Enhanced Topic Modeling

Poovammal E, Madhurima Mukherjee

Abstract*: We belong to an era of digitization where our collective knowledge is continuing to be stored in the form of electronic texts, i.e. blogs, news, scientific articles, web pages, images, audios, videos, social networks. As a result, it is getting more complicated to find out what we actually aim for. To handle this situation there is a rising need for analyzing huge collections of document. Topic modeling is a probabilistic generative modeling that is an efficient text mining technique for finding the hidden semantic structures of contents. 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 more purposes. New models of topic modeling are coming up with advanced inference algorithms. Improvements in algorithms will allow us to retrieve our required data in more efficient and optimized manner. The domain acts as a central concept for multiple on-going researches and we wish to add to it by our own survey. In this paper we have discussed about some methodologies which have been introduced in several papers of topic modeling.

Index Terms: Cloud computing, Data security, User behavior, Decoy technology, Fingerprint authentication, Face recognition.

I. INTRODUCTION

As there is a rising need to handle large volume of digitized data, this offers significant opportunities to the researchers of humanities and many industries. Data mining is the major term by which huge amount of data sets are analyzed and data patterns are retrieved using various techniques of data mining. Data mining comprises numerous domains of database systems, statistics and machine learning. Huge data sets consisting of millions of attributes are explored through data mining. Thus useful information and knowledge are drawn.Inreality, a major portion of data is stored in the form of text databases or through documents, which includes book, e-mail, blogs, web pages etc. These text databases are rapidly rising in electronic form. All the industries including government data also are becoming to be in digitized form.

Information retrieval for every department is becoming an immense necessity but conventional techniques of information retrieval have become ineffective for handling these volumes of increasing text data. It became very difficult to generate adequate queries to analyze and extract required information from data. More efficient tools and techniques needed to manage these complications. Thus text mining or text analysis came in picture and became very popular and a crucial area of data mining.Text mining covers the analysis on any type of text, i.e. newspaper, journal, article, book along with digitized text. Text mining incorporates several methods of analysis. Topic modeling is one of those efficient methods comprised in text mining. Topic modeling has a huge exposure in area of text mining. Specially, it has its main impact for handling large amount of digitized data sets.Topic modeling is a method of text mining that explores how the words are related with each other in a document, where topics are formed by grouping the words together.There are numerous techniques of topic modeling which use multiple sampling algorithms. Word selection as well as topic formation are the main purposes of these various algorithms. The most basic method of topic modeling is latent semantic analysis where frequency of words are looked in a document and topics are created by based on the often occurring words. Another basic topic modeling is latent dirichlet allocation (LDA) [8] where most likely appearing words are searched in a document and they are grouped together to form a topic.The amount of data over Internet are increasing rapidly. To keep a pace with it powerful topic modeling is a real need. It offers an efficient way to manage data. Topic modeling is a probabilistic generative model which is broadly used in Computer Science. After the first proposal of topic modeling it received huge attention and drew widespread interest in it among researchers in numerous research fields. A simple topic modeling concept has been shown infigure 1.

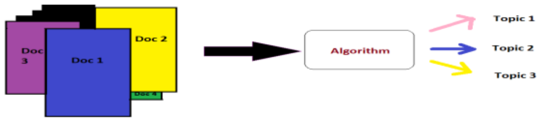

Figure 1. Topic Modeling

The origin of topic modeling is Latent Semantic Indexing (LSI) which is proposed by Scott Deerwesterand et al. in 1990 [9]. LSI served the basic steps for development of topic model, yet it lacks in solid probabilistic foundation.

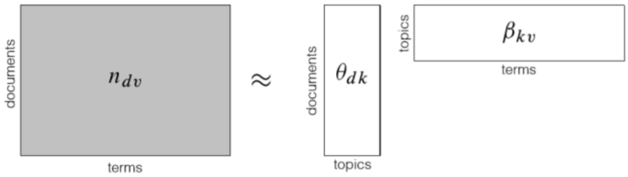

Figure 2. LSI Model

In figure 2 [2], we can see the very early model of topic modeling that is LSI model where ndvrepresentsa

Revised Manuscript Received on December 22, 2018.

Poovammal E, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, E-mail:poovammal.e@ktr.srmuniv.ac.in

Madhurima Mukherjee, Department of Computer Science and EngineeringSRM Institute of Science and Technology, Kattankulathur, madhurimamukherjee_as@srmuniv.edu.in

collection of documents, θdkrepresents each-document topic weights and βkvrepresents each-topic term weights.

It is basically the pioneering work which introduced topic modeling.

By term matrix of the term frequency–inverse document frequency (TFIDF)scoresa collection is treated as a document.

Choosing a number of topics SVD is to be run on the matrix

It will return two matrices :

each-document topic weights

each-topic term weights

After that Thomas Hofmann has proposed a novel approach of Probabilistic Latent Semantic Indexing (PLSI) [8] based on LSI for automated indexing built on the basis of statistical latent class model. Estimation Maximization (EM) is the standard calculation method for estimation of maximum likelihood.

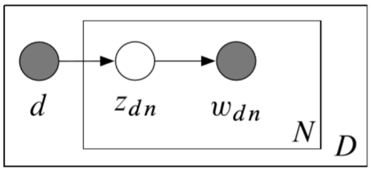

Figure 3. PLSI Model

A PLSI modelis presented by figure 3 [2], where,

singletopicis defined as one distribution over terms.

each document is explained as one distribution over topics.

There are two parametersets those are to be learned with EM.

D. M. Blei and Jordan proposed a probabilistic model, named Latent Dirichlet Allocation (LDA) [8] through which discrete data can be collected. It is a bayesian inference model. Bayesian inference calculates an event occurring probability for observed data set. Firstly, common sense is assumed. Then the output of previous suitable event and that common sense are combined through it. It is an iterative process of allotting words to relevant topics. The modeling will be more improved and accurate if more iterations are applied. LDA is first fully probabilistic [1] generative model in the area of text clustering. It randomly formulates noticeable data. Hidden variables are not observed directly. Here, posterior inference comes in picture. Posterior is calculated through the estimation of former evidence. Structure is thus uncovered and this uncovered structure is used to perform task.In figure 4 [2], LDA model is represented where all the nodes are considered as random variables. Dependence among the variables are indicated by the edges. The nodes which are shaded, are considered as observed and unshaded ones are as hidden. We can see the objectives of LDA.

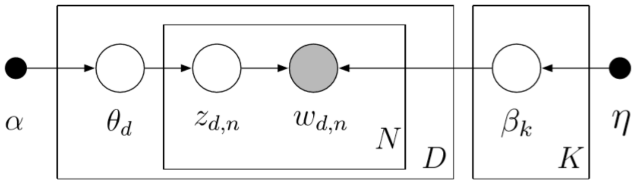

Figure 4. LDA Model

LDA is having two objectives:

Words are allocated to some topics in each document.

Higher probabilities are assigned to some terms in each topic.

Equation (i) is depicted from the above mentioned objectives:

$\log p(\cdot) = \ldots + \sum_{d}\sum_{n} \log p(z\text{dn} \mid \theta\text{d}) + \log p(w\text{dn} \mid \beta\text{zdn}) + \ldots$....equation(i)[2]

Groups of the tightly co-occurring words can be found by trading off these objectives.

LDA is the simplest building block of topic modeling that enables many Applications. Improvements in algorithm will allow us to fit models to massive amount of data. New models are developed along with advanced inference algorithms. New applications, visualizations and various tools techniques are invented based on LDA to detect fundamental topics from documents.In this paper we have discussed about some methodologies which have been introduced in several papers of topic modeling. Andwewill also see how these methods are needful to satisfy users' requirements.

II. RELATED WORKS

2.1.Growth of Probabilistic Topic Models.The origin of topic modeling is "Latent Semantic Indexing" (LSI)and later onResearchers have introduced many methodologies for improvement of topic modeling.

LSI is proposed by Deerwester S [9]. LSI endeavors to prevail over the lexical matching issues by utilizing "conceptual index" technique rather than utilizing individual words. LSI considers that few fundamental structures persist while using of word that is not demonstrated overtly by variation in word choosing. "Singular Value Decomposition" (SVD) is employed to appraise the formation of word usage in documents. Recovery is at that point carried out utilizing the singular-value database retrieved from the truncated SVD. LSI served the basic steps for development of topic model, yet it lacks in solid probabilistic foundation.

Hofmann has proposed a novel approach of "Probabilistic Latent Semantic Indexing" (PLSI) [11] based on LSI for automated indexing built on the basis of statistical latent class model. EM (Estimation Maximization) is the standard calculation method of maximum likelihood estimation, where EM is carried out through two steps of "Expectation" (E step) and "Maximization" (M step) sequentially. Word perplexity gets minimized by this model as it I mainly based on likelihood principle. Each word from a document will be captured from a mixed model indicated by means of multinomial random-variables, which was a significant step forward in the domain of probabilistic text modeling, yet was inadequate as in it presents no probabilistic constitution at the level of the document.

M. W. Berry and Brien have presented a survey [12] on the computational necessities for supervising the databases which are encoded to LSI and in addition to all the utilizations of LSI. The authors described how linear algebra can be employed in efficient way to retrieve the required information. Like, SVD is used for gaining required significant values from the huge number of datasets those who contain a very big number of values. As a result a fewer number of singular vectors are achieved which can be used against user queries. In this way terms and documents are depicted and this is called LSI, which is a promising method to enhance user‘s access to numerous sorts of textual materials.

Chien and Wu presented Probabilistic Latent Semantic Analysis (PLSA)[6] framework . It is basically a Bayesian approach. The authors have focused on taking an advantage of the incremental adaptation of the existing algorithm to tackle the updating issue of new articles. The authors applied Bayesian theory for developing two advanced adaptation paradigms for PLSA. Those are (1) Corrective learning (MAP PLSA), (2) Incremental learning (QB PLSA). Posterior likelihood is enhanced by the consistent employment of these two methods. Thus this algorithm enhance the document modeling by retrieving up-to-date data in its runtime. This method enables dynamic indexing of document together with modeling. For dealing with the domain discrepancy for language processing application PLSA should be adaptive.

Chou and Chen have proposed an algorithm named “Incremental Probabilistic Latent Semantic Indexing” (IPLSI) which is basically a “threshold resilient online algorithm” [7]. Main purpose of event analysis through online is detecting unknown events as well as tracking related documents and from that generation of story line. Here comes the typical challenge and that is threshold-dependency and handling the temporal relationship among the document stream. IPLSI algorithm efficiently handles the latent semantic continuation along timeline and thus improves the event detection quality. The performance evaluation results depict that this algorithm is responsible for cost reduction. And it can earn better performance by increasing the acceptable threshold range. It is really empirically efficient and theoretically sound algorithm.

D. M. Blei and Jordan have proposed a generative probabilistic model, named “Latent Dirichlet Allocation” (LDA) [8] through which discrete data can be collected. It is first fully probabilistic model in the area of text clustering. It improves the limitation of previously introduced LSI and PLSI method. It is basically based on exchangeability assumption in case of words and topics. It can be viewed as a technique of dimensionality reduction. A basic convexity-based variational approach has been presented for inference, demonstrating that it can yield a fast algorithm bringing about reasonable comparative performance with regard to test set likelihood.

2.2. Knowledge-based model. Chen and Liu proposed a knowledge-based model [5] to tackle the problem of incoherence among topics. They have shown that prior knowledge can bemineddynamicallywithout any user input from topics which are already found from a huge number of domains. The proposed model is called as “Lifelong Topic Model” (LTM). It enables indealing with knowledge which are possibly incorrect. A novel lifelong learning algorithm can be represented by LTM which is very much effective for topic discovery and is able to exploit the mined prior knowledge for generating better topic results. LTM can further be utilized to deal with big data.

Autoregressive model. Y. Zheng and Larochelleproposed a new approach [15] for topic modeling. It is an autoregressive method for multimodal data.“Supervised Document Neural Autoregressive Distribution Estimator” (SupDocNADE) is proposed by the authors which is basically a supervised extension of “Document Neural Autoregressive Distribution Estimator” (DocNADE), a type of topic modeling. This SupDocNADE can increase the discriminative ability of unseen topic elements where label information is consolidated in model trained object. Employment of SupDocNADE is also demonstrated to learn joint depiction from visual words, class label information together with annotation words. No iterative method is needed for the computation of an image’s representation. In a very simple manner it can achieve better performance.

III. COMBINATIONS OF LDA.

C. Lin and Ruger have proposed a novel probabilistic model, named “JointSentiment-Topic model” (JST) [3], based on LDA. It can detect topics and sentiment (such as attitudes, opinions, feelings expressed in text etc.) simultaneously from text. Reverse-JST model (A subsequently parameterized model of previously mentioned one) can also be prevailed by reverse sentiment sequence. When there is no hierarchical prior both the model performs similarly. But substantial experiments depict that with inclusion of sentimental prior JST plays reliably better role than Reverse-JST. Moreover, for the different domains the “weakly supervised” behavior of JST makes it exceedingly portable. In addition, those topics discovered by JST are to be surely informative and coherent.

X. Wang and McCallum have proposed a probabilistic generative model, named Group-Topic model [13] of entity relationships that detects groups and topics simultaneously from the textual attributes. Basically it uses the method of Group Latent Dirichlet Allocation (GLDA). Symmetric relations as well as have words as the attributes on relation are mainly focused in this paper. Basically GT model mutually finds latent groups and attribute-clusters which enhance communication among entities. GT model reaches out prior works on group detection by recording pair-wise relations as well as of multiple attributes. It achieves more improved topics and more cohesive groups can be discovered by the joint inference of the Group-Topic model.

YuepengZOU proposed a method [16], named the class frequency weight (CF-weight) to eradicate the negligence of the class frequency word information, which is very significant for classification but is neglected by subsisting supervised topic models. This method will weight words on the basis of “class-frequency” knowledge. The words which are having higher “class-frequency”, for them discrimination is less; similarly those which are having lower “class-frequency”, discrimination is more for them. It will improve the performance of labelled-LDA (L-LDA) and dependency-LDA. And these algorithms can keep up with subsisting supervised models.

Hanqi Wang and Zhuang proposed “identified objective-subjective

LatentDirichlet Allocation" (iosLDA) model [10], influenced by instinctive consideration that for each assigned topics all distinct words are having fluctuating level of discriminatory influence while conveying the sense of "subjective" and "objective". Here each document has two distinctive "Bag-of-Discriminative-Words" (BoDW) representation with respect to both the senses, those which can be utilized in classification of "subjective" and "objective" jointly rather than conventional "Bag-of-Topic" depiction. This analysis gave an account of documents and images, which shows that BoDW depiction is more prescient than conventional models. It also improves the topic modeling performance by means of mixed revelation of unseen topics as well as diverse "objective" and "subjective" properties covered up in each word. This algorithm also has brought down the execution intricacy, particularly when the topics are more expanded.

Chang have proposed parallel LDA [4] which is also a novel approach for large scale applications. PLDA generally smooths out the storage and the computation bottlenecks. It provides fault recovery when distributed computations are very lengthy. Scalability of LDA can be enhanced by significantly decreasing the bottleneck of unparallelizable communication which will help to achieve good load balancing. It will speed up the processing of topic modeling by providing it very quick pace. This pLDA can be applied to huge real-world applications to achieve good scalability and optimized parallel implementation.

IV. HYBRID METHODS

We have already discussed some of the existing methodologies of topic modeling on section 2. It is found that each and every model has its own kind of benefits. Still some models lack in few aspects. For example, though PLSI was a very significant step forward in the domain of probabilistic text modeling, it is not capable to correlate topics. For Probabilistic topic models it has inherited all the limitations of LDA where K-value is static and data is also predefined. Likewise, one algorithm having high efficiency may fail inclassification and prediction part. Some topic modeling may have more computational complexity; some may be time consuming whereas few models produce output with less accuracy.We would like to apply supervised LDA and parallel LDA together to achieve more meaningful topic modeling with quick response time in which Supervised LDA has advantages of collecting data from various sources and correlate related information. This will help in proper classification and prediction of texts. And parallel LDA can speed up the processing of topic modelling by providing it very quick pace so that we can achieve more enhanced and optimized topic modeling.

V. CONCLUSION

Organizing and tracking patterns in data have become very important in humanities, science, industry and culture. LDA is the simplest building block of topic modeling that enables many Applications. Improvements in algorithm will allow us to fit models to massive amount of data. Nowadays there are numerous complex probabilistic models which are basically based on LDA for serving some specific tasks. New models are coming up with advanced inference algorithms. In this paper we discussed and analyzed various methodologies of topic modeling which will help us for incorporating new implementation of topic models which may eradicate few drawbacks of existing topic models.

REFERENCES

1. Blei, D. (2012). "Probabilistic topic models." Communications of the ACM, 55(4), 77– 84.
2. Blei, D. M. (2013). "Probabilistic topic models: Origins and challenges,<http://www.cs.columbia.edu/ blei/talks>.
3. C. Lin, Y. He, R. E. and Ruger, S. (Jun. 2012). "Weakly supervised joint sentiment-topic detection from text." IEEE Trans. Knowl. Data Eng., 24(6), 1134–1145.
4. Chang, Y. W. B. S. Y. C. Y. (2009). "Plda: Parallel latent dirichlet allocation for largescale applications." Springer Verlag Berlin Heidelberg, 301–314.
5. Chen, Z. and Liu, B. (2014). "Topic modeling using topics from many domains, lifelong learning and big data." in Proc. 31st Int. Conf. Mach. Learn., 703–711.
6. Chien, J. T. and Wu, M. S. (Jan. 2008). "Adaptive bayesian latent semantic analysis." IEEE Trans. Audio, Speech, Language Process., 16(1), 198–207.
7. Chou, T.-C. and Chen, M. C. (2008). "Using incremental plsi for threshold resilient online event analysis." IEEE Trans. Knowl. Data Eng., 20(3), 289–299.
8. D. M. Blei, A. Y. N. and Jordan, M. I. (Mar. 2003). "Latent dirichlet allocation." J. Mach. Learn. Res., 3, 993–1022.
9. Deerwester S, Dumais ST, F. G. L. T. H. R. (1990). "Indexing by latent semantic analysis." journal of the American Society for Information Science banner, 41(6), 391– 407.
10. Hanqi Wang, Fei Wu, W. L. Y. Y. X. L. X. L. F. I. and Zhuang, Y. (2017). "Identifying objective and subjective words via topic modeling." IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 2162-237X 2017 IEEE.
11. Hofmann, T. (1999). "Probabilistic latent semantic indexing." in Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 50–57.
12. M. W. Berry, S. T. D. and Brien, G. W. O. (1995). "Using linear algebra for intelligent information retrieval." Vol. 37, 573–595.
13. X. Wang, N. M. and McCallum, A. (2006). "Group and topic discovery from relations and their attributes." Proc. Adv. Neural Inf. Process. Syst., 18, 1449–1456.
14. Ximing Li, JihongOuyang, Y. L. X. Z. T. T. (February 2015). "Group topic model: organizing topics into groups." Information Retrieval Journal, 18, 1–25.
15. Y. Zheng, Y.-J. Z. and Larochelle, H. (2014). "Topic modeling of multimodal data: An autoregressive approach." in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 1370– 1377.
16. Yuepeng ZOU, Ji-hong OUYANG, X. m. L. (2018). "Supervised topic models with weighted words: multi-label document classification." Front Inform Technol Electron Eng, 19, 513–523.
17. Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.
18. Rajesh, M., and J. M. Gnanasekar. "Sector Routing Protocol (SRP) in Ad-hoc Networks." Control Network and Complex Systems 5.7 (2015): 1-4.
19. Rajesh, M. "A Review on Excellence Analysis of Relationship Spur Advance in Wireless Ad Hoc Networks." International Journal of Pure and Applied Mathematics 118.9 (2018): 407-412.
20. Rajesh, M., et al. "SENSITIVE DATA SECURITY IN CLOUD COMPUTING AID OF DIFFERENT ENCRYPTION TECHNIQUES." Journal of Advanced Research in Dynamical and Control Systems 18.
21. Rajesh, M. "A signature based information security system for vitality proficient information accumulation in wireless sensor systems." International Journal of Pure and Applied Mathematics 118.9 (2018): 367-387.