Journal of Korea Multimedia Society Vol. 23, No. 7, July 2020(pp. 883-890)
https://doi.org/10.9717/kmms.2020.23.7.883

Variational Expectation-Maximization Algorithm in Posterior Distribution of a Latent Dirichlet Allocation Model for Research Topic Analysis

Jong Nam Kim†

ABSTRACT

In this paper, we propose a variational expectation-maximization algorithm that computes posterior probabilities from Latent Dirichlet Allocation (LDA) model. The algorithm approximates the intractable posterior distribution of a document term matrix generated from a corpus made up by 50 papers. It approximates the posterior by searching the local optima using lower bound of the true posterior distribution. Moreover, it maximizes the lower bound of the log-likelihood of the true posterior by minimizing the relative entropy of the prior and the posterior distribution known as KL-Divergence. The experimental results indicate that documents clustered to image classification and segmentation are correlated at 0.79 while those clustered to object detection and image segmentation are highly correlated at 0.96. The proposed variational inference algorithm performs efficiently and faster than Gibbs sampling at a computational time of 0.029s.

Key words: Variational Inference, KL-Divergence, Expectation-Maximization, Likelihood

1. INTRODUCTION

Famous search engines like Google, Yahoo, and Bing use different machine learning algorithms to collect and organize information as prompted by the user. These tools have the capacity to organize and categorize information using next-word predictive algorithms, clustering algorithms and others. The information produced based on the query from the user is basically matched to the other information existing in a database. The queried information can be identified based on related terms, topics or clusters of words that are collectively deployed using similarities and differences. The information search process can be intelligently done by using probabilistic generative topic models for machine learning experts. Therefore, this work aims at employing topic models to determine the main objective.

Topic modeling is a technique related to topic clustering which uses statistical models to obtain topics and themes from documents in a corpus [1]. Topic models are basically graphical or probabilistic models that use mathematical functions for mapping to discover the structures of texts [2]. They are commonly used to collect, organize, categorize, and sometimes distinguish texts into organized groups whin in corpora purpose purposely for acquiring sentiments, main ideas from the themes [3]. Generally, they can give inference on text summarization of different documents collectively combined in a corpus. There are different

※ Corresponding Author : Jong Nam Kim, Address: (48513) Yongso-ro 45, Nam-gu, Busan, Korea, TEL : +82-51-629-6259, FAX : +82-51-629-6263, E-mail : jnkim1225 @gmail.com

Receipt date : Jul. 14, 2020, Revision date : Jul. 15, 2020
Approval date : Jul. 16, 2020

† Dept. of IT Convergence & Application Eng. Pukyong National University

※ This research was supported by HK+ Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF - 2017S1A6A3A01079869)

models that have been used to perform topic modeling and most of them are developed from probability and linear algebra. The probabilistic models include Latent Dirichlet Allocation LDA, Correlation Topic Models (CTM), Latent Semantic Analysis (LSA), and probabilistic Latent Semantic Analysis (pLSA) [4]. The models based on algebraic transformation includes Term Frequency-Inverse Document Frequency (TF-IDF), Non-negative Matrix Factorization (NMF), and Latent Semantic Indexing (LSI) [5].

The motivation of this work is to find a fast and efficient machine learning algorithm that determines test similarity for large corpora in a short time. This work is primarily mentioned in [6] where we proposed a probabilistic generative model with Dirichlet prior that determines posterior distribution using a Gibbs Sampler. The algorithm was efficient however, it was quiet slow. In this work we propose a variational bayesian approach to a probabilistic generative model (LDA) that is capable of obtaining the optimal solution by searching the local optima with an approximation of reaching the global optima using the lower bound of the true distribution.

This paper has five sections whereby the second section presents the related works and the third section describes the proposed work. Section four shows experimental results and discussion while the last section presents concluding remarks.

2. RELATED WORKS

Realization of machine learning problems in a field of natural language processing require both statistical models and mathematical functions to discover the themes and their abstract meanings. This situation is referred to as topic modeling whereby different models are developed and used to perform the main goal of this work. An early topic model was described in [7] called as Latent Semantic Indexing (LSI), the second model called probabilistic latent semantic analysis (PLSA) was invented by Hofmann in [8] that has some advantages over the former model. With such a trend of topic models development, the Latent Dirichlet allocation (LDA) model was developed and famously known as a common topic model referred to as a generalization of pLSA invented by Blei et al., in [9]. This model introduces sparse Dirichlet prior distributions over document-topic and topic-word distributions within a corpus. It also encodes the documents in the corpus to cover a small number of topics relative to the number of embedded or related words.

The newly developed models are the general extensions of LDA, such as Pachinko allocation [10], Correlated Topics Model (CTM) [11], Hierarchical latent tree analysis (HLTA) [12], and among others. The Pachinko Allocation improves LDA by modeling correlations between topics, so does CTM. Hierarchical latent tree analysis (HLTA) is an alternative to LDA that models word co-occurrences using latent variables and the states of the latent variables corresponding to clusters of documents referred to as topics.

With topic modeling problems, LDA model is normally trained with two learning algorithms which are Gibbs sampling and variational expectation-maximization (VEM). The former is a Markov Chain Monte Carlo (MCMC) algorithm for obtaining approximate sampling of the posterior when direct sampling is not possible while the latter is the iterative expectation algorithm that maximizes the likelihood by minimizing the entropy of the true posterior distribution. A Gibbs sampler has a lot of advantages on LDA as stated in [9] but Moghaddam et al, in [13] clearly state that, training LDA with Gibbs sampling is computationally expensive and practically slow because of long convergence caused by the difficulty in finding the posterior of each topic assigned to a document. Generally, the samples that initiates the Markov process are often discarded because they may not accurately represent the desired distribution. Based on the drawbacks mentioned thereof, VEM is introduced to train the LDA model for realizing a topic modeling problem.

The VEM algorithm is applicable for complex models with many latent variables. Moreover, it is very useful when the posterior distribution does not rely on point estimates. Generally, when the latent variables are detected, the algorithm treats them as random variables with the same distribution to the prior [14]. Comparatively, Gibbs sampling is guaranteed in the limit to recover the true posterior while VEM does not. In that sense, VEM breaks the links in the graphical model for LDA in order to make the mathematical computation easier at the cost of minimum variance and unbiased estimates [15]. Generally, with VEM, the search of optimal solution sticks to the local optima, and it becomes an approximation after reaching the global optima from the lower bound of the true distribution [16].

Despite the challenging applications of the two learning algorithms on LDA, VEM performs relatively better and faster than the Gibbs sampler. The two algorithms have relative complexity in terms of determining the optimal results based on the dimensions of dataset, training model, and the problem to be realized. This work aims at employing VEM algorithm for topic modeling on LDA and use the posterior distribution for computing documents similarity.

3. PROPOSED WORK

In this work we propose a variational Bayes approach that learns the LDA model to produce posterior probabilities which are used to determine the documents similarity coefficients. The LDA model based on the proposed learning algorithm behaves as a predictive probabilistic topic model that has an ability to create two distinct matrices with similar properties to that obtained by algebraic procedures. We employ variational expectation-maximization algorithm to approximate the posterior probability by minimizing the lower bound of the true probability.

Suppose that the LDA model has a document matrix Φ with the hyper-parameters α and β for the topic document distribution in a corpus D. We define the distribution of topics in every document as $p(\Phi|\alpha)$ which is a multinomial distribution expressed in Eq. 1.

$$p(\Phi|\alpha) = \frac{\Gamma(\sum_{i}^{K}\alpha_i)}{\prod_{i}^{K}\Gamma(\alpha_i)}\prod_{i}^{K}\Phi_i^{\alpha_i - 1}...\Phi_{K-1}^{\alpha_i - 1} \tag{1}$$

where K is the number of topics in a document matrix. We define the posterior distribution of the latent variables that are assumed to exist with similar distribution to the prior in Eq. 2.

$$p(\Phi, t|w, \alpha, \beta) = \frac{p(\Phi, t, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{2}$$

where t and w are latent variables for topics and words in a corpus respectively.

We simplify the numerator by using probability laws and properties to yield an expression represented in Eq. 3.

$$p(\Phi, t|w, \alpha, \beta) = \frac{p(w|t, \beta)p(t, \Phi)p(\Phi|\alpha)}{p(w|\alpha, \beta)} \tag{3}$$

where the denominator is the multinomial dis tribution of documents within a corpus, simply de-as $p(w|\alpha, \beta) = \prod_{i=1}^{N}\beta_{t_n w_n}$. In this case, the parameter t_n and w_n are the n-th topics and words for every document in a corpus. From Eq. 3, we marginalize $p(w|\alpha, \beta)$ and obtain an intractable distribution because the model parameters are inseparable when maximizing the log-likelihood. The marginalization of the above equation results to Eq. 4.

$$p(w|\alpha, \beta) = \left(\frac{\Gamma(\sum_{i}^{K}\alpha_i)}{\prod_{i}^{K}\Gamma(\alpha_i)}\right)\int_{i=1}^{N}\left(\prod_{i}^{K}\Phi_i^{\alpha_i - 1}\right)\left(\prod_{i=1}^{N}\sum_{i=1}^{K}\prod_{j=1}^{T_n}(\Phi_i\beta_{ij})^{w_i^j}\right)d\Phi \tag{4}$$

where T_n is the total number of topics for every document in a corpus. Provided that the above equation is intractable then we introduce VEM that maximizes the log likelihood by minimizing the log-likelihood of the true posterior.

$$\log p(w|\alpha,\beta) = \log\int\sum_t p(\Phi,t,w|\alpha,\beta)d\Phi \tag{5}$$
$$= \log\int\sum_t \frac{p(\Phi,t,w|\alpha,\beta)q(\Phi,t)}{q(\Phi,t)}d\Phi$$

where $q(\Phi,t)$ is the prior distribution for the topics from the topic document matrix. We then, introduce a variational objective function $L(\lambda,\phi,\alpha,\beta)$ in Eq. 5 to determine the lower bound of the log likelihood in Eq.5 above. We apply Jensen inequality in [14] to estimate the lower bound of the posterior.

$$L(\lambda,\phi,\alpha,\beta) \geq \int\sum_t q(\Phi,t)\log p(\Phi,t,w|\alpha,\beta)d\Phi$$
$$-\int\sum_t q(\Phi,t)\log p(\Phi,t)d\Phi \tag{6}$$

where λ and ϕ are the posterior parameters of the objective function, and then we apply the expectation properties based on the expansion of Eq. 6. The resulting expression represents Eq. 7.

$$L(\lambda,\phi,\alpha,\beta) = \int\sum_t q(\Phi,t)\log p(\Phi|\alpha)d\Phi + \int\sum_t q(\Phi,t)\log p(t,\Phi)d\Phi + \tag{7}$$
$$\int\sum_t q(\Phi,t)\log p(w|t,\beta)d\Phi - \int\sum_t q(\Phi,t)\log p(\Phi,t)d\Phi$$

For further simplification of Eq. 7, we express the objective function in terms of expected values for the maximized log-likelihood equation. This results to Eq. 8 comprised of a Shanon Entropy of the prior.

$$L(\lambda,\phi,\alpha,\beta) = E_q[\log p(\Phi,\alpha)] + E_q[\log p(t,\Phi)]$$
$$+ E_q[\log p(w|t,\beta)] + H(q) \tag{8}$$

where $H(q)$ is the Shanon entropy. Moreover, we can easily verify that the log-likelihood of the posterior is equivalent to the summation of the lower bound of the variational function and the Kullback-Leibler divergence (relative entropy) in Eq. 9.

$$\log p(w|\alpha,\beta) = L(\lambda,\phi,\alpha,\beta) + D[q(\Phi,t|\lambda,\phi)||p(\Phi,t|\alpha,\beta,w)] \tag{9}$$

Lastly, we maximize the lower bound in Eq. 9 by minimizing the KL-divergence to obtain the posterior λ and ϕ. After obtaining the posterior we then determine the correlation coefficients among documents with Pearson correlation coefficient in Eq. 10.

$$r = \frac{n(\sum_{i,j=1}^{n} d_i d_j) - \sum_i^n d_i \sum_j^n d_j}{\sqrt{n[\sum_{i=1}^{n} d_i^2 - (\sum_{i=1}^{n} d_i)^2]n[\sum_{j=1}^{n} d_j^2 - (\sum_{j=1}^{n} d_j)^2]}} \tag{10}$$

The procedure of the proposed algorithm is summarized in Fig. 1 whereby the first block with corpus represents a corpus made up by 50 abstracts in the field of computer vision. The second block with LDA represents the topic model that is learned by the variational inference to yield various results stored in the third block. This block with Model Parameters has all the parameters such as term-document matrix, topic-document matrix, and the priors (α and β), to mention a few. We use the posterior in the previous block to find the posterior probabilities of the topic-document matrix. In the last block with Document Similarity, we use Pearson correlation coefficient to determine the document similarities.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental setup and procedures of this work were conducted through a computer with GPU operating on windows 10 installed with R program which was used for analysis. Natural language processing libraries were installed to facilitate model loading, training and prediction. The data used in the experiments were the extracted abstracts from scientific papers published by IEEE journals from 2010 to 2019 in the field of “motion estimation”, “image classification”, “image seg-

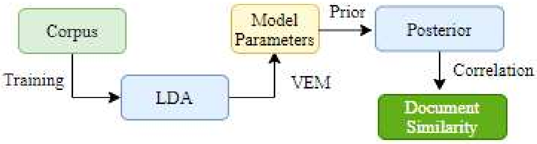

Fig. 1. Procedure of the proposed algorithm.

mentation", "object detection" and "3D reconstructions". The extracted abstracts have different titles, approaches and even writing styles because were written by different authors.

We initially created a corpus and then pre-processed the dataset by transforming to lower case, removing special symbols, numbers, punctuations, white spaces and general errors and formulated a document term matrix which is then trained in LDA model with VEM algorithm. We used a non fixed VEM algorithm to obtain the results as shown in Fig. 2 below.

In Fig. 2, LDA is trained with the VEM algorithm which performs fast within 0.029 milliseconds showing that the assigned topics are image segmentation, motion estimation, image classification, object detection and 3D reconstruction. The above information in the figure can be summarized as shown in Table 1 based on the topic-document probability distribution.

Table 1 summarizes the topic-word distribution with the most frequently used terms for each topic to every document. The terms relevance is arranged in terms of probabilities such that every document has five topics and the total probability for all topics in a document is equal to 1. Table 2 indicates the posterior probabilities for each topic in every document within the corpus.

The highest posterior probability from Table 2 is observed to be 0.99 for abstract25.txt which was

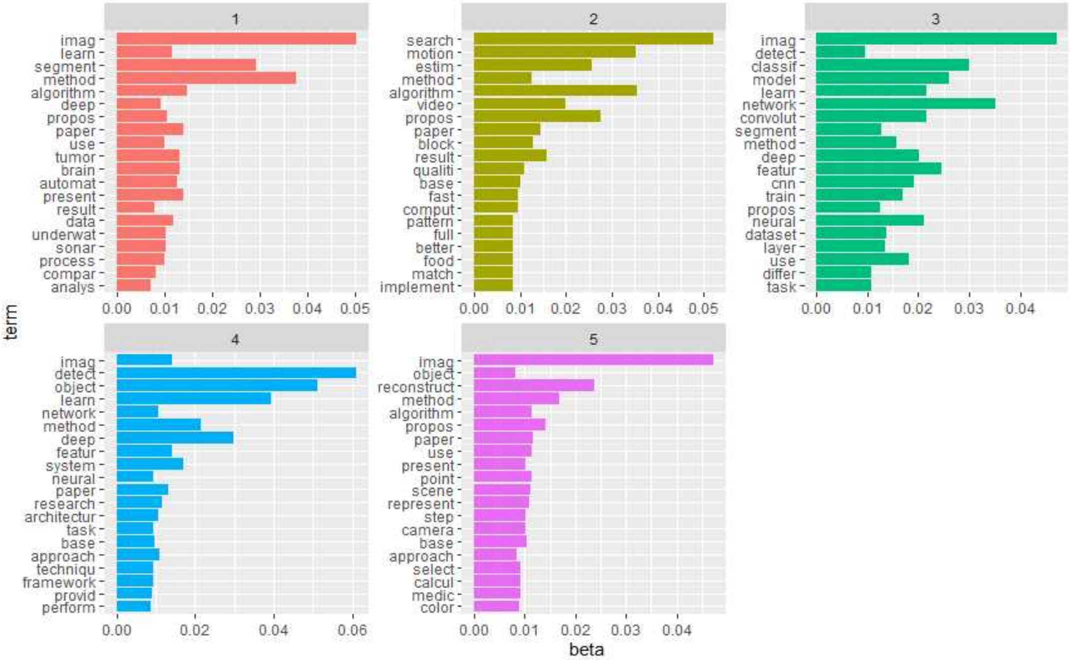

Fig. 2. Frequent terms in five clustered topics using LDA with VEM.

Table 1. Showing topic-word distribution for each assigned topic

Term	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	image	search	image	image	image
2	method	motion	detection	detect	object
3	segment	estimation	classification	object	reconstruct
4	algorithm	method	model	learn	method

Table 2. Topic probabilities by documents (abstracts)

Abstracts	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
10	0.54	0.00	0.46	0.00	0.00
15	0.23	0.77	0.00	0.00	0.00
20	0.09	0.08	0.11	0.60	0.11
25	0.99	0.00	0.00	0.00	0.00
30	0.00	0.99	0.00	0.00	0.00
35	0.77	0.23	0.00	0.00	0.50
40	0.00	0.00	0.00	0.26	0.74

correctly assigned to topic 1. The other probability values lag behind the highest value including 0.77, 0.54, and 0.46. The posterior probabilities from this table are finally used to calculate the correlation coefficients among documents.

Table 3, shows the similarity coefficients among documents based on topic assignment within a corpus. Based on the model output in Fig. 1, the documents clustered to image segmentation category are highly correlated with a coefficient of 0.96, this indicates that object detection and motion estimation fields are closely related. Topics clustered to image segmentation and image classification are positively correlated with a coefficient of 0.41. Moreover, documents assigned to motion estimation and objects detection clusters are positively correlated at a weak correlation coefficient of 0.06. However, documents clustered to motion estimation and 3D reconstruction are negatively correlated to a coefficient of −0.80. From the experimental results above, we noted that the proposed algorithm is fast and efficient for assigning topics in different documents in a corpus. It also defines the true relationship between the topics in every document within a corpus. Additionally, it categorizes the documents based on their differences by yielding negative correlation coefficients. The proposed algorithm is recommended for estimating document similarity at a reasonable pace because it obtains correlation coefficients for related and unrelated documents easily.

5. CONCLUSION

In this paper, we proposed a variational expectation-maximization algorithm that computes the posterior probabilities from LDA model. The characteristics of the proposed algorithm was to esti-

Table 3. Document similarity based on Pearson correlation coefficient

	A1	A5	A11	A15	A21	A25	A31	A35	A41	A45
A1	1.00	0.79	0.41	−0.33	−0.11	−0.25	−0.25	−0.34	−0.25	−0.25
A5	0.79	1.00	0.13	0.29	−0.40	0.40	0.40	−0.28	−0.40	0.40
A11	0.41	0.13	1.00	−0.53	−0.40	−0.40	−0.40	−0.53	−0.41	−0.40
A15	−0.33	0.29	−0.53	1.00	0.04	0.96	0.96	−0.34	−0.33	0.96
A21	−0.11	−0.40	−0.40	0.04	1.00	−0.25	−0.03	0.95	−0.25	−0.25
A25	−0.25	0.40	−0.40	0.96	−0.25	1.00	−0.25	−0.34	−0.25	−0.25
A31	−0.25	0.40	−0.80	0.96	−0.25	−0.25	1.00	0.06	−0.35	0.99
A35	−0.34	−0.28	−0.53	−0.34	0.95	−0.34	0.06	1.00	−0.34	−0.25
A41	−0.25	−0.40	−0.41	−0.33	−0.25	−0.25	−0.35	−0.34	1.00	−0.35
A45	−0.25	0.40	−0.40	0.96	−0.25	−0.25	0.99	−0.25	−0.35	1.00

mate the posterior distribution based on the variational inference. The the proposed framework determined the posterior probabilities used for computing document similarity at a fast speed. The results indicate that the large the number of topics chosen the small the weights for topic assignment which leads to low correlation coefficient values. Additionally, the results indicate that image classification and segmentation are closely related fields while image classification and motion estimation are not closely related fields. Moreover, 3D reconstruction and object detection are closely related fields however, the correlation coefficient value was very small. From the results, we can conclude that the proposed algorithm has good performance in terms of speed compared to that trained with Gibbs sampling even though the results are relatively different.

REFERENCE

[1] T. Yang, J. Torget, and R. Mihalcea, "Topic Modeling on Historical Newspapers," *Proceeding of the Association for Computational Linguistics: Human Language Technologies Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 96-104, 2011.

[2] M. Lamba, "Mapping of Topics in DESIDOC Journal of Library and Information Technology, India: A Study," *Scientometrics,* Vol. 120, No 20, pp. 477-505, 2019.

[3] Y. Zaho and Y. Cen, *Data Mining Applications with R*, Academic Press, Cambridge, Massachusetts, 2013.

[4] D. Blei, "Probabilistic Topic Models," *Communications of ACM,* Vol. 55, No. 4, pp. 77-84, 2012.

[5] D. Lee and H. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, Vol. 401, pp. 788-791, 1999.

[6] J. Mlyahilu and J. Kim, "Generative Probabilistic Model with Dirichlet Prior Distribution for Similarity Analysis of Research Topic," *Journal of Korea Multimedia Society*, Vol. 23, No. 4, pp. 595-602, 2020.

[7] C. Papadimitriou, P. Raghavan, H, Tamaki, and S. Vempala, "Latent Semantic Indexing: A Probabilistic Analysis," *Proceedings of the 1998 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,* pp. 159-168, 1998.

[8] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proceeding of International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval,* pp. 50-57, 1999.

[9] D. Blei, A. Ng, and M. Jordan, "A Latent Dirichlet Allocation," *Journal of Machine Learning Research,* Vol. 3, pp. 993-1022, 2003.

[10] W. Li and A. McCallum, "Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations," *Proceedings of the International Conference on Machine Learning, pp. 577-584,* 2006.

[11] D. Blei and J. Lafferty, "Correlated Topics Model of Science," *The Annals of Applied Statistics,* Vol. 1, No. 1, pp. 17-35, 2007.

[12] T. Liu, N. Zhang, and P. Chen, "Hierarchical Latent Tree Analysis for Topic Detection," *Lecture Notes in Computer Science,* Vol. 8725, pp. 256-272, 2014.

[13] S. Moghaddam and E. Martin, "On the Design of LDA Models for Aspect-based Opinion Mining," *Proceedings of ACM International Conference on Information and Knowledge Management,* pp. 803-812, 2012.

[14] W. Fox and S. Roberts, "A Tutorial on Variational Bayesian Inference," *Artificial Intelligence Review,* pp. 1-11, 2012.

[15] D. Tzikas, A. Likas, and N. Galatsanos, "The Variational Approximation for Bayesian Inference," *IEEE Signal Processing Magazine,* Vol. 25, No. 6, pp. 131-146, 2005.

[16] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning,* Vol. 37. No. 2, pp. 183-233, 1999.

[17] M. Nam, E. Lee, and J. Shin, "A Method for User Sentiment Classification Using Instagram Hashtags," *Journal of Korea Multimedia Society,* Vol. 18, No. 11, pp. 1391-1399, 2015.

Jong Nam Kim

1997. 2: Dept. of Information Telecommunication Eng., GIST (Master)

2001. 8: Dept. of Mechatronics Eng., GIST (Ph.D)

2001. 7∼2004. 2: Researcher at KBS

2004. 3∼Current: Professor at Dept. of IT Conv. and Apps. Eng., PKNU

Interest fields: video compression, multimedia data processing, computer vision, machine learning