The LDA Topic Model Extension Study

Qingquan Yang
Faculty of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, China
949008614@qq.com

Weijiang Li *
Faculty of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, China
522700944@qq.com
* Corresponding Author

Abstract—This article is a literature review that introduces the status quo of research on the LDA in recent years. The topic model of LDA[1] (Latent Dirichlet Allocation) was proposed by D.M. Blei in 2003, which obtained three Bayesian probability model on the extension of the probability of latent indexing (probabilistic Latent Semantic Indexing, pLSI). It consists of the documents, topics and words, which mining implicit subject of the statistical probability model from a semantic document by modeling. In this article, researchers briefly introduce the LDA topic model and the documents generation process, but also introduce several the extended model based on LDA that has a mainstream representative part and advancement, sum up the research methods of extension of the study based on LDA, show the results in the field of study, and give personal suggestions to the future study of topic model based on LDA.

Keywords- LDA; Text Retrieval; Semantics Mining; Topic Model; Document Generation

I. INTRODUCTION

Text retrieval[2], namely the natural language retrieval, is not through any index to the text of the documents according to the documents content, such as keywords, semantics etc. to retrieval, classification, filtering and other operations. With the increase of the number of documents, in the process of text retrieval study, the early text retrieval has been difficult to meet the increasing number of documents for accurate retrieval. Then Xia Lin (1993) introduced the semantic relation graph into the text retrieval, so that the retrieval system can more understand the implicit meaning[3] of the words inputted from the user, and improve the accuracy of the retrieval greatly. Now, text retrieval technology has gradually developed into the query semantic understanding and the specific field[4].

Topic model is a kind of intelligent model[5] based on semantic understanding. It as a new method of text expression has attracted more and more attention in the field of natural language processing system, which excavated the semantic relations[6] among the words of text entry. It set the documents as the probability distribution belong to a series of unspoken theme, and every topic is under the probability distribution of the words in this topic. LDA as a latent semantic indexing topic model which is probability generating, it is a typical representative of the topic development in the field of natural language retrieval. At present, natural language processing has reached a new height in the research for extensions of LDA[7].

II. THE LDA TOPIC MODEL

In 2003, Blei et al. proposed LDA (Latent Dirichlet Allocation) on the basis of the previous research on the topic model, and LDA was an important set of discrete data modeling method. It is based on a common sense assumption: all the text in the document collection share a certain amount of implicit topic. Based on this assumption, it sets the entire document set into a collection of implicit topics, and each test is expressed as the underlying topic of a specific proportion of mixed. The LDA model based on word bag that documents and words are interchangeable, it ignores the order of the words in the document and the document in the order of the corpus, and thus it is easy to text information can be converted to digital information modeling.

A. LDA document generation process

According to the LDA definition of the text sets, the document collection process is as follows:

- Selection of N, N~ Poisson(ζ).Where N represents the length of the document.
- Selecting a multinomial distribution parameter θ, θ ~Dir(α)distribution. Where θ represents the document topic probability distribution, which is column matrix in here.
- Begin to produce N words that belong to the document.

First, choose a topic Z_n, Z_n(obey)~Multinomial(θ) distribution.

There are K topics, which according to the probability of θ parameter selecting one of the topics as Z.

Z is selected from the parameters of the multinomial distribution.

Second, choose W_n from P (W_n | Z_n, β).

Parameter β is multinomial distribution, which is a K * V matrix.

Where K represents topic number, V represents word item number.

β_{ij} = p(W_j = 1|Z_i = 1) represents the probability that comes from a topic Z_i to generate a certain term of W_j.

θ is K-dimension Dirichlet random variable, whose probability density function as shown in algorithm (1):

$$P(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - 1} \quad (1)$$

Joint distribution, which comes from θ, topic: Z and words: W as shown in algorithm (2) :

$$P(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta) \quad (2)$$

Researchers can get the marginal distribution of the document through the integral of θ and calculate the sum of Z, as shown in Equation (3) :

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\Big(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)\Big)d\theta \quad (3)$$

Then calculating the sum of the marginal distribution $P(W|\alpha, \beta)$, researchers can get the probability of the whole corpus, as shown in Equation (4):

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\Big(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta)\Big)d\theta_d \quad (4)$$

B. *Graphical model of LDA*

Researchers can use a graphical model Fig. 1 to show the model of LDA, as shown in the above Equation (4).

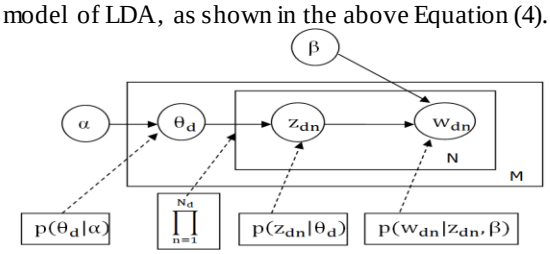

Figure 1. Graphical model representation of LDA.

There are three presentation layers in LDA model. α and β are corpus level parameter, α represents the probability distribution of 'document and topic', β represents the probability distribution of 'topic and words'. θ_d is the documents level variable and represents the probability distribution of 'document and topic' that serial number is d. M is the number of document. Z_{dn} and W_{dn} are lexical item level variable. W_{dn} represents the N th words in documents that serial number is d, it can be directly observed. Z_{dn} is implicit variable and represents W_{dn} topic. The word N is the number of words in a document.

LDA is a kind of unsupervised machine learning technology[8]. It can distinguish the implicit topic information in large lots of document collections[9] and corpus[10]. It apply a method of bag of words which make every one document as a term frequency vector, and turn the text information into the digital information[11] which is easy to be modeled. But the method of bag of words did not consider the order between words and words, which simplifies the complexity of the problem, and provided an opportunity for the improvement of model. Each document represents a probability distribution that consisted by some topics, and each topic represents a probability distribution that consisted by some words[12]. Because of the weak correlation between the random vectors of Dirichlet distribution, it is almost irrelevant between the hypothetical potential topics. This is not consistent with many practical problems, which leads to another legacy[13] of LDA.

III. THE APPLICATION RESEARCH BASED ON LDA

A. *LDA-based document models for ad-hoc retrieval*

In the field of information retrieval, retrieval algorithm which merges some different formats model is an open research strategies. In this article, the researchers attempt to explore how to effectively improve Ad - hoc retrieval using LDA topic model[14].

1) *The main algorithm and plan*

LDA-based document models for ad-hoc retrieval were proposed by Xing Wei et al. From the paper, the authors did an experiment show that directly employing the LDA model hurts retrieval performance. Therefore, they combine the original document model with the LDA model and construct a new LDA-based document model. They formulate the model through a linear combination obtained in one of the following way:

a) Method 1: Linearly combining the original document model and LDA, which is illustrated in (5).

$$p(w|D) = \lambda\left(\frac{N_d}{N_d+\mu}P_{ML}(w|D) + \left(1-\frac{N_d}{N_d+\mu}\right)P_{ML}(w|coll)\right) + (1-\lambda)P_{lda}(w|D) \quad (5)$$

b) Method 2: Additively combining the LDA model with the maximum likelihood estimate of word w in the document D.

c) Method 3: Combining the LDA model with the Dirichlet smoothing part, i.e. the maximum likelihood estimate of word w in the entire collection.

Parameter setting in the paper is for (a), it may be necessary to adjust λ and μ in (b) and (c).

The LDA model has a new representation for a document based on topics. After getting the posterior estimates of θ and φ, the researchers can calculate the probability of a word in a document as Equation (6),

$$P_{lda}(w|d, \hat{\theta}, \hat{\varphi}) = \sum_{z=1}^{K} P(w|z, \hat{\varphi})P(z|\hat{\theta}, d) \quad (6)$$

where, $\hat{\theta}$ and $\hat{\varphi}$ are the posterior estimates of θ and φ respectively.

The LDA model is very complex and cannot be solved by exact inference. Therefore, the researcher used Gibbs sampling and the approximation of $\hat{\theta}$ and $\hat{\varphi}$ can be obtained directly. From a Gibbs sample, they used (7)

$$(n_{-i,j}^{w_i} + \beta_{w_i}) / \sum_{v=1}^{V}(n_{-i,j}^{v} + \beta_v) \quad (7)$$

to approximate $\hat{\varphi}$ and(8)

$$(n_{-i,j}^{d_i} + \alpha_{z_i}) / \sum_{t=1}^{T}(n_{-i,t}^{d_i} + \alpha_t) \quad (8)$$

to approximate $\hat{\theta}$ after a certain number of iterations (burn-in period) being accomplished, where $n_{-i,j}^{w_i}$ is the number of instances of word w_i assigned to topic z=j, not including the current token, α and β are hyper-parameter that determine how heavily this empirical distribution is smoothed, and can be chosen to give the desired resolution in the resulting distribution, $n_{-i,j}^{d_i}$ is the number of words in document d_i assigned to topic a=j, not including the

current token. Thus $\sum_{v=1}^{V} n_{-i,j}^{v}$ is the total number of words assigned to topic z=j; and $\sum_{t=1}^{T} n_{-i,t}^{d_i}$ is the total number of words in document d, not including the current one. Thus(5) will be

$$P(w|D) = \lambda\left(\frac{N_d}{N_d+\mu}P'(w|D) + \left(1-\frac{N_d}{N_d+\mu}\right)P'(w|coll)\right) + (1-\lambda)\left(\sum_{t=1}^{K}\frac{\left(n_{-i,j}^{w_i}+\beta_{w_i}\right)}{\sum_{v=1}^{V}\left(n_{-i,j}^{v}+\beta_v\right)} \times \frac{\left(n_{-i,j}^{d_i}+\alpha_{z_i}\right)}{\sum_{t=1}^{T}\left(n_{-i,t}^{d_i}+\alpha_t\right)}\right) \quad (9)$$

The actual value of $P_{lda}(W|D)$ is an average of the ones from several Markov Chains.

In experiment, There are several parameters that need to be determined in their experiments. For the retrieval experiments, the proportion of the LDA part in the linear combination must be specified . For the LDA estimation, the number of topics must be specified; the number of iterations and the number of Markov chains also need to be carefully tuned due to its influence on performance and running time. They use the AP collection as their training collection to estimate the parameters. The WSJ, FT, SJMN, and LA collections are used for testing whether the parameters optimized on AP can be used consistently on other collections. All parameter values are tuned based on average precision since retrieval is their final task. The parameter selection process, including the training set selection, also follows Liu and Croft (2004) to make the results comparable. Mean average precision is used as the basis of evaluation throughout this study.

From their paper, researchers know the author use symmetric Dirichlet priors in the LDA estimation with α=50/K and β=0.01.Because retrieval results are not very sensitive to the values of these parameters.

Through the experiment and comparison, the researchermade the following conclusions:

Firstly, experiments performed in the language modeling framework, including combination with the relevance model, have demonstrated that the LDA-based document model consistently outperforms the cluster-based approach, and the performance of LBDM is close to the Relevance Model.

Secondly, they had shown that the estimation of the LDA model on IR tasks is feasible with suitable parameters based on the analysis of the algorithm complexity and empirical parameter selections.

Finally, LDA-based retrieval can potentially be used in applications where pseudo-relevance feedback would not be possible.

This paper provides a new train of thought for researchers to study the LDA model, the authors' well file based on the LDA model was used to optimize the ad-hoc retrieval, greatly improving the retrieval performance.

IV. EXTENDED RESEARCH BASED ON LDA

A. *The Author-Topic Model for Authors and Documents*

The Author-Topic model for the authors and documents [15]was proposed by Michal Rosen-Zvi et al. In this paper, it described a generative model for document collections, the author-topic model ,that simultaneously models the content of documents and the interests of authors[16]. This generative model represents each document with a mixture of topics and extends these approaches to author modeling by allowing the mixture weights for different topics to be determined by the authors of the document.

The author-topic model draws upon the strengths of the two models ,one that models documents as a mixture of topics (LDA, Blei et al.,2003), one that models authors with distributions over words. It uses a topic-based representation to model both the content of documents and the interests of authors. As in the author model, a group of authors ,ad ,decide to write the document d. For each word in the document an author is chosen uniformly at random. Then, as in the LDA topic model, a topic is chosen from a distribution over topics specific to that author, and the word is generated from the chosen topic.

The graphical model corresponding to this process is shown in Fig. 2:

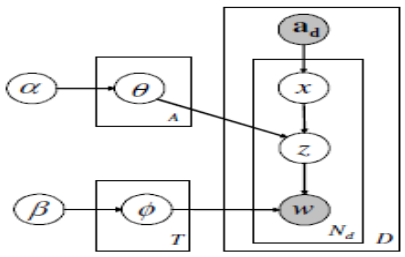

Figure 2. The author-topic model.

As in the author model, X indicates the author responsible for a given word, chosen from ad. Each author is associated with a distribution over, θ, chosen from a symmetric Dirichlet(α) prior. The mixture weights corresponding to the chosen author are used to select a topic Z, and a word is generated according to the distribution φ corresponding to that topic, drawn from a symmetric Dirichlet(β) prior.

The author-topic model subsumes the two models described above as special cases: topic models like LDA correspond to the case where each document has one unique author, and the author model corresponds to the case where each author has one unique topic. By estimating the parameters φ and θ, researchers obtain information about which topics authors typically write about, as well as a representation of the content of each document in terms of these topics.

This article introduced a simple algorithm, which is Gibbs sampling algorithm[17] to estimate parameters of topic models. This is because it provides a simple method for obtaining parameter estimates under Dirichlet priors[18] and allows combination of estimates from several local maxima of the posterior distribution. The LDA model has two sets of unknown parameters, the D document distributions θ, and the T topic distributions φ, as well as the latent variables corresponding to the assignments of individual words to topics z. By applying Gibbs sampling, researchers can construct a Markov chain that converges to the posterior distribution on z and then uses the results to infer θ and φ. Equation (10) is a Markov chain constructed with standard Dirichlet integrals.

$$P(z_i = j | w_i = m, z_{-i}, w_{-i}) \propto \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{DT}_{dj} + \alpha}{\sum_{j'} C^{DT}_{dj'} + T\alpha} \quad (10)$$

For any sample from this Markov chain, being an assignment of every word to a topic, researchers can estimate θ and φ using (11) and (12):

$$\varphi_{mj} = \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \quad (11)$$

$$\theta_{dj} = \frac{C^{DT}_{dj} + \alpha}{\sum_{j'} C^{DT}_{dj'} + T\alpha} \quad (12)$$

In the author-topic model, there are two sets of latent variables: z and x. So researchers draw each (z_i, x_i) pair as a block, conditioned on all other variables. Equation (13) is the conditional probability derived by marginalizing out the random variables φ and θ:

$$P(z_i = j, x_i = k | w_i = m, z_{-i}, x_{-i}, w_{-i}, a_d) \propto \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \frac{C^{AT}_{kj} + \alpha}{\sum_{j'} C^{AT}_{kj'} + T\alpha} \quad (13)$$

These random variables are estimated from samples via Equation (14) and Equation (15):

$$\varphi_{mj} = \frac{C^{WT}_{mj} + \beta}{\sum_{m'} C^{WT}_{m'j} + V\beta} \quad (14)$$

$$\theta_{kj} = \frac{C^{AT}_{kj} + \alpha}{\sum_{j'} C^{AT}_{kj'} + T\alpha} \quad (15)$$

In the process of experiment, the hyper-parameters α and β are fixed at 50/T and 0.01 respectively.

The author-topic model provides a relatively simple probabilistic model for exploring the relationships between authors, documents, topics, and words. The primary benefit of the author-topic model is that it allows researchers to explicitly include authors in documents models, providing a general framework for answering queries and making predictions at the level of authors as well as the level of documents.

V. CONCLUSIONS

The LDA model is a new method of text representation, help to improve the precision of information retrieval system[19]. LDA model at the core of the algorithm has been introduced in this paper. Based on the introduced two kinds of retrieval model based on the LDA, these two kinds of model provides a new train of thought for researchers to research and extension the LDA model. The first model is to modify the LDA model to adapt to the AD-hoc retrieval. The second model is the fusion of the LDA model several features to improve the precision.

REFERENCES

[1] David M. Blei , Andrew Y. Ng , Michael I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research, 3, p.993-1022, 3/1/2003.

[2] Ramage D, Manning C D, Dumais S. Partially labeled topic models for interpretable text mining[C]// In Proceedings of KDD. 2011:457--465.

[3] Min Huang, Mao-sheng Lai. Semantic retrieval research review [J]. Journal of library intelligence, 2008, 52 (6) : 63-66.

[4] Xiaohui Zou, Jing Sun. LDA theme model [J]. Journal of intelligent computer and applications, 2014, 4 (5) : 105-105. The DOI: 10.3969 / j.i SSN. 2095-2163.2014.05.031.

[5] D. Blei and J. Lafferty. A correlated topic model of Science. The Annals of Applied Statistics, 1(1):17–35, 2007.

[6] Aaron Heidel, Hung-an Chang, Lin-Shan Lee, et al. Language model adaptation using latent dirichlet allocation and an efficient topic inference algorithm.[J]. Proc of Interspeech, 2007.

[7] T. Hofmann. Probabilistic latent semantic indexing. In 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 1999: 50-57.

[8] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]// In Proc. of EMNLP. 2009:248-256.

[9] Liao X, Wang Y, Fan X, et al. National security vulnerability database classification based on an LDA topic model[J]. Journal of Tsinghua University, 2012, 52(10):1351-1355.

[10] Blei D M, Mcauliffe J D. Supervised Topic Models[J]. Preparation, 2010:327-332.

[11] Zhang X, Zhou X, Huang H, et al. An Improved LDA Topic Model[J]. Journal of Beijing Jiaotong University, 2010, 34(2):111-114..

[12] Doshi-Velez F, Wallace B, Adams R. Graph-Sparse LDA: A Topic Model with Structured Sparsity[J]. Eprint Arxiv, 2014.

[13] Mimno D, Mccallum A. Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression[J]. University of Massachusetts - Amherst, 2012:411–418.

[14] Xing Wei , W. Bruce Croft, LDA-based document models for ad-hoc retrieval, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, August 06-11, 2006, Seattle, Washington, USA.

[15] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-Topic Model for Authors and Documents[J]. Conference on Uncertainty in Artificial Intelligence Natasa Jovanovic Rieks Op Den Akker & Anton Nijholt, 2012:487-494.

[16] Hu P, Liu W, Jiang W, et al. Latent Topic Model Based on Gaussian-LDA for Audio Retrieval[M]// Pattern Recognition. Springer Berlin Heidelberg, 2012.

[17] Gildea D, Hofmann T. Topic-Based Language Models Using EM[J]. Proceedings of Eurospeech, 1999:2167–2170.

[18] Zhu J, Ahmed A, Xing E P. MedLDA: maximum margin supervised topic models for regression and classification.[J]. Journal of Machine Learning Research, 2009, 13(4):2237-2278.

[19] Gruber A, Rosen-Zvi M, Weiss Y. Latent topic models for hypertext[C]//InUAI.2008.