International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

Distributed and Parallel Expectation Maximization for Big Topic Modeling

Disha Shinde[1], N. V. Alone[2]

[1]Student, ME (CSE), [2]Professor,Gokhale Education Society R. H. Sapat College of Engineering Management Studies and Research, Savitribai Phule Pune University, Nashik, Maharashtra, India.

***Abstract*— As the amount of information increases, the demand for document classification and maintenance is also increases. In today's world Online documentation is rapidly increasing, so the demand for document classification is also increases ,that's why data management and its analysis is also increasing.As the information is important the documentation of the classes of the knowing form is costly. The objective of the system is to classify the documents. EM improves accuracy while using semi-supervised approach in data mining environment. semi-supervised approach is more accurate and effective than any other technique. The advantage of semi supervised approach is "Dynamically New Class Generation*". It also shows that how to improve accuracy. Expectation maximization (EM) algorithm is used by using supervised as well as semi-supervised technique. In the field of probabilistic modeling LDA and various other topic model are used As the data set today becomes ever more vast, there is a pressing need for efficiently parallel these inference algorithms in multi-core and distributed environment. In parallel and distributed manner data is partitioned among different Processor and presumption is considered.**

***Keywords*- Big Model, EM, Big Data, Topic Modeling, LDA**

I. Introduction

Big data is a term for data sets there is so large or complex collection that traditional data processing application software is inadequately deal with big data. Some challenges in front of big data is to capture, search, analysis, querying, transfer, visualization, updating and information privacy. The data mining technique is nothing but the extraction of fruitful knowledge from large amount of corpus. Different data mining tools provide solutions to the problem related to the business, when problem is solved manually then more processing time is required. The important aspects of data mining and a predictive topic modeling is classification. Expectation Maximization is a bunch of iterative algorithm for maximum likelihood words and maximum posterior estimation with unlabeled data. The documents are growing rapidly in size having importance. Generally 90 percent Words of the data which is held by unstructured formats are as follows- : HTML pages, Email, Technical field documents, Business documents, E-Books and Digital stores, Customer care book.

In a Probabilistic topic models having hidden data structure in large collection of documents Used to discover hidden thematic or data structure by different algorithm. The collection of data ids stored and Computerized in the form of multimedia data. The demand for new computational tools is to aid and organize, also to understand the huge amount of information. Here some main tools are used,1) search 2) link. In search engine one must type keywords to find set of documents which are related to them or find out maximum likelihood words [1]. LDA is growing probabilistic model used for collections of diverse data. LDA is a three-tier hierarchical architecture of Bayesian model, in that every item from the collection each item of a collection is moulded as a finite mixture on the large set of data [2]. here, use the language of text collections throughout the paper, referring to entities such as word, documents, and collection. A word is the basic unit of data or information,It represented as vocabulary indexed by (1......N) A document is a sequence of N words represented by w = (wa,wb......wn), where wn is the nth word in the sequence. A collection of M documents represented by D = (fwa,wb......wm)[2].

A. Expectation Maximization

Data mining tools can provide solution to the business problems that were to too time consuming when done manually. EM is a class of iterative algorithm for maximum likelihood and maximum posterior estimation in problem with unlabeled data.The basic Idea of Online Expectation maximization algorithm is to partition of Data stream of D documents into small mini batches with the size Ds, OEM combines Incremental Expectation Maximization (IEM). Xiaosheng Liu and Jia Zeng had studied to handle web-based content to analysis only on one PC [3]. The multi-core PEM algorithms to denote and Forecast LDA parameters in shared memory systems to avoid memory collision and the time required for locking is solved by multicore parallel EM algorithm which denote as well as forecast its parameters for memory sharing system.

Parallel LDA toolkit is now openly available[4]. A commonly relevant algorithm for computing most similar word from partial data available at numerous level of generality. Most EM examples are observed like 1)lost value situations 2)applications 3)truncated data 4)finite structure models[5]. 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 [4].It is simple to implement and contains two steps: E-Step: - This step includes expectation over conditional distribution of the latent data given the observation. M-Step: - This step includes an analogous to complete data weighted maximum likelihood estimation[5] .

B. Literature Survey

Latent Dirichlet allocation is a three-level hierarchical Bayesian model (HBM) that can denote probabilistic word clusters called topics from the document word structure.LDA has no exact inference methods .These two methods are frequently used: Variational Bayes (VB) and collapsed Gibbs sampling (GS) have been two commonly-used approximate inference methods for learning LDA and its extensions, 1) author-topic models (ATM) 2) relational topic models (RTM) [6]. Two distributed algorithm is mostly used in topic model 1) Latent Dirichlet Allocation model 2) Hierarchical Dirichet Process model. In distributed algorithms data is divided among various processors and assumption is done on parallel and distributed manner [7]. Amr Ahmed et al.presented the Profitable parallel framework used for efficiency in the latent variable models over web-based data.This framework cover 3 main challenges: 1) Incorporate the global data. 2) It efficiently stores and retrieve the data from large set of local data. 3) Sequentially incorporating streaming data [8]. The (batch) EM algorithm plays an important role in unsupervised approach, but it sometimes Undergoes from slow convergence. Here show that online variants provide significant speed ups and can even find better solutions than those found by batch EM. EM support four unsupervised tasks: 1) part-of-speech tagging 2) document classification of documents 3) word splitting 4) word union. Here shows that the two flavor of algorithm: online EM incremental EM and stepwise EM, both algorithm contains updating parameter after each iteration. Online algorithms have the capacity to upgrade speed of a learning by making updates more often [9]. Collapsed Gibbs sampling is a well known method of LDA. This method increases the speed of execution in real world dynamic data.

Conventional Gibbs is another technique which contains o(N) operation per sample data and N is the refers to number of topics in the model [10]. The past ten years has seen fast development of latent Dirichlet allocation (LDA)for solving topic modelling problems because it has solution that is three-layer graphical representation as well as two efficient approximate inference methods1)Variational Bayes 2) collapsed Gibbs Sampling [11]. Statistical topic modeling is an growingly useful tool for analyzing large unstructured text or data collections. There is a significant work introducing and developing stagy topic models and their applications. For some applications there may be not internal tasks, such as information retrieval or document classification, In that performance can be evaluated with the help of universal method that measures the platitude capability of a topic model in a way that is correct, computationally capability client, and independent of any decided application[12]. Stochastic variational inference (SVI) used to speed up Bayesian computation to grand or large data. It applies on many types of model like 1) probabilistic factorization 2) statistical network analysis, 3) Gaussian processes. SVI uses stochastic optimization to fit a variational distribution, which is repeatedly sub sampling from the large data set [13].

EM algorithm with Semi Supervised Technique: In this Distributed and parallel EM system is used a semi supervised approach for improvement of accuracy and to reduce the manually process for classifying the document. EM is applied with supervised approach, but the disadvantage of this system is that the whole data required in the labeled format and no dynamic class generation is there and this technique require a lot of time and speed is also degrades its performance. In semi supervised approach it handles both type of data i.e., both labeled and Unlabeled data. First execution is performed on labeled data give it to training data set and after that unlabeled data is classified. Submitted document is categorized in some predefined classes, If any document may get failed then automatically New class generation is done and updated in document list [14].Supervised and Semi-Supervised Technique of EM and comparison between them is as follow: The disadvantages of EM algorithm in supervised technique is that is that, firstly labeling is provided to each data it requires efforts and time and unlabeled data will be useless. In semi supervised approach unlabeled data is useful because it has been used to train the classifier and if in case the document may get failed then predefine classes will update and dynamically new class is generate and automatically generate a updated list in the database [14].

Usually batch LDA algorithms need repeated scanning of the whole structure of data. To process grand corpora having a large number of topics, the training iteration of batch LDA algorithms is often unfit and time-consuming. To speed-up the training speed, active belief propagation (ABP) not passively scans the subset of data collection and searches the subset of topic space for topic modeling. Accuracy is maintained in each iteration to save large training time [15]. In big data, machine learning plays an important role. This Machine learning technique usually divided into two main types.1) predictive or supervised learning approach.2) descriptive or unsupervised learning approach, 3) reinforcement learning, which requires very few time. This technique is useful for learning how to behave in particular situation [16].Fast Online Expectation Maximization For Big Topic Modeling in this review paper contains existing system techniques which has some limitations and in the proposed system .Fast Online EM has main aim to process unlimited documents with unlimited dataset words for lifelong topic modeling activities. We can propose an architecture which can deal with multi core and multi-processor architectures. This architecture will improve the performance of the system [17].

II. SYSTEM ARCHITECTURE

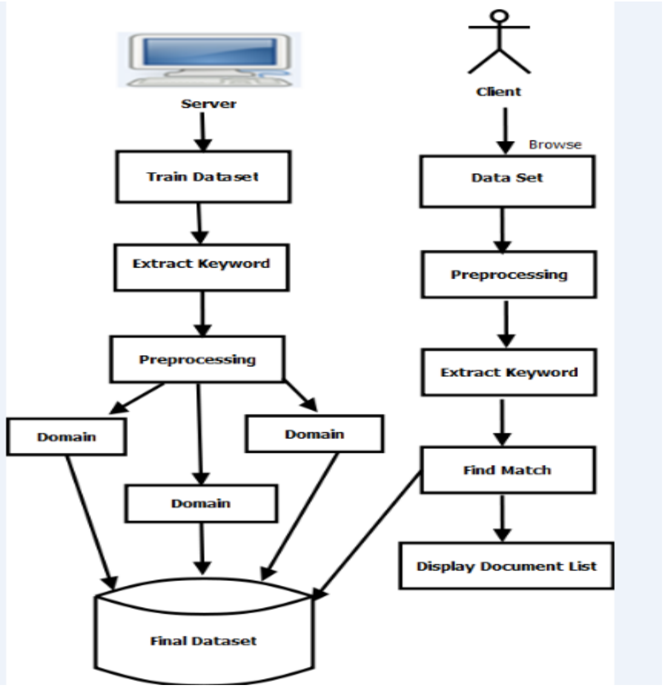


Fig.1. Distributed and Parallel Expectation Maximization for Big Topic Modeling

The contribution works on distributed system where the different domain will be stored in different systems. This means that each domain related keywords will be stored in different systems domain-wise. By implementing this approach, we can have faster retrieval of document as compared to single system approach.

In this proposed system summary can also find out with the help of supervised and unsupervised learning. Unlike previous online algorithms, Fast Online EM(FOEM) is designed to process infinite documents with infinite vocabulary words for some lifelong topic modeling tasks. In Proposed System may extend and deploy FOME on the parallel multi-core and multi-processor architectures for industrial big topic modeling tasks. The contribution works on distributed system where the different domain will be stored in different systems. This means that each domain related keywords will be stored in different systems domain-wise. By implementing this approach, we can have faster retrieval of document as compared to single system approach

A. EM Algorithm

1. Steps for preprocessing of the propose system: In preprocessing it eliminate periods, white space, commas, punctuation mark, stop words and in the preprocessing collect all words on the dataset which has occurrence frequency is more than one time.

2. Shows the frequent words as word sets by matching the words which are in the dataset as well as training dataset documents.

3. Search for maximum likelihood word set or its subset in the list of word sets collected from training dataset with that of subset of frequently occurred word set of new document.

4. Collect the compatible probability values of likelihood word set for each target data.

5. Calculate the probability of the maximum likelihood words

6. Calculate the probability of word with the help of EM algorithm.ie, with the help of E- step and M-step.

7. Collect the document in the Domain which having maximum probability.

B .Dataset

Name Of Dataset	Size Of dataset	URL of Datase
PUBMED	67.8 MB	https://catalog.data.gov/dataset/pubmed

In this Propose System used PUBMED data set is used as a input to the system to find out maximum likelihood words, Which is related to Medical field. In this dataset contain Various Types of Keyword Related to Health Care like human name, disease name, Patient details etc. This dataset require near about 67.8 MB space.

III. Analysis Of System

A. Mathematical Model

1. Problem Description

Let S be a Parallel and distributed EM

Such that

S= {S0,S6,Fs,D,K,F,M,R,C | ∅s}

Where,

D Represents set of d Documents

K Represents Keywords

F Represents set of Frequency

M Represents domain;

R Represents records

C Represents Set of Category

2. Initial State S0: User asks to browse dataset

3. End State S6: User gets Result In the form of Document.

Input: Dataset

Output: Document

State Diagram

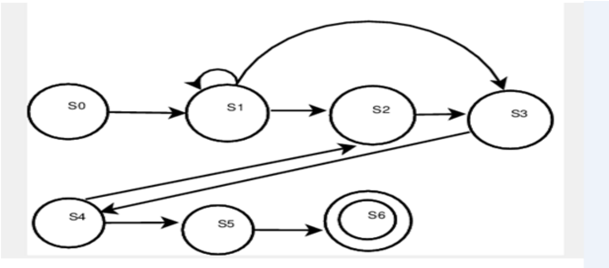


Fig 2.State Diagram

IV. Result And Discussion

System Result: In this dataset all the keyword and word count is visible in descending order, PUBMED data set is related with medical field. This result is parsing of data base.

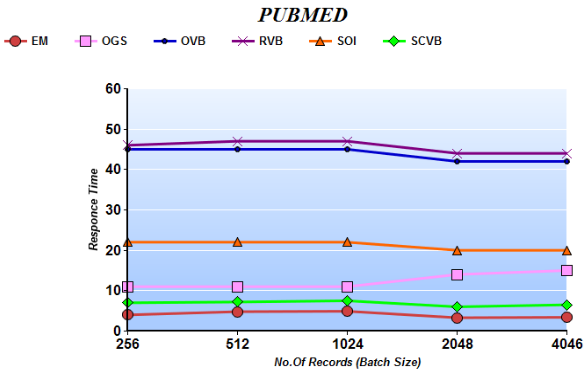


Fig. 3 Results of Response Time.

Fig.3 shows comparison between EM and other LDA algorithm. Expectation Maximization algorithm deal with infinite documents with less Response Time.

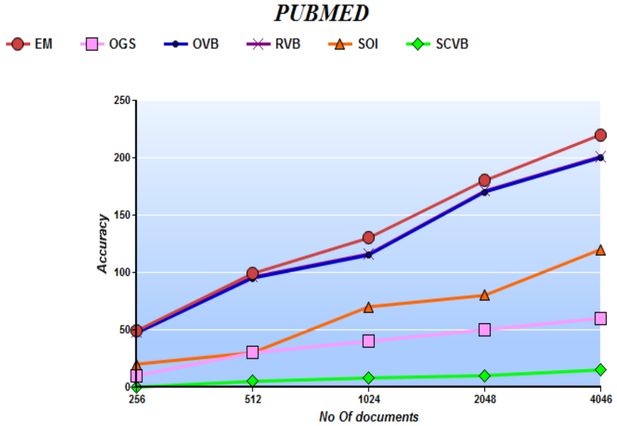


Fig 4. Result of Accuracy

In fig.4 shows that EM algorithms accuracy of result is higher than other LDA algorithms.

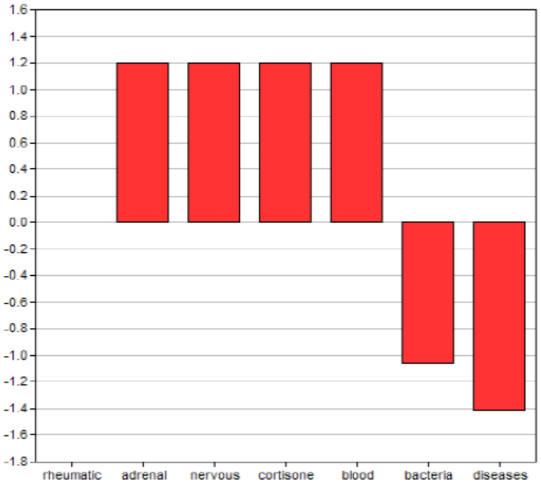
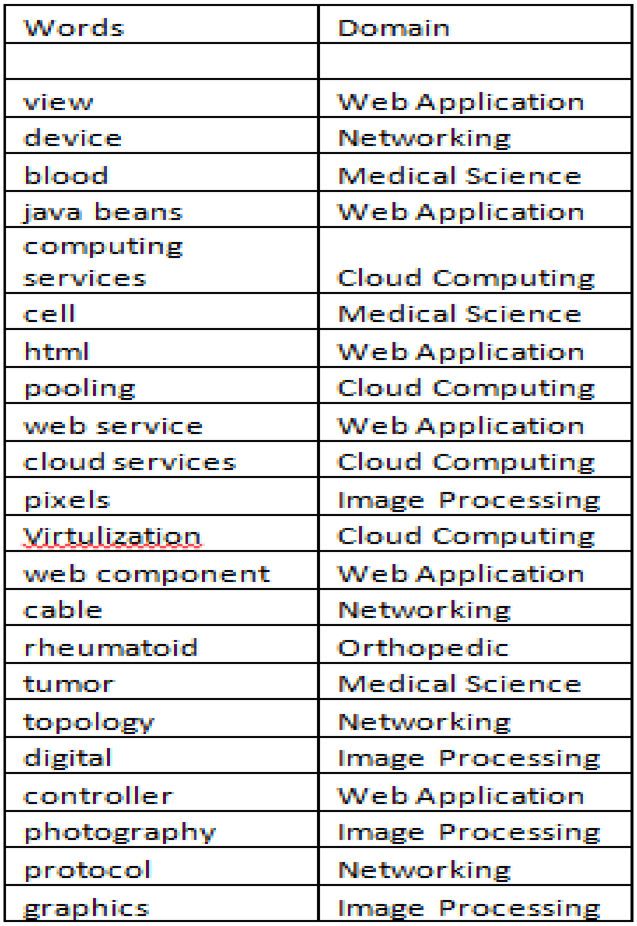


Fig.4. Graph showing weights of words from PUBMED dataset

In this dataset all the keyword and word count is visible in descending order, PUBMED data set is related with medical field. In this graph the weight of present in PUBMED dataset is represented. X-axis represents weight and Y-axis represents words. Words like rheumatic gives a weightage of 0.0 while other words like adrenal, nervous, cortisone, blood etc. shows weight of 1.204.



Words	Domain
view	Web Application
device	Networking
blood	Medical Science
java beans	Web Application
computing services	Cloud Computing
cell	Medical Science
html	Web Application
pooling	Cloud Computing
web service	Web Application
cloud services	Cloud Computing
pixels	Image Processing
Virtulization	Cloud Computing
web component	Web Application
cable	Networking
rheumatoid	Orthopedic
tumor	Medical Science
topology	Networking
digital	Image Processing
controller	Web Application
photography	Image Processing
protocol	Networking
graphics	Image Processing

Fig.5. Dataset keywords and domain

Above table shows the words that occur in PUBMED dataset and the domain which fits into that particular keyword.

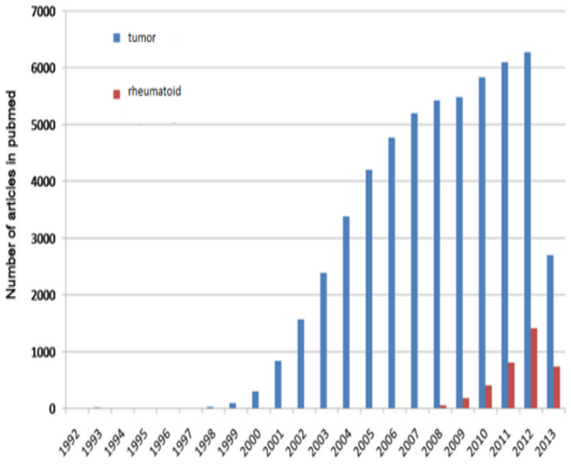


Fig.6. Graph showing the results of words and their frequency occurring in article.

A PubMed search with either the key word "tumor" or "rheumatoid" showed that the reference of these words has increased tremendously in the last decade. The results show the number of words that occur in PUBMED dataset along with its domain. Domain categorization becomes helpful because it makes us search documents from that particular domain only rather than searching the whole data. This improves the response time and thus the data can be retrieved faster.

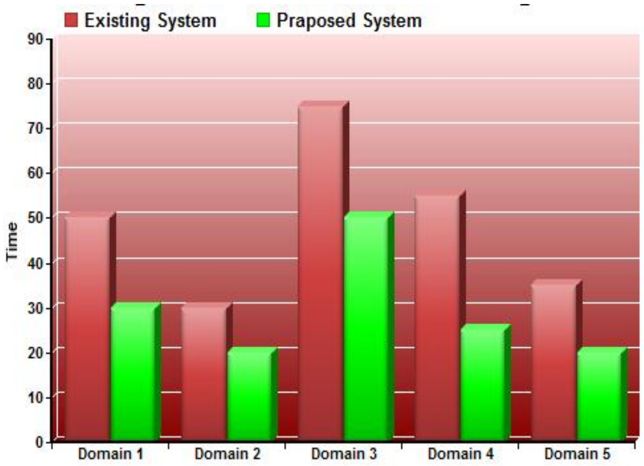


Fig.7 Shows Comparison Between Existing System And Proposed System.

V. CONCLUSION

In this approach here, developed a topic modeling architecture which works on client – server architecture where the particular file requested by client is processed by server by counting its number of keywords and bringing it to appropriate domain by the weights obtained for the keyword. This approach works on distributed system thus making the process faster. The EM algorithm is used for retrieval of most likelihood words from the large collection of dataset. In this system PEM algorithm is used for Multiprocessing environment and this algorithm is used for remove memory conflicts and system locking time. This parallel expectation-maximization is much more scalable than other LDA algorithm which is more accurate and efficient than other algorithm. The LDA toolkit is available on the web as a open source software. The contribution works on distributed system where the different domains are stored in different systems. This means that each domain related keywords are stored in different system systems domain-wise. By implementing this approach, we can have faster retrieval of document as compared to single system approach. Future work may include modifying the existing EM algorithm by combining the Quad Tree approach and the EM algorithm which gives a clustering method that not only fits the data better in the clusters but also tries to make them compact and more meaningful.

Acknowledgment

Wish to express my sincere thanks and the deep sense of gratitude to respected guide Prof. Mr. Nilesh V. Alone in Department of Computer Science of R.H. Sapat College of Engineering and Research Centre, Nasik for the technical advice, impetus 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] D. M. Blei, Introduction to probabilistic topic models, Commun. ACM, vol. 55, pp. 7784, 2012.

[2] 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. 11211134, May 2013.

[3] 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.

[4] 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. 669679.

[5] 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.

[6] N. de Freitas and K. Barnard, Bayesian latent semantic analysis of multimedia databases, Univ. Brit. Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2001-15, 2001.

[7] D. Newman, A. Asuncion, P. Smyth, and M. Welling, Distributed algorithms for topic models, J. Mach. Learn. Res., vol. 10, pp. 18011828, 2009.

[8] A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. Smola, Scalable inference in latent variable models, in Proc. 5th ACM Int. Conf. Web Search Data Mining, 2012, pp. 123132.

[9] 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. 669679.

[10] 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. 569577.

[11] J. Zeng, X.-Q. Cao, and Z.-Q. Liu, Residual belief propagation for topic modeling, in Proc. 8th Int. Conf. Adv. Data Mining Appl., 2012, pp.739752.

[12] 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. 139146.

[13] Bhawna Nigam et.al,"Document Classification Using Expectation Maximization with Semi Supervised Learning,"International Journal on Soft Computing (IJSC) Vol.2, No.4, November 2011.

[14] S. Mandt and D. M. Blei, Smoothed gradients for stochastic variational inference, in Proc. Int. Conf. Mach. Learning, 2014, pp. 2438 2446.

[15] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, Online belief propagation for topic modeling, arXiv:1210.2179 [cs.LG], 2012.

[16] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

[17] Disha R. Shinde ,, 2 A.S. Vaidya,"A SURVEY: FAST ONLINE EXPECTATION MAXIMIZATION FOR BIG TOPIC MODELLING",Volume 5,Issue 4- 2016.

Author's Profile

Disha R. Shinde,ME Second year student,Department of Computer Engineering Gokhale Education Society's, R.H.Sapat College Of Engineering Management Studies and Research, Nashik (dishashinde06@gmail.com)

Prof. Mr. Nilesh Alone, Department of Computer Engineering Gokhale Education Society's, R.H.Sapat College Of Engineering Management Studies and Research, Nashik (nilesh.alone@gmail.com)