three phase method proposed in this paper can process Chinese and English words. We use the ICTCLAS package (ICTCLAS) for word split of the Chinese texts. After a series of text preprocessing, we obtain the matrix of text word vectors. The LDA based approach and Gibbs sampling are used to obtain the Document-Topic matrix and the Topic-Word matrix. Then, the system can be evaluated to further determine the best number of topics, and perform the k-means++ clustering.

8.2. Experiments and analysis

In the following, we discuss the design of experiments. We specify the data sets and indexes in Section 8.2.1, and give the experimental settings in Section 8.2.2. The experimental analysis is made from the two perspectives as follows. First, we will respectively compare the clustering accuracy and the running time of the three clustering solutions in order to validate the effectiveness and efficiency of our clustering approach based on the best topic model. Second, we will compare and examinate the clustering accuracy and the running time when the initial centers in solutions 1 and 2 are the same.

8.2.1. Data sets and indexes

We use two data sets for validating our approach: Corpus 20_Newsgroups (Corpus 20_Newsgroups) and Fudan Chinese Corpus (Fudan Chinese Corpus). Corpus 20_Newsgroups (Corpus 20_Newsgroups) includes approximately 20,000 newsgroup documents nearly across 20 different newsgroups. Fudan Chinese corpus includes 19,637 documents and 20 categories such as Agriculture, Communication, Sports, Politics, and so on.

We use the following indexes for our experiment evaluation: Precision, Recall and F-Measure, which are respectively defined as follows.

$$Recall(r,s) = \frac{n(r,s)}{ns} \tag{3}$$

$$Precision(r,s) = \frac{n(r,s)}{nr} \tag{4}$$

Where r represents a category in the clustering results, s is the true category in the corpus. $n(r,s)$ is the number of documents contained in both r and s. nr and ns are the numbers of documents in r and s, respectively.

The F value between categories r and s can be further defined as follows.

$$F(r,s) = \frac{2 * Recall(r,s) * Precision(r,s)}{Recall(r,s) + Precision(r,s)} \tag{5}$$

The total index for clustering evaluation is defined as follows.

$$F = \frac{1}{n}\sum_{i=1}^{K} n_i \max\{F(i,j) | 1 \leqslant j \leqslant K\} \tag{6}$$

Where j represents a true category in the corpus, n is the total number of all the documents in the corpus, and n_i is the number of documents in a clustered category i.

8.2.2. Experimental settings

We respectively perform the three solutions mentioned in this paper, where Solution 1, i.e.. the k-means++ clustering based on the best topic model, is the one we proposed in this paper for the first time. The three solutions are respectively achieved by running our clustering system. The system is run on a laptop with Memory 4 GB and Frequency 2.1 GHz. For a given experiment, the system will be evaluated 10 times. We calculate and output the average value of the experimental results as the final result of the experiment.

We evaluate the clustering effectiveness by obtaining their F-values of clustering. F-measure is a comprehensive index for evaluating clustering accuracy. The larger the F-value is, the more effective the document clustering is. Clustering efficiency is examinated by the running time that it takes to respectively perform each of the three solutions.

For the corpus 20_Newsgroups, we know that the best number $M = 12$ when we specify $N = 80$ in the Section 5, so Solutions 1 performs the k-means++ clustering over the 12-dimensional vectors, while the other two solutions are achieved over the 80 topic space. For the corpus Fudan Chinese corpus, the best number $M = 10$ when we specify $N = 120$, so Solutions 1 performs the k-means++ clustering over the 10-dimensional vectors, while the other two solutions are achieved over the 120 topic space.

8.2.3. Comparing clustering effectiveness

This section is to compare the clustering effectiveness among the three solutions. Fig. 3 illustrates the clustering results about their F-measure for the two data sets.

First, we find that Solution 1 has the largest average F-scores for both the two corpus. This means that our clustering approach based on the best topic model (i.e. Solution 1) outperforms the other two solutions in clustering accuracy. It shows that document clustering based on the best topic model is more effective than clustering based on the original topic model. Compared to Solution 3, the F-measure values by the solutions 1 and 2 are higher than that by Solution 3. The initial centers of both the solutions are chosen based on the best topic model by k-means++. This means that the initial centers chosen based on the best topic model can improve the effectiveness of document clustering, and have significant influence on clustering accuracy.

What is interesting is the difference of the average F-measure values between Solution 1 and 2. They have the different average F-measure values even if they are seeded with the same initial centers. Because the clustering based them is made over the different dimensional vector space, this means that some insignificant topics participated in clustering interfere with clustering accuracy in Solution 2.

By the experiments, we also found that the additional topics (i.e.. the topics that do not belong to the set of the topics in the best topic model) are not helpful to improve the accuracy of clustering. On the contrary, they possibly make the accuracy lower. Therefore, the pruned topic set is already sufficient for clustering.

Another finding is that all the three solutions for the Fudan Chinese corpus have the larger average F-scores than for Corpus 20_Newsgroups. One of the reasons is possibly because each of the topics discovered by LDA in the Fudan Chinese corpus are

semantically more compact, which is beyond the scope of this paper, and remains to be explored in the future work.

8.2.4. Comparing clustering efficiency

We evaluate the clustering efficiency based on the three solutions by obtaining their running time of clustering. Fig. 4 illustrates the running time of the three clustering solutions based on the two data sets.

It is obvious that Solution 2 has the largest running time for both the two corpus. In order to deeply understand the difference residing in the three solutions, we need to analyze their constituents of running time for each solution. There are two constituents of running time for solutions 1 and 2: the running time of determining the best M (denoted as T_d), and the running time of performing the k-means++ clustering (denoted as T_k). Solution 3 only has the running time of performing the k-means++ clustering.

Solutions 1 and 2 have the closer T_d, but the clustering of Solution 2 is made over the N-dimensional vectors rather than M-dimensional vectors. As shown in the previous section, because N is generally greater than M, the T_k consumed by Solution 2 is commonly larger than the T_k consumed by Solution 1. So the running time of Solution 2 is often larger than that of Solution 1. The running time of Solution 2 is larger than that of Solution 3 because they consume the closer T_k, but Solution 2 needs the extra time T_d for determining the best topic number.

What is worthy of noting is that the running time of solutions 1 and 3 starts looking perhaps a little bit closer. Although Solution 1 needs the extra time T_d for determining the best topic number, its clustering runs over the M-dimensional vectors compared to Solution 3 whose clustering is made over the N-dimensional vectors. So their total running time depends on how many dimensions are reduced, and how many topics are selected. For examples, for the `Fudan` Chinese corpus, the running time of Solution 1 is a little bit larger than that of Solution 3. In contrast, the running time of Solution 3 is a little bit larger than that of Solution 1 for Corpus 20_`Newsgroups`.

If Figs. 3 and 4 are analyzed together, then we will have some more meaningful observations. By using our document clustering approach (Solution 1) based on the best topic model, we can improve the clustering accuracy and effectiveness almost without consuming more running time (sometimes even have the lower running time than the other solutions), compared to the k-means++ clustering based on the topics discovered by LDA (Solution 3). Solution 1 completely outperforms Solution 2 in running time and clustering accuracy. So according to the analysis above, we argue that our clustering approach based on the best topic model can improve the performance of topic based document clustering.

9. Conclusion

In this paper, we propose a three-phase approach for topic based k-means++ document clustering. We propose a method to determine the best topic model and the best number of topics by the concept of significance degree of topics and the topic selection criterion. The clustering solution 1 is used based on our three phase approach for the first time. The other two solutions are used for clustering effectiveness and efficiency comparison. The empirical experiments show that our clustering approach based on the best topic model improves the performance of topic based document clustering.

The future work includes the three aspects as follows. In our current work, we separately address topic model and clustering, and do not consider the close correlation between topic model and clustering. In fact, document clustering is also helpful for improving the quality of topic modeling, and therefore both them are closely correlated and benefit each other. So we will need to deeply explore the correlation between topic model and clustering for further improving the effectiveness and efficiency of document clustering. Second, new clustering approaches should be explored to further improve the clustering efficiency. Third, we will pursue the application of our approach in a large volume of data sets such as big data sets in electric power systems.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under (61001197, 61372182), National Key Basic Research Program of China (2009CB320704), and the Fundamental Research Funds for the Central Universities.

References

Agarwal, P. K., & Mustafa, N. H. (2004). K-means projective clustering. In Proceedings of PODS2004.

Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text clustering algorithms. *Mining Text Data*, 77–128.

Agha, M. E., & Ashour, W. M. (2012). Efficient and fast initialization algorithm for k-means clustering. *International Journal of Intelligent Systems and Applications* (1), 21–31.

AlSumait, L., Barbara, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. In *Proceedings of ECML/PKDD'09* (pp. 67–82).

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms (pp. 1027–1035).

Blei, D. M. (2012). Introduction to Probabilistic Topic Models. *Communications of ACM, 55*(4), 77–84.

Blei, D., & Lafferty, J. (2006). *Correlated topic models. Advances in neural information processing systems* (Vol. 18). Cambridge, MA: MIT Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bolstad, W. M. (2010). *Understanding computational bayesian statistics*. John Wiley.

Cai, D., He, X., & Han, J. (2011). Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge and Data Engineering, 23*(6), 902–913.

Cao, J., Xia, T., Li, J., Zhang, Y. D., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7-9), 1775–1781.

Corpus 20_`Newsgroups`. <http://qwone.com/~jason/20Newsgroups/>.

Drake, J., & Hamerly, G. (2012). Accelerated k-means with adaptive distance bounds. In *Proceedings of 5th NIPS workshop on optimization for machine learning*, December.

Elkan, C. (2003). Using the triangle inequality to accelerate k-means. *Proceedings of ICML'03*, 147–153.

Emre Celebi, M., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications, Expert Systems with Applications, 40*, 200–210.

Fudan Chinese Corpus. <http://www.datatang.com/data/43318>.

Fuglede, B. & Topsoe, F. (2004). Jensen–Shannon divergence and Hilbert space embedding. In *Proceedings of international symposium on information theory* (ISIT'04).

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. Alexandria, VA: SIAM-ASA.

Gao, J., & Hitchcock, D. B. (2010). James–Stein shrinkage to improve k-means cluster analysis. *Computational Statistics & Data Analysis, 54*(9), 2113–2127.

Gruber, A., Rosen-Zvi, M., & Weiss, Y. (2007). Hidden topic markov models. *Proceedings of AISTATS'07*, 163–170.

Hamerly, G. (2010). Making k-means even faster. In *Proceedings of SIAM international conference on data mining*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of SIGIR'99*.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1), 177–196.

Hu, Y., Boyd-Graber, J., & Satinoff, B. (2011). Interactive topic modeling. *Proceedings of ACL'11*, 248–257.

ICTCLAS. <http://ictclas.org/>.

Jiang, M. F. et al. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Letters, 22*, 691–700.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 881–892.

Kumar, A., & Srinivas, S. (2010). Concept lattice reduction using fuzzy k-means clustering. *Expert Systems with Applications, 37*(3), 2696–2704.

Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. *Proceedings of SIGKDD'99*, 16–22.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of international conferenceon machine learning* (ICML'06).

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*, 129–137.

Lu, Y., Mei, Q., & Zhai, C. X. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval, 14*(2), 178–203.

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems, 2*, 849–856.

Nguyen, C. D., Nguyen, D. T., & Pham, V. H. (2013). Parallel two-phase k-means. In *Proceedings of computational science and its applications* (ICCSA 2013) (pp. 224–231).

Pena, J. M., Lozano, J. A., & Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters, 20*(10), 1027–1040.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2004). A two-phase k-means algorithm for large datasets. *Proceedings of the Institution of Mechanical Engineers, 218*(10), 1269.

Putthividhya, D., Attias, H. T., & Nagarajan, S. (2009). Independent factor topic models. In *Proceedings of the 26th international conference on machine learning* (ICML'09), Montreal, Canada.

Wang, L., Wei, B., & Yuan, J. (2011). Topic discovery based on LDA_col model and topic significance re-ranking. *Journal of Computers, 6*(8), 1639–1647.

Wei, J. T., Lee, M. C., Chen, H. K., et al. (2013). Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications, 40*(18), 7513–7518.

Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. In *Proceedings of UAI'13*.

Xu, W., & Gong, Y. (2004). Document clustering by concept factorization. *Proceedings of SIGIR'04*, 202–209.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of SIGIR'03*, 267–273.

Zhang, X. et al. (2012). A density-based method for Initializing the k-means clustering algorithm. In *Proceedings of 2012 international conference on network and computational intelligence*.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications, 41*(4), 1476–1482.