The Auto Annotation Latent Dirichlet Allocation

Yingzhuo Xiang [a,*], Dongmei Yang [b] and Jikun Yan [c]

National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu, China

[a]xiangyzh@vip.qq.com, [b] 410887907@qq.com, [c]yanjk@126.com

Keywords: LDA, auto annotation, NLP, text modeling.

Abstract. In this paper, we introduce the Auto-Annotation LDA models (aaLDA), a statistical model of non-labeled documents. This model generates the annotation of LDA automatically. We derive the annotation of LDA using a k-means methods combined with a pre-processing of the corpus. In this paper, we use aaLDA models to categorize "zhongwenshilei" corpus, which is a famous Chinese corpus. Then we make a compare with the traditional LDA methods.

1. Introduction

With the Big Data Era coming, most information is stored as text format in the disk, such as emails, web pages and so on. There is a growing need to analyze large collections of electronic text. The document corpus is complexity, which leads to considering applying hierarchical statistical models based on topics. A topic usually represents an underlying semantic theme. And a document with a large number of words is considered as a combination of some topics. These topic models are powerful to describe a document collection, which facilitates tasks like searching, clustering, and browsing.

2. Related works

A famous topic model, latent Dirichlet allocation (LDA) [1], has been studied a lot by many scholars. The goal of LDA is to infer topics that maximize the likelihood (or the posterior probability) of the collection. As the model is complex, [1] used a variational inference method to solve the model. Instead of variational inference method, [2] proposed to use Gibbs Sampling method to estimate the posterior probability. As Gibbs Sampling is easy for programing and the results is similar to variational inference method, it has been widely used to solve such problems. To make the results better, [3] used annotation, which is known to be useful if the annotation is proper. In Gibbs Sampling, the annotation is a pre-assignment of some specific words to a specific topic. However, which words should be pre-assigned is an important issue. If we have read the corpus, we can annotate the words that can mostly describe a specific topic. But it needs lots of time and Manual labor. This paper proposed the Auto Annotation LDA model (aaLDA model) aiming to solve this problem. The aaLDA can annotate some words to specific topics automatically using a k-means cluster [4]. In the next part of this paper, we will introduce our aaLDA model, and we compare aaLDA model to LDA model in "zhongwenshilei" corpus. We find that aaLDA model can have 6% increase in NMI (Normalized Mutual Information) [5], [6] compared with LDA model.

3. Auto Annotation LDA models

Before introducing aaLDA model, we first introduce some definitions and notations. We use N to present the total number of unique words in the corpus D. Our goal is to cluster the corpus D into k clusters. Under the LDA model, each document and response arises from the following generative process:

Draw topic proportions $\theta|\alpha \sim Dir(\alpha)$;

For each word

Draw topic assignment $z_n \mid \theta \sim Mult(\theta)$;

Draw word $w_n \mid z_n, \beta_{1:K} \sim Mult(\beta_{zn})$

The graphical model representation of LDA is shown in figure 1. The LDA model has three levels. The parameters α and β are corpus- level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

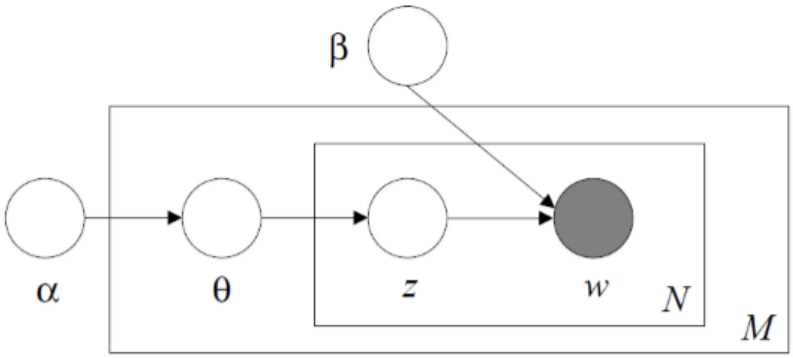

Fig. 1 graphical model of LDA

In order to solve this model, [2] propose a Monte Carlo method, called Gibbs Sampling. This method using Eq. 1 to assign words to topics, but with counts that are computed from the subset of the words seen so far rather than the full data. The chain is then run for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Eq. 1. Because the only information needed to apply Eq. 1 is the number of times a word is assigned to a topic and the number of times a topic occurs in a document, we can pre-assign the time of specified words to a specific topic in order to improve the performance of the clustering results. Our method uses Gibbs Sampling to solve LDA, so we can pre-assign the words that we got in the pre- processing step to specific topics.

$$P(z_i = j|\boldsymbol{z}_{-i}, \boldsymbol{w}) \propto \frac{n_{-i,j}^{(wi)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta} \frac{n_{-i,j}^{(di)}+\alpha}{n_{-i,\cdot}^{(di)}+T\alpha} \quad (1)$$

Our aaLDA model is composed of three part, Fig.2. The first part is a pre-processing of corpus. In this step, we use a LDA to cluster corpus into a large number of topics. Here we clustering each document into 50 topics, which means we use 50 topics to present a document. After the first part, we got a matrix of T*N dimensions, which T represents the number of topics we used to present the document, and N means the number of the vocabulary size. In the second part, the matrix is processed with a k-means cluster, which will cluster the T topics we got in the previous step into k classes. Each class is a vector of N dimensions. Then we can get each cluster's center point, which is represented as an N dimension vector. We use the top m words that have large weigh which means a large value in each dimension. The word we picked will be pre-assigned into the topic. In the third part, we use the annotated words and Gibbs Sampling to re-process the corpus. In our experiment, we set m as 10 and the annotated initial value is 1000.

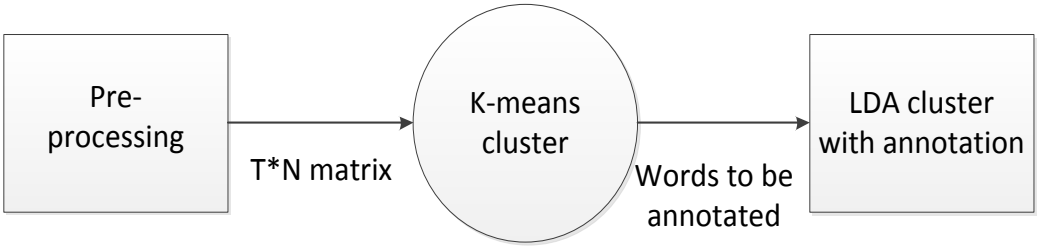

Fig. 2 auto annotation LDA model

4. Experiment

In order to test the aaLDA model, we used “zhongwenshilei” corpus, which is a famous Chinese news corpus that has been categorized into ten classes. Each class is represent as politics, environment, education, military, economic, culture, history, entertainment, and agriculture and computer science. We use JAVA to implement our algorithm. As described above, we use Gibbs Sampling to solve the LDA model. In aaLDA model, the first LDA pre-processing used 500 times burn-in, and 500 times sampling to make the Gibbs Sampling convergence. In the k-means part, we use cosine distance to measure the distance of each vector. In the third part, we use 500 times burn-in and 500 times sampling, the same as the settings in part one. To make the compare fair, we set the normal LDA 1000 times burn- in and 1000 times sampling, which is the same as the sum of burn-in times and sampling times in aaLDA. In our experiment, we measure the average NMI of the two models for 50 times each. The average NMI of aaLDA model is 0.73 while the average NMI of normal LDA is 0.67. See table 1.

Table 1 the experiment results

	LDA	aaLDA
Average $NMI_{50\ times}$	0.67	0.73

The reason why aaLDA can work better than normal LDA is due to the proper annotation of the specific words to specific topics. Because k-means cluster can figure out the center point of each cluster, we use this feature to find out which proper words should be annotated. As the center point can be best to describe each cluster, we consider the words we pick to annotate is to be the best to describe the topic. Another reason may be that this method can make up the short comings of the LDA model, which may sampling a word having high tf-idf more times, especially if the word cannot present the topic properly. In this case, LDA may not performs well. Our method can well deal with this case, by using a k-means cluster to pre-assign the word that can describe the topics proper a higher initial value. We also have done some experiment to find out how many times of burn-in and sampling will performs better. We have tested some specific number of burn-in times and sampling times. We found 500 times burn-in and 500-times sampling is enough for our experiment. That is the reason we choose these numbers.

5. Conclusion

In this paper, we proposed auto annotation LDA model, which can clustering corpus better than a normal LDA. We use a k-means cluster to find out the words that will be annotated in the LDA cluster. Some reason why our approach can work better has been listed. We will take some further research in which kind of words to be annotated can make annotation LDA perform better. Some other corpus will be tested, such as a corpus written in the languages of English or German. Different kind of corpus may lead to different results. We hope to find a better way to find words to be annotated, and we hope to increase the NMI up to 0.80.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. JMLR, 3:993–1022, 2003.

[2] Griffiths T L, Steyvers M. Finding scientific topics [J]. Proceedings of the National academy of Sciences of the United States of America, 2004, 101(Suppl 1): 5228-5235.

[3] Blei, D. M., & Jordan, M. I. (2003, July). Modeling annotated data. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 127-134). ACM.

[4] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm [J]. Applied statistics, 1979: 100-108.

[5] Lancichinetti, A., Fortunato, S., & Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 11(3), 033015. doi:10. 1088/1367-2630/11/3/033015

[6] http://en.wikipedia.org/wiki/Mutual_information