K-means Document Clustering Based on Latent Dirichlet Allocation

Peng Guan, School of Economics & Management, Nanjing University of Science & Technology, 200
Xiaolingwei Street, Nanjing, China, 210094,86+25+84318560, guanpeng1983@163.com
Yuefen Wang, School of Economics & Management, Nanjing University of Science & Technology, 200
Xiaolingwei Street, Nanjing, China, 210094,86+25+84318560, yuefen163@163.com
Bikun Chen, School of Economics & Management, Nanjing University of Science & Technology, 200
Xiaolingwei Street, Nanjing, China, 210094,86+25+84318560, rome3@163.com
Zhu Fu, School of Economics & Management, Nanjing University of Science & Technology, 200
Xiaolingwei Street, Nanjing, China, 210094,86+25+84318560, fuzhu886@163.com

ABSTRACT

K-means is a popular algorithm in document clustering, which is fast and efficient. The disadvantages of K-means are that it requires one to set the number of clusters first and select the initial clustering centers randomly. Latent Dirichlet Allocation (LDA) is a mature probabilistic topic model, which aids in document dimensionality reduction, semantic mining and information retrieval. We present a document clustering method based on LDA and K-means (LDA_K-means). In order to improve document clustering effect with K-means, we discover the initial clustering centers by finding the typical latent topics extracted by LDA. The effectiveness of LDA_K-means is evaluated on the 20 Newsgroups data sets. We show that LDA_K-means can significantly improve the clustering effect in contrast to clustering based on random initialization of K-means and LDA (LDA_KMR).

INTRODUCTION

As Big Data applications undergo rapid development, it has become easier to store and analyze mass datasets. However, people could easily be lost in huge information produced from blogs, BBS and mobile terminals every day. Document clustering is an important information processing method which can automatically organize large amount of documents into a small number of meaningful clusters and find latent structure in unlabeled document collections. Document clustering is often used in intelligence analysis field to resolve the issue of information overload[1].

K-means algorithm is one of the partitioned-based clustering algorithms[2], which has been popularly used in such areas as information retrieval[3]and personalized recommendation[4].However, in traditional K-means algorithm, initial clustering centers are selected randomly. Hence, the cluster results are too dependent on the initial clustering centers, especially when documents are represented with a bag-of-words (BOW) model. When using raw terms as features, documents are often represented as high-dimensional and sparse vectors–a few thousand dimensions and a sparsity of 95% to 99% is typical [5].In such situations, if we use K-means in document clustering, we must solve two problems: One is how to reduce the document dimensionality and capture more semantic of documents as efficient as possible; another is how to discover the initial clustering centers that can represent most semantic information of the latent clusters.

For the first problem, we will use Latent Dirichlet Allocation (LDA)[6] to model documents. In fact, document clustering and LDA are highly correlated and LDA can mutually benefit document clustering. As one of the basic topic models, LDA can discover latent semantic structure in document corpus. The latent semantic structure is able to put words with similar semantics into the same group. The topics can capture more semantic information than raw term features. So, through LDA model, each document will be represented into a topic space to reduce the noise in similarity measure. Hence the latent grouping structure of the documents can be identified more effectively.

For the second problem, we present an algorithm of discovering initial clustering centers based on LDA. Given a document corpus, we will input results from LDA to K-means. However, we cannot simply cluster documents into the topic structure, because there always exist some disturbance topics whose semantic meaning is fuzzy. Hence, we propose a method which can find typical topics and discover the initial clustering centers for K-means. This is the major contribution of this article.

RELATED WORK

K-means algorithm is based on the objective function of prototype. How to optimize initial clustering centers of K-means is very important for improving clustering effect. In general clustering literature[7-10], some improved algorithms are proposed. See[11]for a broad overview. Unlike general clustering, the document clustering must consider the processing to document dataset, such as document dimensionality reduction[12].One popular method is based on topic model, including Latent Semantic Indexing(LSI),probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet allocation (LDA), etc. The basic ideaof methods based on topic model is to transform documents from raw word space to latent topic space.

LDA is a completely generative probabilistic topic model. In generative probabilistic topic model, we treat our data as arising from a generative process that includes hidden variables. This process defines a joint probability distribution over both the observed and hidden random variables[13]. So, in LDA, the latent variables are represented as polynomial distribution over words, exiting in the whole corpus. A document could be represented as a polynomial distribution over topics.

Through topic extraction, we can reduce the document feature dimensionality[14] and measure the similarity between the documents. Since Blei and coauthors first represented the LDA in 2003, many improvement works were made. The Topic Over Time model (TOT) was represented for topic evolution with the change of time[15]. An unsupervised LDA model based on WordNet used the sense of word as latent variable, which could solve the problem of Word Sense Disambiguation (WSD)[16]. In addition, some future directions, such as evaluation and model checking, visualization and user interfaces et al. were discussed in detail by Blei (2012).

In fact, many researches[17,18]have applied LDA to document clustering. Millar et al. (2009) presented a document clustering and visualization method based on LDA and self-organizing maps (LDA-SOM) [19]. Xie and Xing (2013) integrated document clustering and LDA. They proposed a multi-grain clustering topic model (MGCTM)[20] which integrated document clustering and topic model into a unified framework and jointly performed the two tasks. However, in the process of LDA-SOM and MGCTM, how to identify disturbance topics is still a problem unsolved. Especially for MGCTM, they need to set the global and local topic numbers, which increases the complexity of the problem. In addition, all document clustering methods based on K-means above, have a same problem that the initial clustering centers remain randomly selected as traditional K-means dose. So this article present a new method to discover the initial clustering centers for K-means.

LDA_K-means

Algorithm of Discovering Initial Clustering Centers for K-means

Traditional K-means algorithm needs toselect the initial clustering center randomly. It is clear that it always lacks semantic correlation between clustering center and cluster members finally. In addition, a single document in corpus contains few words that could represent the cluster semantic. So it usually spends many iterations to find appropriate clustering centers. However, our method will improve this situation by using LDA. We assume that there must exist some latent clusters in a document set. By

LDA, we get some topics of document set. As we discussed in section 1, the typical topics can give import semantic information to the latent clusters. So, it is important to find the typical topics.

We first select the typical topic by document support number of topic, noted as a parameter C. Then, we select initial clustering centers through support documents of the typical topic. The algorithm is described as follows.

Algorithm of discovering initial clustering center for K-means

(1) Input a corpus with M documents.

(2) Generate document-topic matrix θ with T topics by LDA.

(3) For each topic:

(i) Compute the mean value of support degree of document to topic, noted as P.

(ii) Select document as a support document whose document support degree is greater than P.

(iii) Compute the document support number C and the mean value of support document set noted as $\vec{D}$·

(iv) If a topic whose document support number is greater than or equal to $C_0 = \frac{M}{T}$ exists, we select this topic as a typical topic.

(v) Repeat process (iv) until no new typical topic is generated.

(5) Compute the number of typical topics as cluster number K and choose vector $\vec{D}$ of each typical topic as initial clustering center.

(6) Output the cluster number K and initial clustering center set $\{\hat{D}_1, \hat{D}_2, ..., \hat{D}_K\}$.

Note that the role of parameter C_0 is to find out the typical topics. Usually, different value of C_0 generates different number of typical topics. We assume that a topic is a typical topic whose document support number is greater than the mean value $\frac{M}{T}$.

LDA_K-means algorithm

We have determined the initial clustering centers already. The next step is document clustering using K-means. Through the empirical comparison made by Lee (1999),Jensen-Shannon divergence performed better than Euclidean distance to measure distributional similarity[21]. So, we use Jensen-Shannon divergence instead of Euclidean distance in LDA_K-means algorithm.

LDA _K-means algorithm

Input:

- Document-topic matrix $\theta = \{\theta_1, \theta_2, ..., \theta_M\}$ and T topics generated by LDA model, where θ_i means the document-topic vector of the *i*th document.
- The initial clustering center set $\{\hat{D}_1, \hat{D}_2, ..., \hat{D}_K\}$ generated by algorithm of discovering initial clustering center for K-means.
- i=1.

Process:

a) Compute the Jensen-Shannon divergence between θ_i and clustering center $\{D'_1, D'_2, ..., D'_K\}$. Distribute θ_i to the nearest cluster according Jensen-Shannon divergence;

b) For the new clusters, recalculate the clustering centers $\{D''_1, D''_2, ..., D''_K\}$.
i=i+1;

c) Repeat the steps a) and b), until the distance between new center and the original center is zero or less than a specified threshold.

Output: the final clustering center and all documents belonging to each cluster.

Algorithm Complexity

In topic model and document clustering, we often concern about algorithm complexity. If a document cluster model suffers from potentially high computational costs, we will consider whether it is worth to use, especially for large size of document collections.

Liu and Croft (2004) used a three-pass K-means algorithm primarily motivated by its efficiency[22]. They showed that the running time for each pass/iteration grew linearly with the number of documents (N) and the number of classes (K), i.e., O(KN).Wei and Croft (2006) gave a more detailed analysis about LDA algorithm perplexity due to the large size of document collections. They showed that the complexity of each iteration of the Gibbs sampling for LDA was also linear with the number of topics (K) and the number of documents (N), which was also O(KN)[23].

The time-consuming part of algorithm of discovering initial clustering center for K-means is linear with K, C and N, where K is the topic numbers , C is document support number of topic ,N is the number of documents. So, the time-consuming of LDA_K-means for each pass is linear with the total number (N) of documents to be clustered.

Experimental Evaluation

We experimentally evaluate the performance of LDA_K-means using 20 Newsgroups dataset[24]. Nowadays, it has become the standard dataset in the field of machine learning. In the rest of this section, we will describe information of experimental dataset first. Then, we gave experimental measures. Finally, we compared the clustering result between LDA_K-means and clustering based on random initialization of K-means and LDA (LDA_KMR).

The 20 Newsgroups dataset is a supervised dataset collected from 20 different netnews newsgroups. We choose 6 newsgroups as a cluster structure by artificial judge: alt.atheism, comp.sys.ibm.pc.hardware, rec.motorcycles, rec.sport.baseball, sci.space and talk.politics.gun. For computing the evaluation measure easily, we randomly select 300 documents from each newsgroup. Table 2 gives statistics of the data sets.

Category	Labels	# of documents	# of terms
alt.atheism	1	300	4632
comp.sys.ibm.pc.hardware	2	300	3143
rec.motorcycles	3	300	3664
rec.sport.baseball	4	300	3745
sci.space	5	300	4633
talk.politics.gun	6	300	4288

Table 2. Statistics of Data Sets

Collection	**F_score**	**AA**	**Entropy**
alt.atheism	0.692	0.684	0.384
comp.sys.ibm.pc.hardware	0.791	0.81	0.341
rec.motorcycles	0.803	0.798	0.332
rec.sport.baseball	0.832	0.858	0.301
sci.space	0.793	0.805	0.372
talk.politics.gun	0.775	0.761	0.376
Mean value	0.781	0.786	0.351

Table 3. LDA_K-means Cluster Results for 20 Newsgroups Data Sets

We use supervised method to evaluate the cluster results with three popular measures that are *Average Accuracy (AA), F-Score* and *Entropy*[25].

Experimental Results and Comparison

First, we obtain experimental corpus using natural language processing tools NLTK[1]. The processing

[1] NLTK 3.0: http://www.nltk.org/

contains removing the stop words and extracting the stem. Then, LDA was performed using the open source machine learning tool Gensim[2].

Parameter setting for LDA model: the topics are extracted from a single sample at the 2000th iteration of the Gibbs sampler; the hyper-parameters are $\alpha = 0.01, \beta = 0.05$; topic number is 25 which is determined by Bayesian method (Griffiths and Steyvers, 2004).

Another important parameter for LDA_K-means is C_0. According to 1800 documents and 25 topics, $C_0 = \frac{1800}{25} = 72$. Table 3 shows the LDA_K-means cluster results for 20 Newsgroups data sets. It is shown that rec.sport.baseball gets the best cluster result and alt.atheism gets the worst. From the result of LDA, we find that the data set rec.sport.baseball has more clear topic terms. So, it is easier to find out the typical topic for cluster. That's why we get better cluster result from data set rec.sport.baseball.

For evaluating the effectiveness of parameter C_0, we give different values to C_0. Figure 3 shows LDA_K-means cluster results for 20 Newsgroups data sets under different values of C_0. We just describe the mean value for evaluation measures. Form figure 3 we can see that we get almost the same results when C_0 equals to 70 or 80. So, we affirm that it must be an interval for C_0, in which we can get best cluster results. In fact, the topic number for LDA changes when running LDA each time. So, we can give parameter C_0 a variation ranges. Cluster results are considered to be effective in this variation ranges.

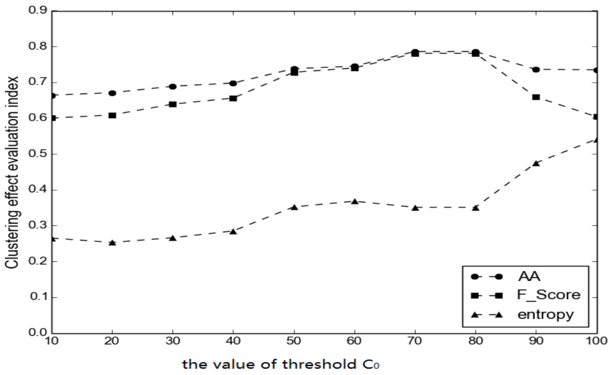

Figure 3.LDA_K-means Cluster Results Under Different Values of C_0

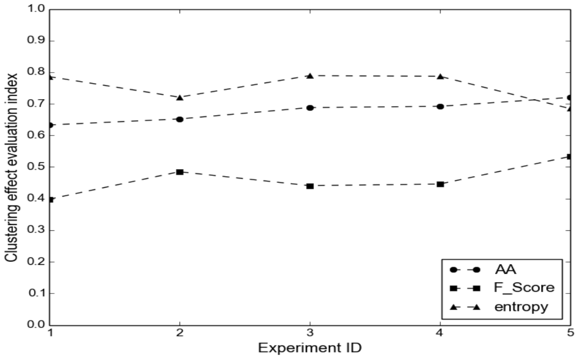

Figure 4.LDA_KMR Cluster Results for 20 Newsgroups Data Sets in Five Experiments

In this section, we compare LDA_K-means and LDA_KMR. In fact, some researches show that in K-means document clustering, LDA performs better than LSI and pLSI[20]. But they don't consider the initial clustering center when using K-means. So, in this paper, we only compare the document clustering algorithm based on LDA and K-means. The difference is whether initial clustering center is randomly selected before K-means.

Figure 4 shows the results of LDA_KMR clustering repeated five times. Since the LDA_KMR clustering selects the initial clustering center randomly, so we show the average value of the cluster results.

Measure	LDA_K-means	LDA_KMR	t	df	Sig. (2-tailed)	Mean Difference
F_score	0.781	0.454	-14.537	4	0.000	-0.327
AA	0.786	0.672	-6.649	4	0.003	-0.114
Entropy	0.351	0.76	16.042	4	0.000	0.409

Table 3. Comparison Between LDA_K-means and LDA_KMR. The Evaluation Measure are F-Score, Average Accuracy and

[2] gensim 0.10.2:Python Package Index. https://pypi.python.org/pypi/gensim

Entropy. Sig.(2-tailed) is the p-value with a 95% Confidence According to the t-test.

From table 3, LDA_K-means is significantly more effective than LDA-KMR in respect of measure Entropy. LDA_KMR has high Entropy and great fluctuation in experiments shown as figure 4. Entropy of LDA_KMR is greater than 0.6 in five random experiments. But, we can also get Average Accuracy to 0.72 occasionally shown in figure 4. That means, although we can get good cluster result from LDA-KMR, the structure of cluster is chaos. Besides, LDA_K-means performs significantly better than LDA-KMR on measure F-score. Based on the empirical results reported in table 3, figure 3 and figure 4, we employt-test for each evaluation measure. From the result in table 4, we can see LDA_K-means performs significantly better than LDA-KMR on each measure with a 95% confidence according to the t-test.

Conclusion and Future Work

Based on the experimental results, we can make the following conclusions. First, experiments performed on the 20 Newsgroups data sets have demonstrated that LDA_K-means works in a stable manner for different newsgroups data sets. It will fluctuate, however, with change of topic numbers. Secondly, we have shown that LDA_K-means performs better than LDA_KMR on 20 Newsgroups data sets. More importantly, compared with LDA_KMR, LDA_K-means achieves smaller value of the evaluation measure Entropy, which means that each cluster from LDA_K-means is derived from fewer topics than LDA_KMR. This demonstrates that algorithm of discovering initial clustering center for K-means can make the latent clustering center more focus on a certain topic. Hence it can improve the cluster effect for K-means.

We have applied this method to six document collections in 20 Newsgroups data sets with good results. However, there are some challenges in applying this method. Most significantly, reasonable values for the number of topics in LDA must be discovered, for which we chose Bayesian method. However, different number of topics still resulted when using Bayesian method each time. Secondly, while our outcomes were obtained in our processing of LDA_K-means using 20 Newsgroups data sets, different types of text, such as scientific literature, Weibo (microblog) short-form text, or other forms of text data should be experimented to further and more thoroughly evaluate LDA_K-means. We intend to address this issue in future work.

ACKNOWLEDGMENTS

This research was supported in part by National Natural Science Foundation of China grant 71373124.

REFERENCES

[1]Eppler, M. J., &Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The information society, 20*(5), 325-344.

[2]MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297.

[3]Manning, C. D., Raghavan, P., &Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, p. 496). Cambridge: Cambridge university press.

[4]Kim, K. J., &Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. *Expert systems with application*s,34(2), 1200-1209.

[5]Dhillon, I. S., &Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, *42*(1-2), 143-175.

[6]Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

[7]Arthur, D., &Vassilvitskii, S. (2007, January). K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.

[8]Khan, F. (2012). An initial seed selection algorithm for K-means clustering of georeferenced data to improve replicability of cluster assignments for mapping application. *Applied Soft Computing*, *12*(11), 3698-3700.

[9]Xiao, J., Yan, Y., Zhang, J., & Tang, Y. (2010). A quantum-inspired genetic algorithm for K-means clustering. *Expert Systems with Applications*, *37*(7), 4966-4973.

[10]Reddy, D., Jana, P. K., & Member, I. S. (2012). Initialization for K-means clustering using Voronoi diagram. *Procedia Technology*, *4*, 395-400.

[11]Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the K-means clustering algorithm. *Expert Systems with Applications*, 40(1), 200-210.

[12]Jun, S., Park, S. S., & Jang, D. S. (2014). Document clustering method using dimensionality reduction and support vector clustering to overcome sparseness. *Expert Systems with Applications*, *41*(7), 3204-3212.

[13]Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM,55(4)*, 77-84.

[14]Griffiths, T. L., &Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228-5235.

[15]Wang, X., & McCallum, A. (2006, August). Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*(pp. 424-433). ACM.

[16]Boyd-Graber, J. L., Blei, D. M., & Zhu, X. (2007, June). A Topic Model for Word Sense Disambiguation. In *EMNLP-CoNLL* (pp. 1024-1033).

[17]LIU, Z., WANG, D., FENG, S., ZHANG, Y., & FANG, D. (2011). An Approach of Latent Semantic Space Partition and Web Document Clustering.*Journal of Chinese Information Processing*, *1*, 60-65.

[18]Wang, L. D., Wei, B. G., & Yuan, J. (2012). Document clustering based on probabilistic topic model. *DianziXuebao(Acta Electronica Sinica)*, *40*(11), 2346-2350.

[19]Millar, J. R., Peterson, G. L.,& Mendenhall, M. J. (2009, March). Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps. In *FLAIRS Conference* (Vol. 21, pp. 69-74).

[20]Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. *arXiv preprint arXiv:1309.6874.*

[21]Lee, L. (1999, June). Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 25-32). Association for Computational Linguistics.

[22]Liu, X., & Croft, W. B. (2004, July). Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 186-193). ACM.

[23]Wei, X., & Croft, W. B. (2006, August). LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 178-185). ACM.

[24]Ken Lang.(1995). 20 Newshroup DataSet[Websites]. Retrieved from http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html

[25]Zhou, Z. (2005). Quality Evaluation of Text Clustering Results and Investigation on Text Representation. ME thesis. Institute of Computing Technology Chinese Academy of Sciences.