Global Journal of Advanced Engineering Technologies *Volume 5, Issue 4- 2016*
ISSN (Online): 2277-6370 & ISSN (Print):2394-0921

A SURVEY: FAST ONLINE EXPECTATION MAXIMIZATION FOR BIG TOPIC MODELLING

[1] *Disha R. Shinde ,,* [2] *A.S. Vaidya*
[1]*Student, ME (CSE), Gokhale Education Society R.H.Sapat College of Engineering management studies and Research , Savitribai Phule Pune University,Nashik, Maharashtra, India.*
[2]*Professor, Gokhale Education Society R.H.Sapat College of Engineering management studies and Research, Savitribai Phule Pune University, Nashik, Maharashtra, India*

Abstract: *Expectation maximization algorithm can compute the most likelihood word. The FOEM Fast Online Expectation maximization denotes the topic distribution from unseen documents. The different models like LDA, Lifelong topic model, Topic modeling has been commonly used for to find the topic from topic collection. We first describe the latent dirich let allocation (LDA) which is the simplest kind of topic model. Gibbs sampling is one of the most impotent method in LDA model. FOEM is more efficient for some lifelong topic modeling activities. With these algorithms, various case studies have been performed and results are analyzed for making subjective and objective studies in topic model.*
Keywords: *Expectation Maximization algorithm; LDA; big topic modeling, Gibbs Sampling, FOEM*

I. INTRODUCTION

Probabilistic Topic models are a collection of algorithms whose objective is to discover the hidden words in large collection of documents.LDA is one of the most important topic model paradigms. In this article we review the LDA has found many applications like natural language processing, machine learning, and computer vision. Batch LDA algorithm corporate the Expectation Maximization algorithm, Gibbs Sampling (GS) Gibbs sampling is commonly used for statistical inference, especially Bayesian inference. It is a Random number algorithm and it is alternative of deterministic algorithm such as Expectation Maximization algorithm ,collapsed variation byes(VBS) , belief Propagation (BP) The BP algorithm is easy to understand and implement ,faster and accurate than two approximate methods like VB and GS in topic modeling activities.

We need lifelong topic modeling algorithms that denotes a large number of parameter of LDA from big data collection without big topic modeling. However previous batch algorithm had to sweep repeatedly the all dataset until convergent, so they have large time and space complexity.

It is popular for maximum likelihood word estimation; EM algorithm naturally is an Betterment algorithm, in the sense that it increases the likelihood at each iteration. It is simple to implement and contains two steps:

- ***E-Step: -*** *This step involves expectation over conditional distribution of the latent data given the observation.*
- ***M-Step: -*** *This step involves an analogous to complete data weighted maximum likelihood estimation.*

Big topic modeling has shown possible business values in real world business application such as search engine etc.

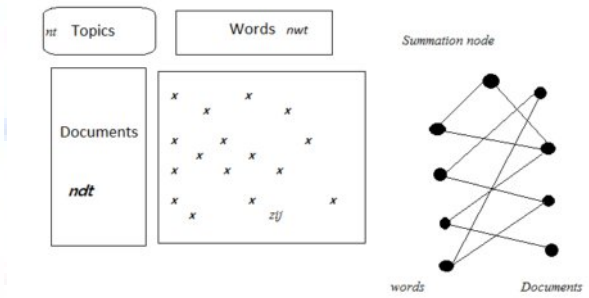

Figure 1: General structure of Topic Mining

More specifically, big topic modeling requires handling following activities:

- *When the data collection is too large to fit in a memory;*
- *When the number of LDA parameters is too big to fit in memory;*
- *When the number of extracted topics is very large;*
- *When the vocabulary size in data collection is very large.*

The above four activities can be categorized into two problems:

- *Big Data*
- *Big Model.*

II. RELATED WORK

A) Probabilistic Topic Modelling

Asuncion and Max Welling had introduced a Latent Dirichlet analysis or Topic Modeling; It is a high-dimensional Latent Variable Framework for sparse count data. This paper represents the close connection between Gibbs sampling, variation inference, and maximum a posteriori estimation approaches [1]. Jia Zeng introduced a topic modeling toolbox which is based on Belief propagation Algorithm. The BP algorithms for learning LDA-based topic models. The current version includes BP algorithms for latent Dirichlet allocation (LDA),

- *Author topic models (ATM),*
- *Relational topic models (RTM),*
- *Labeled LDA (LaLDA)*

Toolbox is an ongoing project [2]. D. Newman and Arthur Asuncion had described an approach distributed algorithms for two widely-used topic models, which are Latent Dirichlet Allocation model, and the Hierarchical Dirichlet Process model. [3]. In distributed algorithm data is partitioned across separate processors or individual processors and Inference (speculation) is parallel. The algorithm is simple to implement and easy to understand, It can be viewed as an approximation to Gibbs-sampled LDA. [3].

Jia Zeng showed the concept of Speeding up Topic Modeling [4]. Fast batch LDA algorithms have attracted intensive research interests recently. The proposed active belief propagation algorithm speedup training LDA significantly with a comparable topic modeling precision to the state-of-the-art batch LDA algorithms. [4].

David M. Blei had brought forward the concept to develop a new algorithm for finding most likelihood words in large Datasets, where the need was to deal with other needs like such as more than one objectiveness, user priorities, etc in their paper [5]. Probabilistic Topic model is a collection of algorithms whose objective is to discover the hidden thematic structure in large documents.LDA is simplest kind of probabilistic Topic model; it incorporates Meta data in to analysis of documents [5].

Statistical topic modeling is useful tool for analyzing large unstructured Document collections. There is a significant body of work introducing and developing sophisticated topic models and their applications. Evaluation is an important issue: In the unsupervised nature of topic models makes model selection difficult. [6].Chong Wang and David M. Blei had proposed one new algorithm to recommend scientific articles to users of an online community. Finding relevant Documents has become more difficult. So Newly formed online communities of researchers sharing citations provides a new way to solve this problem [7].

B) Expectation Maximization

The basic Idea of Online Expectation maximization algorithm Is to partition of Data stream of D documents into small mini batches with size Ds, OEM combines Incremental Expectation Maximization (IEM). Xiaosheng Liu and Jia Zeng had studied To handle web-scale content analysis on just a single PC [7]. Authors propose the multi-core parallel expectation-maximization (PEM) algorithms to denote and Forecast LDA parameters in shared memory systems by avoiding memory access conflicts reducing the locking time among multiple t heads. Parallel LDA toolbox is made publicly available as open source software [8].

A commonly applicable algorithm for computing maximum likelihood word from incomplete data available at various level of generality. Many EM examples are observed like missing value situations, applications to, censored or truncated data, finite mixture models, variance component estimation, hyper parameter estimation, iteratively reweighted least squares and factor analysis[9].

The EM algorithm forecast the parameter of model reputedly, from some Initial level. Each Iteration consists of two steps like E- step and M- step [8]. Radford M. Neal and Geoffrey E. Hinton provided a top-level overview of Expectation Maximization, The EM algorithm provide maximum likelihood Approximation for data in which some variable are Unobserved[10].

Masa Aki Sato and shin isshii had developed a new online algorithm for the NGnet which is come from Batch EM algorithm. This survey paper show that batch algorithm is equivalent to online EM algorithm [9].the online EM algorithm considered as stochastic estimated method to find out maximum likelihood words [11].

In the paper “Online EM Algorithm for Latent Data Models”, the authors had reviewed that they had made a special effort to abstract some combinatorial and algebraic properties, and some common data-structural tools that were at the base of those techniques [12]. The Author had proposed a generic online (also called adaptive or recursive Algorithm) version of the Expectation-Maximization (EM) algorithm applicable to latent variable models of independent observations. this approach is more directly connected to the usual EM algorithm and does not depend on merger with

respect to the complete data distribution .This helped them try to present a few recent results in a unifying framework so that they could be better understood and deployed also by non-specialists. In this paper they had surveyed the newest developments in the area of fully dynamic algorithms for Expectation Maximization. Throughout the paper, they had tried to present all the algorithmic methodologies within a single structure by using the mathematical and combinatorial features and the data structural tools that lie at their base [12].

C) LDA

David M. Blei and Andrew Y. Ng had described the. Latent Dirichlet Allocation (LDA) generative probabilistic topic model for collection of Corpora. It is a three level Bayesian model [11]. In this paper author report results in document modeling, text classification, and collaborative filtering [13].

David M. Blei and Andrew Y. Ng Proposed a generative model for text and collection of discrete data that improves several previous models like byes /unigram, mixture of unigram, and Hofmann's aspect model which is known as Probabilistic latent semantic indexing[14].

Yee Whye , David Newman and Max Welling proposed the collapsed variation Bayesian inference algorithm for LDA, which shows that it is computationally efficient, and easy to implement and significantly more accurate than standard variation Bayesian inference for LDA .The large scale of applications current inference procedures such as variation Bays and Gibbs sampling had been found shortage. Latent Dirichlet allocation (LDA) is a Bayesian network which is more popular than other algorithm [15].

To overcome the above problem of Gibbs sampling method Ian Porteous, David Newman, Alexander Idler proposed the new method that is Fast Collapsed Gibbs Sampling method for LDA and which is widely used LDA model. This new method results in speedups on real world text corpora [16].

Collapsed Gibbs sampling method for the widely used latent Dirichlet allocation model was a general probabilistic framework used for modeling sparse vectors for counting data, which includes bags of words for text, bags of features for images, or ratings of items by customers. This new method results in significant speedups on real world text corpora [17].

III. CONCLUSION

Unlike previous online algorithms, Fast Online EM is designed to process infinite documents with infinite vocabulary words for lifelong topic modeling activity. After some experiments the FOME is superior to the state-of-the-art online LDA algorithms in terms of 1) speed2) space 3) accuracy. We can propose an architecture which can perform FOEM parallel multi-core and multi-processor architectures. This architecture will improve the performance of the system.

IV. ACKNOWLEDGMENT

Wish to express my sincere thanks and the deep sense of gratitude to respected guide Mrs. A.S. Vaidya in Department of Computer Science of R.H. Sapat College of Engineering and Research Centre, Nasik for the technical advice, encouragement and constructive criticism which motivated to strive harder for excellence. We also wish acknowledgment to the people who gives support direct or indirectly to the paper writing.

REFERENCES

[1] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in Proc. 25th Conf. Uncertainty Artif. Intell., 2009, pp. 27–34.

[2] J. Zeng, "A topic modeling toolbox using belief propagation," J.Mach. Learn. Res., vol. 13, pp. 2233–2236, 2012.

[3 D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," J. Mach. Learn. Res., vol. 10, pp. 1801–1828, 2009.

[4] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, "A new approach to speeding up topic modeling," arXiv:1204.0170 [cs.LG], 2012.

[5] D. M. Blei, "Introduction to probabilistic topic models," Commun.ACM, vol. 55, pp. 77–84, 2012.

[6] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. M. Mimno,"Evaluation methods for topic models," in Proc. 26th Annu. Int.Conf. Mach. Learning, 2009, pp. 139–146.

[7] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011.

[8] X. Liu, J. Zeng, X. Yang, J. Yan, and Q. Yang, "Scalable parallel EM algorithms for latent Dirichlet allocation in multi-core systems," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 669–679.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statist. Soc. Ser. B, vol. 39, pp. 1–38, 1977.

[10] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," vol. 89, pp. 355– 368, 1998

[11] M. Sato and S. Ishii, "On-line EM algorithm for the normalized gaussian network," Neural Comput., vol. 12, pp. 407–432, 2000

[12] O. Capp_e and E. Moulines, "Online expectation-maximization algorithm for latent data models," J. Royal Statist. Soc., Ser. B, vol. 71, no. 3, pp. 593–613, 2009.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[14] J. Zeng, W. K. Cheung, and J. Liu, "Learning topic models by belief propagation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 5, pp. 1121–1134, May 2013.

[15]Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," in Proc. Annu. Conf. Neural Inf. Process. Syst., 2007, pp. 1353–1360.

[16] A. Ahmed, M. Aly, J. Gonzalez, S. Narayana murthy, and A. Smola, "Scalable inference in latent variable models," in Proc. 5th ACM Int. Conf. Web Search Data Mining, 2012, pp. 123–132.

[17] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed Gibbs sampling for latent Dirichlet allocation," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 569–577.