International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Improved Topic Modeling with Parallel-Supervised LDA

Madhurima Mukherjee, Poovammal E

Abstract: ***In the modern era of digitalization, our day-to-day life is entirely dependent on digital platform-from raising our voice in social media to online shopping. Our collective knowledge is continued to be accumulated in the form of electronic texts, blogs, news, images, audios, videos and in many more ways and on account of this there is a greater need of analyzing these huge contents to get rid of the difficulties in searching the object, we aim for. Topic modelling is an efficient machine learning techniques for discovering the hidden semantic structure of contents. "Latent Dirichlet Allocation" (LDA) is a generative probabilistic topic modelling, which is the basis of other generative topic modelling techniques. New models are coming up with advanced algorithms in order to improve the topic modelling. Existing models have their own limitation. In case of obtaining more accuracy, the processing time of topic modelling goes high while in consideration of achieving more speed, accuracy gets low. Most of the algorithms implemented earlier cannot perform well in above-mentioned area. In this paper, we would like to introduce parallel-supervised LDA model where supervised Latent Dirichlet Algorithm (sLDA) and parallel Latent Dirichlet Algorithm (pLDA) are applied together to obtain high accurate results with quicker response time.***

Index Terms: ***Digitalizatio, Latent Dirichlet Allocation, parallel Latent Dirichlet Algorithm, supervised Latent Dirichlet Algorithm***

I. INTRODUCTION

In this era, there is a rising imminent demand for organizing and retrieving useful information for increasingly vast amounts of data. This is really a very sound opportunity for the researchers of many domains including science humanities to research on it.

Data mining is the key, by which large dataset can be analyzed and as per our requirement we can obtain the data patterns using multiple data mining techniques. Data mining offers a number of domains of database systems, machine learning, artificial intelligence and statistics. Through data mining, millions of attributes are analyzed and useful information are obtained. All kinds of databases are nowadays rapidly transformed into digital version. All industries including government data are also now in electronic form. As more information is becoming available, accessing of the particular data, which we are looking for, is getting more complicated. Therefore, powerful tools and techniques are needed for organizing, searching and understanding those data. An automated analysis is required for multimodal data, which are available there in internet. So far, numerous machine-learning algorithms have been employed in handling data. Among them, topic models are very powerful as well as popular because they are capable of discovering the latent structure, which is embedded over documents. Moreover, topic models proffer us low dimensional depiction in the matter of large-scale data.We get to know, how the words are related to each other in a document and based on that how topics are generated. A simple topic-modelling concept has been shown in fig 1, where we can see by applying topic modelling on multiple documents topics are generated.

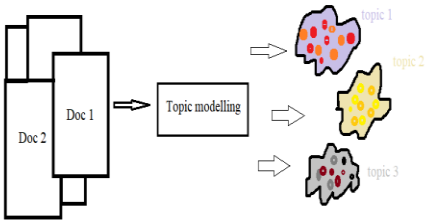

Figure. 1: Topic Modelling

1.1. Variations of topic modelling:

In a natural way, topic modeling is discovering thematic structure in large volume of data and annotating those according to the structure. It finally uses those annotations for visualization, organization, summarization and many purposes. For example, a stack of unorganized books is made to turn into an automatically organized library.

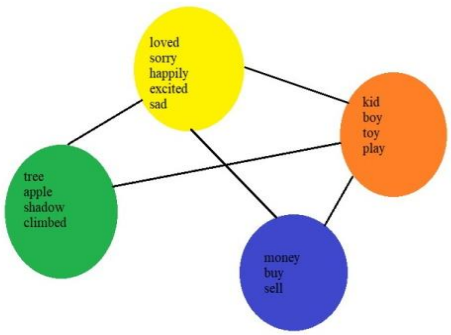

Figure. 2: Network diagram of topic selection

Manuscript published on 30 September 2019
* Correspondence Author
Madhurima Mukherjee*, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur. Email:madhurimamukherjee_as@srmuniv.edu.in
Poovammal E, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur.

A long time ago, there was a huge apple tree. A little boy loved to come and play around it everyday. He climbed to the treetop, ate the apples and took a nap under the shadow. He loved the tree and the tree loved to play with him. Time went by. The little boy had grown up and he now longer played around the tree every day.
One day the boy came back to the tree and he looked sad. 'Come and play with me", the tree asked the boy.

"I am no longer a kid. I do not play around trees any more," the boy replied.
"I want toys. I need money to buy them."

"Sorry.....but I do not have money, but you can pick all my apples and sell them. So, you will have money."

The boy was so excited. He grabbed all the apples on the tree and left happily. The boy never came back after he picked the apples.
The tree was sad.

Figure.3: Sample article

Topic modeling is utilized in various ways, such as through network model, sequential model etc. Some simple examples are described here to get a better understanding of topic modeling. In figure 2 a simple network model is depicted to find out the connection in between topics. The topics are found from first part of a story "The Boy and The Apple Tree" whichis demonstrated in figure 3 . The most frequent words are found from that part of stories and related words are grouped together to form topics. In fig. 2 a connection between those topics are shown.
Not only from texts but also from images, videos and other forms of digitized data annotations can be done automatically through topic modeling. When topic modeling is applied on an image, the algorithm returns a list of words which are related to that given image. Fig. 4 is an image of nature. Suppose, topic modeling algorithm is applied on the image. The output will be a list of words related to the image. For example, the outputs may be sky, hill, river etc.

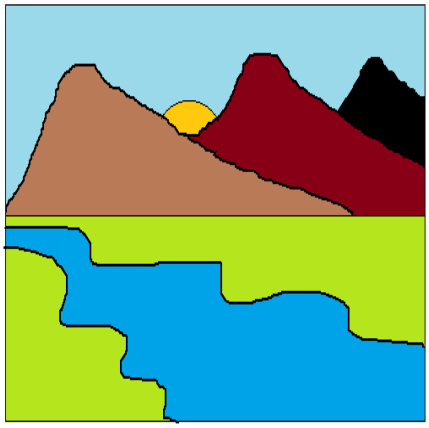

Figure. 4. Example of nature

The topic modeling which are described till now are static. For static topic modeling the input is raw text of a paper or an static image. In dynamic scenarios, topic modeling is used to analyse the evolution of topics over time. Basically, it is assumed that the words are exchangeable and documents are made in group together by time slice. Here, topics are found through events from large collection of data. From millions of data, i.e, telegrams, cable, weather report etc., interesting and significant events are found.

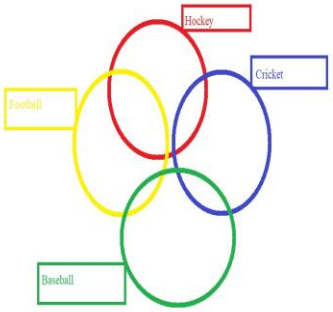

Figure. 5: Mixed membership model

A mixed membership model is shown in fig. 5. Four groups, such as 'Football', 'Cricket', 'Hockey', 'Baseball' are demonstrated in an overlapping manner. Suppose the players who play baseball they also can play football as well as cricket as shown in fig. 5. Likewise, the whole system is modelled. Topic modeling plays a major role in these kinds of systems, such as, a big social media having vast amount of increasing data where more than 3.5 million nodes capture people in overlapping network.
Topic modeling has emerged as a very effective technique in retrieving of information. For biological data retrieval there is no better choice than topic modeling. Capturing of genetic signature of ancestors becomes easier through topic modeling. In Evolutionary biology and genetics topic modeling is very popular. Likewise, in every domain including medical, social media data retrieval, topic modeling has wisely proved its efficiency.

1.2. Growth of topic modelling:

Latent Semantic Indexing (LSI) [1] is the first building block of topic modeling, which is proposed by Deerwester et al. in 1990. It is the earliest model of topic model, shown in fig. 6, where n_{dv}demonstrates a collection of documents,θ_{dk}demonstrates each-document topic weights and β_{kv}
demonstrates each-topic term weights.

- By term matrix of the term frequency-inverse document frequency (TFIDF) scores a collection is treated as a document.
- Choosing a number of topics single value decomposition (SVD) has to be run on the matrix.
- It will return two matrices : each-document topic weights and each-topic term weights.

LSI is the pioneering work that has introduced topic modelling. But it had few limitations, so one by one more improved topic models are kept on coming up with more powerful methods. The first probabilistic topic model was Probabilistic Latent Semantic Analysis (PLSA) [2], which is based on LSI for automated indexing built based on statistical latent class model. Maximum likelihood is estimated in this method. After that in 2003 D.M. Blei and Jordan proposed a novel method of Latent Dirichlet Allocation (LDA) [3] , where the most likely appearing words are found and grouped together for generating topics. It is basis of other topic modelling, where the words are obtained in document and a topic assignment is repeatedly chosen from the topic proportions. Then, a word is drawn from that corresponding topic.

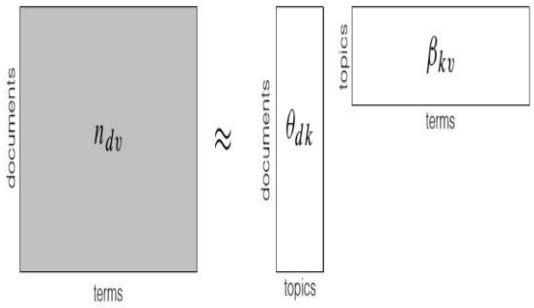

Figure. 6. LSI model

It is the first probabilistic generative topic model [4], which enables numerous applications. Enhancements of the conventional method is very much needed to keep pace with the increasing massive data and handling it in more efficient and accurate way. Therefore, new models are coming up with advanced algorithms in order to improve the topic modelling. Existing models have their own limitation. In case of obtaining more accuracy, the processing time of topic modelling goes high while in consideration of achieving more speed, accuracy gets low. Most of the algorithms implemented earlier cannot perform well in above-mentioned area. In this paper, we would like to introduce parallel-supervised LDA model where supervised Latent DirichletAlgorithm (sLDA) and parallel Latent Dirichlet Algorithm (pLDA) are applied together to obtain high accurate results with quicker response time.

II. RELATED WORKS

If we look overthe topic modelling background, we can find "Latent Semantic Indexing" (LSI) is the origin of topic modelling and afterwards with newly invented methodologies topic modelling techniques are improved.

Deerwester proposed "Latent Dirichlet Indexing" (LSI) [1], which served the basic steps in topic modelling. "Singular Value Decomposer" (SVD) is applied to evaluate the generation of word usage in in documents. In spite of serving the basic steps of topic modelling also, this method lacks in solid probabilistic basis.

Thomas Hoffmann [2] then proposed a powerful technique called "Probabilistic Latent Semantic Analysis" (PLSI). This method is used for supervised learning and it is having a sound statistical basis. This method is been accepted as a powerful and very significant in order to retrieve information and it is holding a wide area of application.

Chou and Chen [5] proposed "Incremental Probabilistic Latent Semantic Indexing" (IPLSI). This algorithm can gain better performance by maximizing the acceptable threshold range. This method is apt in handling temporal relationship among the documents. Event detection quality is been improved by using this sound method.

D.M.Blei and Jordan [3] proposed "Latent Dirichlet Allocation" (LDA) and it is the first generative probabilistic model. Discrete data can be calculated by this method. The limitations of previously invented LSI and PLSA methods have been improved by this novel method. It is an unsupervised technique, where words in the documents are modelled through this. The aim is to infer topics by which likelihood is maximized.

Based on LDA numerous combinations of algorithm have come up with more improved topic modelling. Likewise, Changhua Lin and et al. [6] proposed a method for sentiment analysis based on LDA model and named as "Joint-Sentiment Topic Model" (JST). This method is weakly supervised. Therefore, it is extremely portable. The topics discovered by JST model is more informative and coherent.

X.Wang and et al. [7] proposed a novel model, named Group-Topic model (GT) of entity-relationships. The algorithm is used for the technique is Group Latent Dirichlet Allocation (GLDA). Simultaneously groups and topics are detected from the textual attributes. GT model searches latent group and attribute-clusters, by which communication is enhanced among entities. More improved topics and cohesive groups are discovered through this model.

Chen and Liu proposed [8] "Lifelong Topic Model" (LTM) which is a knowledge-based model. It deals with the possible incorrect knowledge. The authors have shown irrespective of any user input the knowledge can be mined dynamically. LTM is extremely useful in topic discovery by handling the incoherence problem efficiently among topics.

Hanqui Wang and et al. [9] proposed, "identified objective-subjective Latent Dirichlet Allocation" (iosLDA) model, where each document having two unique "Bag-of-words" (BoDW) representation with respect to "subjective" and "objective" senses. And this is moresignificant than the traditional "Ba of Topic" depiction. This algorithm has come up with more improved topic modelling by means of performance and execution intricacies.

III. METHODS

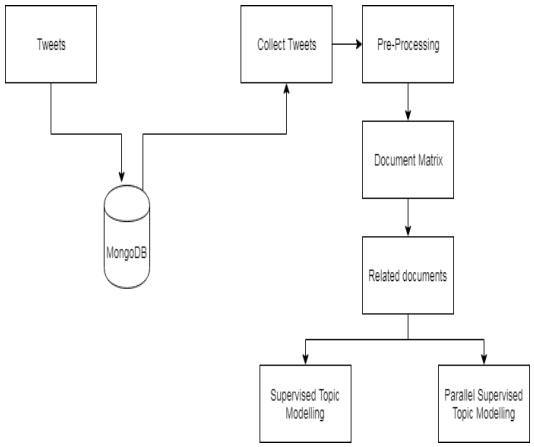

Figure. 7. Block diagram of the proposed model

The methods that are involved in topic modelling are:

- **Known source document collection:** For downloading twitter, data twitter account is configured and access token is generated. Tweets on trending topic are downloaded using Tweepy API. Thus, related documents from internet blogs are collected.
- **Pre-processing:** In order to ensure the output of topic modelling is noiseless, pre-processing steps are accomplished for cleaning the collected data. At first the punctuations and special characters are removed.

Stop words are also removed because these are clutter and do not carry any meaningful information. For removing stop words, tokenization of the text has been done. The string of text are splitted into individualwords or tokens. Stop words are removed and lemmatization is performed. In this way data are cleaned up and pre-processed.

- **Document-Term matrix:** The corpus is the combination of all text documents. The text corpus is converted into a matrix representation for applying mathematical model on it. Repeating term patterns are required to be found in case of LDA model in the whole document-term matrix. “Genism” is one of the efficient python libraries for handling text data.
- **Parallel-supervised LDA:**Firstly, Latent Dirichlet Allocation (LDA) is applied on text data. LDA is basis of other topic modelling, where the words are obtained in document and a topic assignment is repeatedly chosen from the topic proportions. Then, a word is drawn from that corresponding topic. Now, in order to apply supervised Latent Dirichlet Allocation (sLDA) [10] , response variables are added on top of LDA to each collected document associated with it. The responses and the documents are together modelled for finding out the latent topics. These latent topics are capable to obtain the best prediction of the response variables in order to predict for future unlabeled documents. On top of sLDA, parallel Latent Dirichlet Allocation (pLDA) [11]is applied in order to increase the execution speed of the topic modelling. By applying this supervised-parallel LDA, more accurate output can be gained with faster response time.

IV. RESULT AND DISCUSSION

In this paper, we have proposed a new model of topic modelling, which is implemented by following the methods discussed in section 3 of this paper.

Figure. 8. Collected twitter data

From fig. 8 and fig. 9, we can observe that the input dataset are preprocessed. Raw data, which are downloaded from twitter, are shown in figure 8. Likewise, in fig. 9 we can see that the data are preprocessed properly. Strings of texts are tokenized. All special characters, punctuations, stop words are removed from the data. Lemmatization is performed. As a result, we have obtained stream of cleaned words.

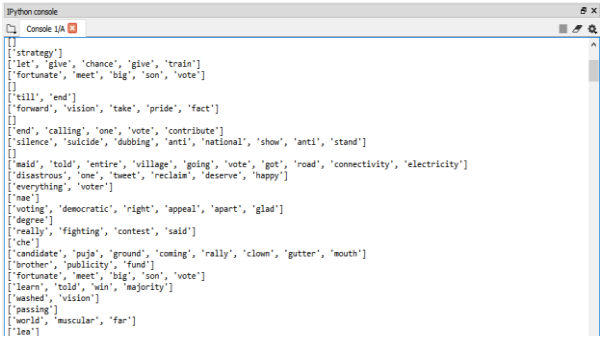

Figure. 9. Preprocessed data

After pre-processing of the data, LDA is applied on the dataset. The output is shown in fig. 10. And fig. 11, We can see that multiple Topics are generated by the related words together.

In fig. 11 we can see the visualization graph of LDA topic modeling, where mostly used words are obtained. But LDA has its own limitation. As, LDA is an unsupervised technique it lacks in categorizing and predicting properly.

Figure. 10. LDA output

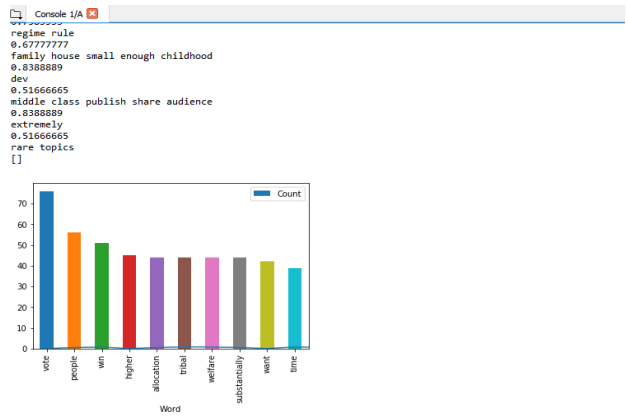

Figure. 11. Visualization graph of LDA

Now, in top of LDA supervised LDA (sLDA) and parallel LDA (pLDA) are applied to get more relevant output, which will be efficient in categorizing and will overcome the limitations of LDA. In Fig. 12 and fig. 13, we can find the result of topic modelling after applying sLDA on it. Response variables are added which will be helpful for categorization and prediction as well. pLDA speeded up the response time.

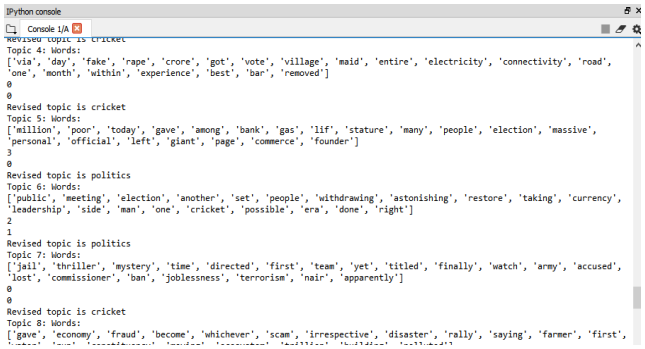

Figure. 12. Parallel-supervised LDA output

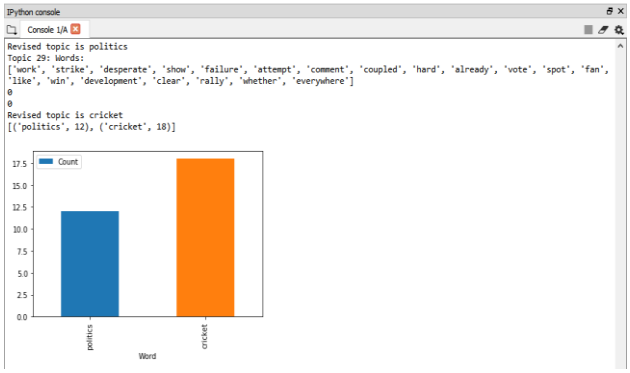

Figure. 13. Visualization graph of parallel-supervised LDA

From the table no. 1, we can compare the execution time of LDA and our implemented model:

Table no. 1. Comparision of execution time of topic modelling

Method name	Execution start time	Execution end time
LDA model	23:43:55.581317	23:44:00.091722
sLDA+pLDA model	23:58:37.963739	23:58:42.004589

It is taking less execution time in topic modelling when parallel LDA and supervised LDA are applied into dataset. Therefore, we can see that the result achieved by applying parallel-supervised LDA is highly accurate in categorization and the response time is also faster.

V. CONCLUSION

Enhancements of the conventional method is very much needed to keep pace with the increasing massive data and handling it in more efficient and accurate way. Therefore, new models are coming up with advanced algorithms in order to improve the topic modelling. Existing models had their own limitation. In case of obtaining more accuracy, the processing time of topic modelling goes high while in consideration of achieving more speed, accuracy gets low. Most of the algorithms implemented earlier cannot perform well in above-mentioned area. In this paper, we proposed and implemented parallel-supervised LDA model where supervised Latent Dirichlet Algorithm (sLDA) and parallel Latent Dirichlet Algorithm (pLDA) are applied together. And we have seen that high accurate results with quicker response time can be achieved by this model.

REFERENCES

1. Deerwester S, Susan T. Dumala, George W. Furnas, and Thomas K. Landauer, RicharHarshman,"Indexing by latent semantic analysis", *journal of the American Society for Information Science banner*, vol. 41, no. 6, pp. 391– 407, 1990.
2. T. Hofmann, "Probabilistic latent semantic indexing", *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, pp. 50-57, 1999.
3. D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet allocation", *J. Mach. Learn. Res.*, vol. 3, pp. 993-1022, Mar. 2003.
4. D. M. Blei, L. Carin, D. Dunson, “Probabilistic topic models” , *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55-65, Nov. 2010.
5. T.-C. Chou, M. C. Chen, "Using incremental PLSI for threshold-resilient online event analysis", *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 3, pp. 289-299, Mar. 2008.
6. C. Lin, Y. He, R. Everson, S. Ruger, "Weakly supervised joint sentiment-topic detection from text", *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134-1145, Jun. 2012.
7. X. Wang, N. Mohanty, A. McCallum, "Group and topic discovery from relations and their attributes", *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18, pp. 1449-1456, 2006.
8. Z. Chen, B. Liu, "Topic modeling using topics from many domains lifelong learning and big data", *Proc. 31st Int. Conf. Mach. Learn.*, pp. 703-711, 2014.
9. Hanqi Wang, Fei Wu, Weiming Lu, Yi Yang, Xi Li, Xuelong Li and Yueting Zhuang, "Identifying objective and subjective words via topic modeling", *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, vol. 29*, no. 3, pp. 718 – 730, 2017.
10. J. D. Mcauliffe, D. M. Blei, "Supervised topic models", *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, pp. 121-128, 2008.
11. Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, Edward Y. Chang, "Plda: Parallel latent dirichlet allocation for largescale applications", *Springer Verlag Berlin Heidelberg*, pp. 301–314, 2009.