BVG Trust

*J. Sci. Trans. Environ. Technov.* 2018, 12(1) : 12-16

Scientific Transactions in Environment and Technovation

# Business analysis and intelligence powered by Big Data Topic Modeling with Mahout

https://doi.org/10.56343/STET.116.012.001.003

http://stetjournals.com

**R.Anitha*, K.Savima and T.Tamilselvi**

Department of Computer Science, S.T.E.T. Women's College, Sundarakkottai, Mannargudi - 614 016, Thiruvarur (DT), Tamilnadu, India.

## Abstract

Topic modeling is a form of text mining, a way of identifying patterns in a corpus. When a corpus is run through a tool that groups words across the corpus into 'topics'. Topic modeling for big data provides a key opportunity to address the needs of data-driven businesses in a way to deliver genuine value to business users by simplifying search and summary processes via the vast amount of information. The approach captures the evolution of topics in a sequentially organized corpus of documents into two main phases, mapping and reducing phases. In the mapping phase the probabilistic on each word, in collected documents, is calculated by using collapsed space of latent variables and parameters for summarizing words in each topic, and reducing phase to utilize the various results from map phase while predicting a new topic model from a given trained models.

**Keywords:** Big Data, Topic modeling, Mahout, Business Intelligence

Received : April 2017 Revised and Accepted : July 2018

## INTRODUCTION

Topic modeling aimed at automatically extracting and analyzing the knowledge from unknown data of the many collected documents for the synthesis of information. Blei *et al*. (2003) they described the topic models by using a hierarchical Bayesian to discover connected topics, trends within topics and primary semantic patterns from large datasets. The given probabilistic Steyvers and Griffiths ,2005 ; Michael *et al*., 2010) is used to assign the value on the distribution of words from several related topics for generated model for making a new document based on different distributions over topics. The trend of Topic models Steyvers *et al*., (2005) is applied to many types of analysis of documents such as extract words from email, check the similarity patterns of scientific abstracts and related contents from newspaper for discovering patterns of word and linking documents that exhibit similar patterns. Topic models have emerged as a powerful new technique for finding useful structure in an otherwise unstructured collection. Many new researches are emerged on technologies related with topic models. Blei *et al*. (2010) proposed the survey result of the probabilistic topic models with suitable algorithms to solve the problem of managing collected documents. The topic model as applied for generating a model from text documents and images in each object which contains a set of instances by using the methodology of a confidence-constrained rank minimization (CRM) with recovery probability Behmardi and Raich (2012). Heli *et al*. (2014) applied user-topic model from original and rewet interest on micro blogs by using Gibbs sampling for the inference of the parameters and Negoesscu and Gaticaperex, (2010) proposed an implementation on Flicker that has more than 30 million users and over 3 billion photos and public tags with probabilistic topic model by learning unsupervised discovery of similar users and group from tag-based strategies. The Latent Dirichlet allocation (LDA) is the most popular topic model that uses a conjugate prior to the multinomial and the Dirichlet distribution is a convenient choice for simplifying the problem of statistical inference Blei *et al*., (2003) ; Michael *et al*., (2010) ; Blei *et al*., (2010) and Heli *et al*., (2014). In Jing Jiang, (2009) proposed a hybrid model on Hidden Markov Models (HMMs) within LDA topics for joining on topics and syntactic structures for each topic. The multi-document summarization proposed by Bian Jinqiang *et al*. (2014), utilized the topic distribution of each sentence with topic based on the corpus for calculating the posterior probability of the sentences. The online topic model (OLDA) was discovered Alsumait *et al*. (2008) which has interesting patterns for analyzing a fraction of data at a time. The Collapsed Variational Bayesian (CVB) Inference for LDA is a new inference algorithm that utilized the advantages from variational Bays and collapsed Gibbs sampling. Nakano *et al*. (2011) developed a CVB nonparametric latent source discovery method for music signal analysis. James Foulds *et al*. (2013) proposed a stochastic algorithm

*Corresponding Author :
email: *anithaaocsqa@gmail.com*

P - ISSN 0973 - 9157
E - ISSN 2393 - 9249

for collapsed variational Bayesian inference for LDA to improve the efficiency and computation time when compared to the previous methods. Furthermore, for processing the large volume of datasets Hadoop is a large-scale distributed batch processing infrastructure. While it can be used on a single machine, its true power lies in its ability to scale hundreds or thousands of computers, each with several processor cores. However the architecture is also designed to improve the distribution of large amounts of work efficiently across a set of machines and the data are distributed to all the nodes of the cluster as it is being loaded in. Wichian Premchaiswadi and Walisa Romsaiyud (2013) proposed a model tuning and optimizing the parameters of Hadoop Map Reduce for reducing the computation time. They applied a collapsed variation Bayesian (CVB) inference algorithm for LDA to generate a new topic models on Hadoop Map Reduce framework to improve the performance on searching and reducing the response time.

**Background**

This section describes three important techniques that are closely related to our work. They are the Hadoop Distributed File System (HDFS), Machine Learning with Apache Mahout and Collapsed Variation Bayesian Inference for LDA.

**The Hadoop Distributed File System (HDFS)**

Hadoop (Jason venner *et al*., 2009 ; Tom White, 2012 ; Srinath Perera and Thilina gunarathne, 2013) comes with a distributed file System called Hadoop Distributed File System (HDFS). Hadoop enables distributed parallel processing of large volume of datasets and splits them into smaller pieces for processing across clusters of commodity servers that provide scalable and reliable data storage and the detailed account (Fig.1)

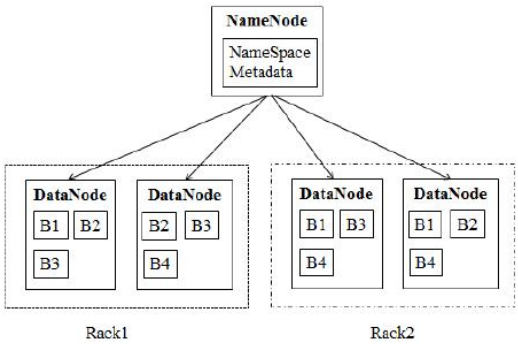


**Fig.1.** Shows the HDFS clustering Name Node

This HDFS cluster consists of a Name Node which manages the file system namespace and regulates access to files by clients and Data Nodes that store the distributed file system for both reading and writing. HDFS splits the large input data set into smaller data blocks and the default block size equal 64 MB., and replicate blocks to other nodes for improving the performance of the fault-tolerance (such as replicate the B1 file on Data Node both Rack1 and Rack2). A main objective of such system is to improve the performance of complex data analytical tasks by confirming the potential of their approach. Map Reduce (Tom White, 2012 ; Srinath Perera and Thilina gunarathne, 2013) is a programming model for data processing that can run on Hadoop. A Map Reduce job splits a large data set into independent chunks and organizes them into key, value pairs for parallel processing. The following figure illustrates that Map Reduce programs are executed in two main phases, mapping and reducing phases. The Map phase divides the input into ranges based on the input format and creates a map task for each range in the input. The Reduce phase collects the various results from Map phase and combines at a final result. This parallel processing improves the speed and reliability of the cluster, returning solutions more quickly and with more reliability.

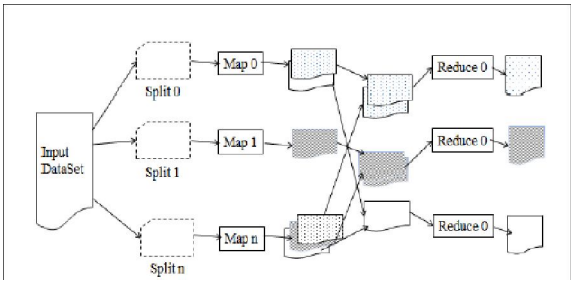


**Fig.2.** Show the Data set Mapping

**Machine Learning with Apache Mahout**

Apache Mahout Tom White, (2012) is a library of scalable machine learning algorithms Which are implemented on top of Hadoop and uses the Map Reduce paradigm. Mahout can be applied for solving problems of machine learning covered with supervised and unsupervised learning approaches. Example techniques include; collaborative filtering, clustering, classification and frequent item set mining. Several researches applied the Mahout for improving the performance, For example: RPig framework was proposed for a scalable advanced data analysis solution with Mahout for machine learning and statistical analysis on the Telecom domain (Mingxue Wang, 2012). Esteves, (2011), evaluate the experiment by running KMeans on a Hadoop system and testing

on Amazon EC2 Cloud instances for gaining in running on a multi node cluster and then Hammond and Varode (2013), developed prototype tools for studying the context of text classification, recommender systems and decision support on Cloud. The performance of k-means was improved and fuzzy c-means for clustering a noisy realistic and big dataset on Cloud (Esteves and Chunming Rong, 2011). This paper focuses on a generating a dynamic topic model for large volume of dataset stored in data file system (dfs) based on batch oriented paradigm shift and applied the collapsed variation Bayesian inference for LDA to discover and integrate a topic modeling for big data and business analytics. However, this paper uses a Mahout-distribution version 0.7 that covered the basically machine learning techniques, and using the Apache Ant version 1.9.4 for building files as targets and extension points dependent.

**The architecture overview and proposed method**

This section describes major procedures employed in each step of data processing so as to clearly identify the architecture by proposing a method to assign the related documents from many topics modeled on Hadoop.

***The architecture of the proposed method***

The system architecture comprises five main parts; collecting the different data sources, splitting the large file size to smaller size that using the line format for each file pattern, assigning the word frequent on each topic, grouping the set of related word to represent a related topics and generating the related document from weight. More details are shown in the following figure.

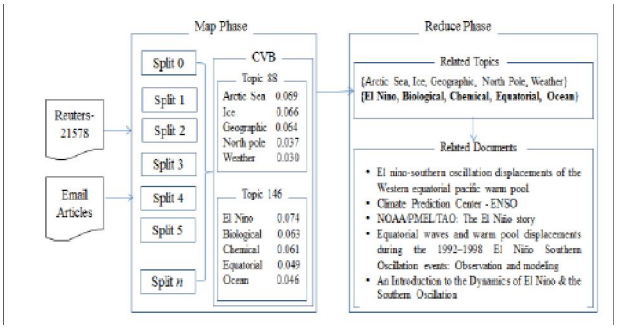


**Fig.3.** Shows the Reuters-21578 dataset

Data Source that comes from the Reuters-21578 Text categorization collection dataset and Email articles text corpus. Data split provides the HDFS splits the large input data set into smaller data blocks in 64 MB. with default in the Map phase; Collapsed Variation Bayesian (CVB) inference algorithm for finding the word frequency in each topic. In the Reduce phase, grouping the set of words that are related with topics. Final Model for predicting the new topic model for text classification to estimate a multinomial Naïve Bayes classifier and to apply the classifier to the test documents.

***The Extraction Topic on the Predictive Model***

Generating a new topic model is required for mixing the data set across the cluster network. The feature of this capability is the dynamic words, topic and documents from probabilistic scores. However, the data were analyzed and processed into a predictive model come from the Reuters- 21578 Text categorization collection dataset (Blei *et al*., 2003 ; Steyers and Griffiths, 2005) and Email articles text corpus used in the experiments are shown in Pseudo code I.

**Pseudo Code**

1: ext-CVB for generating a topic modeling Input: The different data sources Set: Nm: Number of Map Task job; Nsem: Number of Sentence in Document; T: The number of topics; D: Document; W: Word in a document; Z: The distribution over topics; Tnew: The new topics; C: Class; 1. Read the data files into the system;

2. In the Map Phase;

3. int length = IO.readInt();

4. String sentence = IO.readString();

5. int full = sentence.length();

6. int wordcount = 0;

7. for(int i = 0; i == length; i++)

8.if(Nsen =Character.isLetter(sentence.charAt(i)))

9. wordcount= wordcount + 1;

10. Assign the Nm = Nsen; Defines the number of sentence equal with the number mapped in each slot.

11. Calculate word count;

12. Split the corpus D into {D0, …., Dn}, corresponding from distribute of Nodes.QTopics {T0,…, Tn} and counts

13. Calculated the probabilistic of word under topic that distribution over words within a document.

14. P (wi) = $\sum p\ (wi\ |zi = j)\ P\ (zi = j)$; $T\ j$=1 Each word wi in a document where the index refers to the ith word token; Tnew⇓

15. Draw a new topic

16. Count of tuple (k*, v) +1); For the key pair and value.

17. In the Reduce Phase;

18. Combine (k*, v) +1); combine the key and value from the Map phase.

19. Sort tasks, sub corpus Dp

20. Merge changes of the global count

21. { key(k,v); +1} by key (k,v), output tuples (k,v); list;

22. Modify the new count of tuple (k*, v); the number of list ;ßkv F

23. For key(k,v),

24. Generate the class with classification rule $\log 1 \leq k \leq nd(tk \mid c)$]; find the$\sum$

25. Cmap = avgmax [logP(c) + weight for displaying the relative frequent of class C

Pseudo code I illustrated in this paper is divided into 2 phases, Map and Reduce phases. For the Map phase, the collected data from data files system that have the large volume of data sets are split into a smaller data files by using the end of sentences and store each block in a map slot. The Collapsed Variation Bayesian (CVB) inference algorithm calculates a word in document where the index refers to the word token. The Reduce phase, merge the key and value from a map phase that refer to a set of word in the topic and sort the word frequency. The classification rule with Bayes is used for generating the new topic is related to all documents.

**Experimental Evaluation**

For the evaluation process, the data files in 3 scenarios; 1) ext-CVB1: Reuters-21578 and following with Email articles, 2) ext-CVB2: Email articles and Reuters-21578, 3) ext-CVB3: Random from both of 2 data files types in a batch processing. However, the rationale idea behind the experiment was to mix the collection of data for studying the coherent topics across multiple data sets. Running the experiment, from ext-CVB1, 2 and 3 in order to revise those parameters and latent variables were well coupled. However, the experiment displays the result from mixing or sequence of data sets. The parameters in each classification group are not directly affected with the latent variables.

**Conclusion and future work**

The main objective of the paper is to generate the new topic model on top of word probabilistic on each topic for retrieving the similarity documents score based on improving accuracy and computation time. The paper proposes a novel method which can generate a new topic model from large volume of data sets for supporting the analysis services on business. The primar objective of this study is to enhance the features of a collapsed variation Bayesian (CVB) inference algorithm for Latent Dirichlet allocation (LDA) on Hadoop. Moreover, a local optimum is guaranteed by our proposed method. The method was executed on Reuters-21578 text categorization collection corpus on Hadoop clustering with 64 nodes, with comparison to a standard CVB as a baseline. The empirical results demonstrate that our method; (1) is able to classify the documents from words frequency by using statistics score and predict the new topic models, (2) is capable of extending the features on CVB in order to collapse space of latent variables and parameters, and (3) can be applied on an Apache Mahout with Hadoop Map Reduce paradigm on 64 nodes clusters to improve the performance in terms of different data file sizes and corpus. However, in future we will study more on the performance of computation time and cost on the different data types from many data sets.

**REFERENCES**

AlSumait, L., Barbara, D. and Domeniconi, C. 2008. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking, *Int'l IEEE Conf on the 8th Data Mining (ICDM 2008),* P. 3-12. https://doi.org/10.1109/ICDM.2008.140tu

Behmardi, B. and Raich, R. 2012. On Confidence-Constrained Rank Recovery. *In : Topic Models, IEEE Trans. On Signal Processing,* 60 : 5146-5162. https://doi.org/10.1109/TSP.2012.2208634

Bian Jinqiang, Jiang Zengru and Chen Qian, 2014. Research on Multidocument Summarization Based on LDA Topic Model. *Int'l IEEE Conf on the 6th Intelligent Human-Machine Systems and Cybernetics (IHMSC 2014),* P. 113-116. PMid:24708959 https://doi.org/10.1109/IHMSC.2014.130

Blei D., Andraw, Y. and Michael Jordan, I. 2003. Latent Dirichlet allocation. In : *I nternational Journal Machine Learning Research,* 3 : 993-1022.

Blei D., Carin L. and Dunson, D., 2010. Probabilistic Topic Models, *IEEE Magazine. On Signal Processing Magazine,* 27 : 55-65. PMid:25104898 PMCid:PMC4122269 https://doi.org/10.1109/MSP.2010.938079

Esteves, R.M. and Chunming Rong, 2011. Using Mahout for Clustering Wikipedia's Latest Articles: A Comparison between K-means and Fuzzy C-means in the Cloud. *Int'l IEEE Conf on Cloud Computing Technology and Science (CloudCom* 2011), P. 565-569. https://doi.org/10.1109/CloudCom.2011.86

Esteves, R.M. 2011. K-means Clustering in the Cloud—A Mahout Test, *Int'l IEEE Conf on Advanced Information Networking and Applications* (WAINA 2011), P. 514-519. PMid:21694451 https://doi.org/10.1109/WAINA.2011.136

Hammond, K. and Varde, A.S. 2013. Cloud Based Predictive Analytics: Text Classification, Recommender Systems and Decision Support, *Int'l IEEE Conf on Data Mining Workshops* (ICDMW 2013), P. 607-612. https://doi.org/10.1109/ICDMW.2013.95

He Li, Jia Yan, Han Weihong and Ding Zhaoyun, 2014. Mining user interest in microblogs with a user-topic model, *IEEE Trans. On Communications,* 11: 131-144. https://doi.org/10.1109/CC.2014.6911095

James Foulds, Levi Boyles, Christopher Dubois, Padhraic Syth and Max Welling, 2013. Stochastic collasped variational Bayesian inference for latent Dirichlet

allocation, *Int'l Conf on 19th Knowledge discovery and data mining (ACM SIGKDD 2013),* P. 446-454. https://doi.org/10.1145/2487575.2487697

Jing Jiang, 2009. Modeling Syntactic Structures of Topics with a Nested HMM-LDA, *Int'l IEEE Conf on the 9th Data Mining (ICDM 2009),* P. 824-829. https://doi.org/10.1109/ICDM.2009.144 PMCid:PMC3207638

Michael W. Berry and Jacob Kogan, 2010. *Text Mining: Applications and Theory*, Johm Wiley & Sons. https://doi.org/10.1002/9780470689646

MingXue Wang, 2012. RPig: A scalable framework for machine learning and advanced statistical functionalities. *Int'l IEEE Conf on the 4th Cloud Computing Technology and Science* (CloudCom 2012), P. 293-300. https://doi.org/10.1109/CloudCom.2012.6427480

Nakano M., Le Roux J., Kameoka H., Nakamura T., Ono N. and Sagayama S., 2011. Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model, *Int'l IEEE Conf on the Applications of Signal Processing to Audio and Acoustics (WASPAA 2011),* P. 325-328. https://doi.org/10.1109/ASPAA.2011.6082324

Negoescu R. and Gatica-Perex D., 2010. Modeling Flickr Communities Through Probabilistic Topic-Based Analysis, *IEEE Trans. On Multimedia,* 12: 399-416. https://doi.org/10.1109/TMM.2010.2050649

Srinath Perera and Thilina Gunarathne, 2013. *Hadoop MapReduce CookBook,* Packt Publishing, 2013.

Steyvers M. and Griffiths T., 2005. Probabilistic topic models: In : *Latent Semantic Analysis A Road to Meaning, Lawrence Erlbaum Associates.*

Steyvers M., Griffiths T., Blei D. and Tenenbaum J.B., 2005. *Integrating topics and syntax, Advances in Neural Information Processing* 17, Cambridge, MA:MIT Press.

Tom White, 2012. Hadoop : *The Definitive Guide*, 3rd ed. O'Reilly Media, United States of America.

Wichian Premchaiswadi and Walisa Romsaiyud, 2013. Optimizing and Tuning MapReduce Jobs to Improve the Large-Scale Data Analysis Process, *International Journal of Intelligent Systems*, 28: 185-200. https://doi.org/10.1002/int.21563