

# Classification of Academic Topics Based on LDA

516030910536 Feixiang Xu  
516030910520 Hengkui Cao  
Shanghai Jiaotong University

**Abstract:** We studied and applied supervised latent Dirichlet allocation (sLDA), a statistical topic model of labelled documents. We use sLDA on our real-world problem: classification of Chinese academic topics. We adopt approximate maximum-likelihood procedure for parameter estimation, and then we use the fitted model to predict response values for new academic topics. When applying to the Chinese classification problem, we have solved some derivative problems. We quantify the Chinese documents by converting them to vectors based on continuous bag of words. Besides, due to the limitation of direct segmentation of academic text, we tried to promote the result of the segmentation by adding fixed match of words. We discovered new academic terminology by text mining based on Hidden Markov Model (HMM).

■

## 1. INTRODUCTION

LDA allows us to discover connected topics and trends within topic. Unsupervised LDA has previously been used to construct features for classification. The hope was that LDA topics would turn out to be useful for categorization, since they act to reduce data dimension. However, when the goal is prediction, fitting unsupervised topics may not be a good choice. Consider predicting a movie rating from the words in its review. Intuitively, good predictive topics will differentiate words like excellent, terrible, and average, without regard to genre. But topics estimated from an unsupervised model may correspond to genres, if that is the dominant structure in the corpus. The distinction between unsupervised and supervised topic models is mirrored in existing dimension-reduction techniques.

In our work, there are a large number of Chinese academic topics. Each of these academic topics belongs to a specific field. Figure 1 shows some examples of the relationships between academic topics and their corresponding fields.

Our goal is to train models to predict which field the new academic topics belong to, based on the existing relationships between a large number of academic topics and fields, so that automatic classification can be achieved. In order to extract the potential information of the topics, we must first carry out Chinese word segmentation. The problem is that the word segmentation system does not recognize academic proper nouns, so that a satisfactory word segmentation effect cannot be obtained. In order to solve this problem, we use Hidden Markov Chain (HMM) to first automatically discover new academic words that are not in the word segmentation system library. After segmentation, we have done some common treatments in the NLP field, such as removing stop words, converting Chinese words into word vectors, and building a corpus. Then get a model by training based on supervised LDA. Finally, we tested the model with some data that was not involved in the training and got the accuracy.

## 2. MOTIVATION AND BACKGROUND

Researchers need tools to explore and browse large collections of scholarly literature. When faced with access to millions of articles in their fields, they must not be satisfied with simple search. They want interacting with the scholarly literature in a more structured way: finding articles similar to those of interest, and exploring the collection based on the underlying topics. Topic models are probabilistic models for uncovering the semantic of a document collection based on a hierarchical Bayesian analysis of the original texts.

Formally, a topic is a probability distribution over terms in a vocabulary. Informally, a topic represents an underlying semantic theme; a document consisting of a large number of words might be concisely modelled as deriving from a smaller number of topics. Such topic models provide useful descriptive statistics for a collection, which facilitates tasks like browsing, searching, and assessing document similarity. Most topic models, such as latent Dirichlet allocation (LDA), are unsupervised. Only the words in the documents are modelled, and the goal is to infer topics that maximize the likelihood (or the posterior probability) of the collection. This is compelling with only the documents as input, one can find patterns of words that reflect the broad themes that run through them and unsupervised topic modeling has many applications.

## 3. PROBLEM FORMULATION

### 3.1 Overview

The problem can be divided into three major parts: the establishment of the terminology library, Chinese text processing, and model training based on sLDA.

### 3.2 Detail

The existing dataset contains a large number of academic topics and their one-to-one correspondence fields, we need to train a model to predict the field to which an academic topic belongs, so as to achieve automatic classification of the topics. In our specific project, the magnitude of data is on the order of 100,000, and the number of fields is six, that is, the academic topics should be divided into six categories.

## 4. PROPOSED METHODS

### 4.1 supervised Latent Dirichlet Distribution

LDA is a topic model, which can give the theme of each document in the document set as a probability distribution. By analyzing some documents and extracting their themes (distribution), you can make themes according to the theme (distribution). Cluster or text classification. At the same time, it is a typical word bag model, that is, a document is composed of a group of words, and there is no order relationship between words and words. In addition, a docu-

工程与材料科学部	新型亚稳材料的设计原理、实验合成与结构调控
化学科学部	化学键活化与可控性重组研究
医学科学部	天然药源分子及其新作用特点
工程与材料科学部	城市水质转化规律与保障技术
信息科学部	通信网的网络理论和技术
信息科学部	可信软件的基础理论、方法和技术研究
信息科学部	片上系统的互连问题与高端IP核研究
地球科学部	冻土与寒区工程
生命科学部	精神药物成瘾和记忆机制
生命科学部	高等植物生殖生物学研究
工程与材料科学部	表界面纳米工程学
工程与材料科学部	复杂装备的数字化设计
管理科学部	分布系统的协调优化与风险管理
信息科学部	半导体低维结构中的量子调控
生命科学部	细胞命运决定的分子网络
化学科学部	化学工程中复杂系统的结构
化学科学部	生命科学中的分析新原理与新方法研究
数理科学部	星系的形成和星系的活动
数理科学部	北京谱仪上的新强子态和新物理现象研究
工程与材料科学部	高分子材料多层次结构及结构调控
管理科学部	基于行为的若干社会经济复杂系统建模与管理
化学科学部	功能材料的结构化学
生命科学部	北方草地全球变化生态学
数理科学部	高超声速飞行器复合材料的热力耦合问题
生命科学部	水稻粒重比较重要农艺性状的功能基因组学研究
工程与材料科学部	能源高效节约和可再生转化利用的多相流理论基础
信息科学部	控制科学中若干关键基础问题的研究
数理科学部	飞秒光物理与介观光学研究
1.png 信息科学部	数学机械化方法及其在信息技术中的应用

Fig. 1. some examples of the relationships between academic topics and their corresponding fields

ment can contain multiple topics, and each word in the document is generated by one of the topics.

LDA is about doing this: according to a given document, the theme distribution is reversed. In the LDA model, a document is generated in the following way:

- (1) Sampling from Dirichlet Distribution  $\alpha$  to generate the subject distribution  $\theta_i$  of the document.
- (2) Sampling a document from the polynomial distribution  $\theta_i$  of the topic  $i$ . The subject of the  $j$ th word  $z_{i,j}$ .
- (3) Sampling from the Dirichlet distribution  $\beta$  to generate the word distribution  $\phi_{z_{i,j}}$  corresponding to the subject  $z_{i,j}$ .
- (4) Sampling from the polynomial distribution  $\phi_{z_{i,j}}$  of the word and finally generating the word  $z_{i,j}$ .

Among them, the Beta-like distribution is the conjugate prior probability distribution of the binomial distribution, and the Dirichlet distribution (Dirichlet distribution) is the conjugate prior probability distribution of the polynomial distribution.

In supervised latent Dirichlet allocation (sLDA), we add to LDA a response variable connected to each document. As mentioned, examples of this variable include the number of stars given to a movie, the number of times an on-line article was downloaded, or the category of a document. We jointly model the documents and the responses, in order to find latent topics that will best predict the response variables for future unlabeled documents. sLDA uses the same probabilistic machinery as a generalized linear model to accommodate various types of response: unconstrained real values, real values constrained to be positive (e.g., failure times), ordered or unordered class labels, non-negative integers (e.g., count data), and other types.

Fix for a moment the model parameters:  $K$  topics  $\beta$  1:K (each  $\beta_k$  a vector of term probabilities), a Dirichlet parameter  $\alpha$ , and response parameters  $\eta$  and  $\delta$ . (These response parameters are described in detail below.) Under the sLDA model, each document and response arises from the following generative process:

- (1) Draw topic proportions  $\theta | \alpha \sim Dir(\alpha)$
- (2) For each word, draw topic assignment  $z_n | \theta \sim Mult(\theta)$ , then Draw word  $w_n | z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$
- (3) Draw response variable  $y | z_{1:N}, \eta, \delta \sim GLM(\bar{z}, \eta, \delta)$ ,  $\bar{z} := (1/N) \sum_{n=1}^N z_n$

The distribution of the response is a generalized linear model

$$p(y | z_{1:N}, \eta, \delta) = h(y, \delta) \exp \left\{ \frac{(\eta^T \bar{z}) y - A(\eta^T \bar{z})}{\delta} \right\}$$

## 4.2 HMM for New Words Discovery

In the Chinese word segmentation of the HMM model, our input is a sentence (that is, a sequence of observations), and the output is the state value of each word in the sentence.

HMM is a triple (pi, A, B): Initialize probability vector:  $\Pi = (\pi_i)$  State transition matrix:  $A = (a_{ij})$   $Pr(x_i | x_{j-1})$  The transition probability is a very important knowledge point of the Markov chain. People who have learned probability theory in the university know that the biggest feature of the Markov chain is the state of the current  $T=i$  state Status(i), only with  $T=i$ . Related to the  $n$  states before the moment. That is: Status(i-1), Status(i-2), Status(i-3),...Status(i-n)

Furthermore, the HMM model has three basic assumptions as a premise of the model, including a finite historical hypothesis, that is,  $n=1$  of the Markov chain. That is, Status(i) is only related to Status(i-1), and this assumption can greatly simplify the problem. Looking back at the transfer matrix, it is actually a two-dimensional matrix of  $4 \times 4$  (4 is the size of the state value set). In the confusion matrix, each element is actually a conditional probability. According to the three observations of the three basic assumptions of the HMM model (see the comments on the independence of observations), the observations only depend on the current state value. Is:  $P(\text{Observed}[i], \text{Status}[j]) = P(\text{Status}[j]) * P(\text{Observed}[i] | \text{Status}[j])$ . The value of  $P(\text{Observed}[i] | \text{Status}[j])$  is obtained from the confusion matrix.

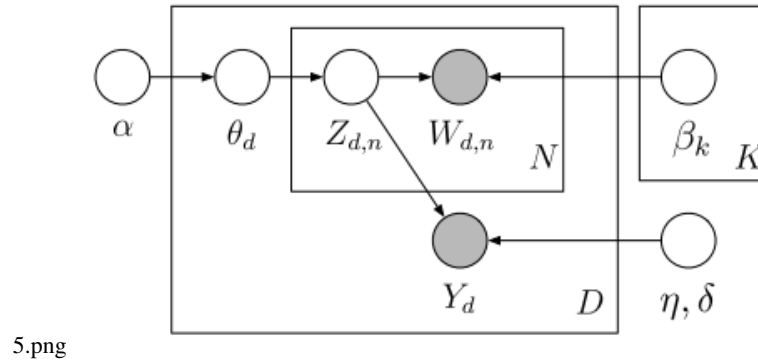


Fig. 2. Probabilistic model diagram model of sLDA

In the new word recognition, the ternary array is linked by a viterbi algorithm, the triplet is the input, and the output is the state sequence value. With the ternary array of HMM, it is equivalent to having a certain hidden Markov model. In the new word recognition, the new word can be identified by using the viterbi algorithm.

## 5. EXPERIMENT

### 5.1 Segmentation

At first we used the jieba library to segment the academic topics. Jieba is a mature method commonly used in Chinese word segmentation in the field of NLP. Soon we found that there was a problem with the result of the jieba segmentation. There is a lack of terminology in jieba's library, which is an important component of the academic topics. Thus the result contains many inappropriate segmentation, which has an adverse effect on the results.

Thus we started to explore effective improvements. We found that jieba reserved a library for us. By artificially adding vocabulary to it, jieba will remain the integrity of the words in it when segmenting.

However, manually adding words can be an extremely hard work, due to the large amount of the words. So we used a clever and efficient method, that is using Hidden Markov Chain (HMM) to conduct the segmentation. According to the principle of HMM segmentation we just explained, we can get a segmentation with new words been automatically discovered.

Figure shows some examples of the segmentation. From the examples we can see that much of the terminologies have been separated out and completely retained. This indicates that the applying of HMM in segmentation does work, comparing to the directly segmentation without HMM. Professor Jiang suggested that the IEEE Association has a list of a large number of academic vocabulary, which contains enough academic vocabulary in various fields to be used for our word segmentation. However, what we are doing is the classification of Chinese academic topics. The vast majority of the existing information is in English. It is difficult for us to find references to Chinese. Thus, we choose to find our own way to solve this problem.

### 5.2 Text Processing

This part of work contains some common operation on Chinese text processing in NLP field. The purpose of these processes is to gradually transform the Chinese text into a mathematical expression that can be quantified. After segmentation and the removing of stop-

- 1 总论 总论分析 总论分析
- 2 心血 心血管 血管 系统 内科 科学
- 3 数理 数理统计 统计
- 4 摩擦 摩擦学
- 5 机械 机械学 学科
- 6 固体 体力 力学
- 7 物理 物理学 理学
- 8 大地 大地测量 测量 测量学 物理 大地 大地测量 测量 动力 大地 大地测量 测量
- 9 环境 化学
- 10 恒星 物理 与 星际 物质
- 11 超导 材料
- 12 传感 传感器 传感器 技术 与 测试 测试仪 仪器
- 13 固体 无机 无机化学 机化 化学
- 14 信息 理论与 信息 信息系统 系统
- 15 控制 控制论
- 16 宇宙 宇宙学
- 17 地球 地球物理 物理 物理学 理学 与 空间 物理
- 18 凝聚 凝聚态 物性 电子 子结构 结构 电学 磁学 和 光学 性质
- 19 肿瘤 肿瘤学
- 20 化学 化学工程 学工 工程 及 工业 工业化 工业化学 化学
- 21 工程 水力 水力学 力学
- 22 基础 数学
- 23 高分 高分子 分子 合成
- 24 分子 遗传 遗传学
- 25 固体 体力 力学
- 26 高分 高分子 分子 科学
- 27 消化 消化系 消化系统 系统 内科 科学
- 28 工程 热物 物理 与 能源 利用 用学 学科
- 29 气动 气动力 动力 动力学 力学
- 30 核能 物理 物理学 理学 和 量子 子理 理论
- 31 金属 金属材料 材料 学科
- 32 热力学 力学 与 统计 物理 物理学 理学 含混 混沌
- 33 金属 金属腐蚀 腐蚀 与 防护
- 34 核物理 物理
- 35 细胞 生物 生物学 及 发育 生物 生物学
- 36 物理 物理学 理学
- 37 遗传 遗传学
- 38 能源 化工
- 39 基因 基因工程 工程
- 40 金属 金属材料 材料 学科

Fig. 3. some examples of the segmentation results

words, we got a satisfying segmentation result, just like the figure shows. This provides us with easy-to-use materials for our next text processing. The library gensim provides us some convenient approaches for text-vector converting. We then establish a corpus for the text, which consists of the vectors that the text has been converted to. The process of converting Chinese text to vectors is based on the calculation of TF-IDF values. TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and data mining. TF means Term Frequency, and IDF means Inverse Document Frequency. TF-IDF is a statistical method used to assess the importance of a word for a file set or one of the files in a corpus. The importance of a word increases proportionally with the number of times it appears in the file, but it also decreases inversely with the frequency it appears in the corpus. The main idea is that if a word or phrase appears in an article with a high frequency and rarely appears in other articles, the word or phrase is considered to have a good class distinguishing ability and is suitable for classification.

The figure shows part of the corpus. It can be seen that the corpus is filled with the word vector, which proves that our processing work is successful. These word vectors can be used for subsequent training.



Fig. 4. part of the corpus

### 5.3 Training and Testing

After the corpus is established, it can be trained based on the correspondence between word vectors and fields in corpus. The training is based on LDA, and the principles of LDA have been introduced earlier. In our project, the six categories to be divided into total are deterministic, so it is appropriate to use supervised learning, i.e. sLDA. As mentioned in the introduction to the model principle, we used Gibbs Sampling for the posterior distribution of LDA. The speed of Gibbs Sampling itself is very fast, which makes the training process not take too long. Hundreds of thousands of pieces of data are trained and only used for a few minutes, which reflects the high efficiency of the LDA model in practical applications. During training, we specifically divided some of the data as a test set, and this part of the data did not participate in the training. When the model training is completed, we apply the model to the test set and get a preliminary accuracy of 0.624, as shown in the figure.

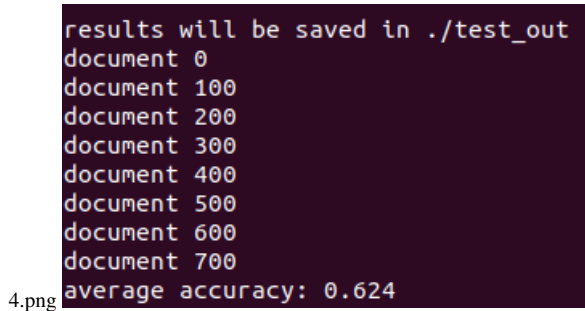


Fig. 5. test results

This result indicates that the model may not be satisfactory if applied directly to the classification. However, considering that we have divided all academic topics into a total of six fields, it is understandable that the accuracy of the results is not high. In fact, this result is enough to show that it is very effective when we combining HMM, sLDA and text processing. The results of 0.624 can fully prove that the sLDA topic model does work in our real-world problem solving.

## 6. CONCLUSION

We studied the principle and application of supervised latent Dirichlet allocation (sLDA) and tried to apply it to solve the field-classification problem of Chinese academic topics in the real-world. In the process of dealing with Chinese academic topics, we solved the derived problems of academic terminology and the

conversion of Chinese text into vectors. We uses a much effective method of discovering new words by applying Hidden Markov Chain (HMM) to improve word segmentation. We trained the processed data based on the sLDA principle and test the resulting model. The final result proved that our application of sLDA for Chinese academic topic classification is effective.

[Mcauliffe and Blei 2008] [Bodrunova et al. 2013] [Bickel and Doksum 2015] [Blei and Lafferty 2009] [Yu 2010] [Morwal et al. 2012]

## REFERENCES

- Peter J Bickel and Kjell A Doksum. 2015. *Mathematical statistics: basic ideas and selected topics, volume I*. Vol. 117. CRC Press.
- David M Blei and J Lafferty. 2009. Topic models. Text mining: theory and applications. (2009).
- Svetlana Bodrunova, Sergei Koltsov, Olessia Koltsova, Sergey Nikolenko, and Anastasia Shimorina. 2013. Interval semi-supervised LDA: Classifying needles in a haystack. In *Mexican International Conference on Artificial Intelligence*. Springer, 265–274.
- Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*. 121–128.
- Sudha Morwal, Nusrat Jahan, and Deepti Chopra. 2012. Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing (IJNLC)* 1, 4 (2012), 15–23.
- Shun-Zheng Yu. 2010. Hidden semi-Markov models. *Artificial intelligence* 174, 2 (2010), 215–243.