IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review On Temporal Information based Topic Modeling Techniques

Ketaki D. Bhaskar

Student

Department Of Computer Engineering,

Matoshri College Of Engineering & Research Centre, Eklhare, Nashik, India

Abstract : In variety of domains large numbers of documents are generated every day. Mining text document and extracting useful information is challenging task. A group of words in a document describes the topic discussed in the document. Lot of work has been done for mining topic from a document set. A document containing time information helps to perform analysis of time series documents like collection of news articles, series of scientific papers, posts or tweets on social media sites, etc. By analyzing time specific document the topic discussed in particular temporal period can be extracted. Topics are evolved over time and are correlated. This paper includes the study of various topic modeling techniques and temporal topic analysis techniques. Based on the study of existing system, a new system is proposed for temporal topic modeling and forecasting the trends.

IndexTerms **- Text mining, Topic forecast, Topic discovery, Cluster labeling, topic modeling, Label identification**

I. INTRODUCTION

Large number of data is generated in variety of applications. Availability of such large volume dataset in the text form generates a need to manage such quantitative data in automated manner. As the information size increases, extracting useful information is a challenging task. The text mining techniques can be applied on such data and useful information is extracted. In case of text documents the relationship among documents is extracted by finding group of similar words. The actual content discussed in the document is analyzed based on its vocabulary words. The group of words represent topic discussed in those documents. This process is known as topic modeling.

The topic modeling helps to

1. Arrange all data in appropriate form like clustering.
2. Understand and summarize the document collection information.
3. Annotate the document for easier searching
4. Discover the hidden patterns in data

Such topic modeling is useful in variety of domains. For example: articles or post published on social media, articles published on news channel, paper published in publishing domain, etc. This helps to analyze which areas is higher importance.

The topics are evolved with respect to time and there is a correlation among topics. The articles published in certain time period defines the trend in topic. As time proceeds the change in topic can be seen. The analysis of topic with the help of time information is temporal topic analysis. Based on the temporal information the topic and its relationship can be extracted. The analysis of time specific topic information helps to predict the topic that may occur in future. The prediction of future topic is topic forecasting.

The problem of temporal topic modeling and trend forecasting from text document is three fold:

1. How to extract summarized information from text document
2. How to find trends in topic
3. How to forecast the topic that may occur in future.

Lot of work has been done in mining text document. The summarized information from text document is extracted using clustering, association rule mining, latent semantic modeling, Gibbes sampling etc. These techniques are studied independently in literature.

The system works on the processing of structured document. The structured document contains time information as well as text content. The system analyzes the document and extracts topic information using ensemble technique. The ensemble technique uses more than one mining technique for information collection. The proposed system uses: Gibbs sampling with Latent Dirichlet Allocation and Expectation Maximization algorithm for topic structure analysis. Based on the extracted topic information topic tend forecasting is performed. For topic trend forecasting regression model is used. As an output system find topics from documents, its relationship, topic dictionary and topic tend forecasting.

Following section describes the related work in text mining and topic modeling domain followed by problem formulation. In section IV the proposed system details are mentioned followed by conclusion.

II. RELATED WORK

C.Andreieu , Nando de Freitas , Arnaud Doucet, and Michael I. Jordan.[2] proposed machine learning in probabilistic way. For this , it adopts threefold system. Very first it uses Monte Carlo method. Secondly it uses modern Markov chain Monte Carlo method of simulations. Markov chain Monte Carlo techniques are frequently used to solve integration and optimization issues as far as large dimensional spaces are concern.

John Lafferty, et.al.[3] proposes a technique that helps to model the evolution of topic over time. This technique finds multinomial distributions of topic among multiple documents at each time epoch. Kalman filters and nonparametric wavelet regression technique are used to find approximate posterior inference over the latent topics.

David M. Blei, et. al, [4] proposes a topic modeling technique based on latent Dirichlet allocation (LDA). This topic modeling technique finds the correlation among topics and generates correlated topic model (CTM). This technique proposes mean-field variational inference algorithm. This algorithm is used for approximation of posterior inference in this model.

Tomoharu Iwata, et. al. [5] proposes a technique to perform sequential analysis of dynamic topics based on evolving time scale. This technique is called as Multiscale Dynamic Topic Model (MDTM). This technique considers the long as well as short time dependency and topic relationship is extracted. This technique uses expectation maximization (EM) algorithm. The topic model is updated as per the dynamic document update.

Dahua Lin ,et.,al.[6] proposes a segmented topic model (STM). A structured Document is input to the system. This paper follows different structure of document. The document is segmented based on the paragraph. Each paragraph is treated as a one sentence. This technique is useful for analyzing posts on social media where each post contains multiple statements grouped together in a paragraph. This technique uses collapsed Gibbs sampling and marginalized posterior of a two-parameter Poisson-Dirichlet process. This technique simultaneously processes the topic distribution and segment topic distribution under same latent space.

Amr Ahmed et, al.[7], addresses the problem of modeling time-varying document collections. This technique works on finding infinite dynamic topic models (iDTM). This technique finds topic evolved over time, the topic word distribution i.e. topic specific word dictionary and analyze topic trends adaptation over time. This technique uses Gibbs sampler. This system is useful to generate summarized information from topics and generate bird eye view information of each topic.

Hurtado, Jose, et. al,[8] proposes a system that mainly deals with structured document analysis, topic discovery and trend analysis. This paper uses association mining rules for finding patterns in a document. Using this technique, frequent patterns are extracted. It uses inclusion/exclusion operations. This technique uses NLP approach to find important words in a document. The extracted patterns are then refined. Based on the refined patterns topic community is identified using temporal frequency analysis and its correlation. Using the temporal topic information, topic forecasting is done.

Jun Song, et. al.[9], proposes The hierarchical topic evolution model (HETM) technique. The HTEM is proposed to process time-stamped documents. The timestamp information is used to measure the dependencies among multiple documents. (HTEM) captures the relation between evolving topics using a nested distance-dependent Chinese restaurant process. It uses Gibbs sampler for document information extraction. This technique organizes the time series documents in hierarchical manner. The topics near at the root note of the hierarchy are more abstract while topics at leaf nodes define more specific topics.

X.wang , Andrew McCallum. 2006 [10] proposed LDA- Style topic model. This model not only works for low dimenstional structured data but also analyze that how structure changes over time. Word co-occurrences and document's timestamp is considered for topic finalization. This paper deals with very particular datasets such as personnel email (9 - months) , research papers (17 years of NIPS research papers) , presidential state-of-the-union addresses (Over last 200 years). Timestapm predictions and interpretable trends are fetched.

Thomas L. Griffiths and Mark Steyvers. 2004[11] proposed a system to analyze the abstracts from PNAS website having various articles. For this Markov chain Monte Carlo algorithm is used. Small datasets is used as an input to the system. This dataset contains corpus of abstracts from PNAS from 1991 to 2001. Number of topics are fetched that represents particular corpus. Also hot topics are also extracted by analyzing the topic dynamics by using words assignments to particular topics to highlight the semantic content of documents from the corpus. This system is dedicated one.

Wei Li and Andrew McCallum. 2006 [12] this system proposed PAM (pachinko allocation model) over LDA. This is because topic correlation is not captured in LDA. Whereas Pachinko Allocation Model calculates and captures arbitrary , sparse as well as nested topic correlation using DAG (Direct Acyclic Graph). DAG has leaves which are nothing but the words and their nodes a nothing but the correlation amongst children. PAM shows improved performance in document classification.

Hida, Rem, et.al.[1] proposes a study on static and dynamic relationship among topics. The system takes structured document as an input. The document contains time information and the text content such as news article, research papers, etc. For each time epoch i.e. defined time span the relationship among text document is identified. The relationship is identified based on the word collection and topic information. The relationship among documents at different time epoch is also studied called as dynamic topic analysis. For modeling topic information ensemble approach is used. This technique uses Latent Dirichlet Allocation with Gibbs Sampler and Expectation maximization(EM) algorithm is used. The System considers frequently occurred words as a candidate words for topic modeling and do not contribute in topic dictionary creation and topic trend forecasting.

Paper	Description	Analysis
C.Andreieu , Nando de Freitas , Arnaud Doucet, and Michael I. Jordan.[2]	It proposed machine learning in probabilistic way using three fold system.	For the very first time MCMC methods are used for large scale dataset classification.
John Lafferty, et.al.[3]	It proposed Kalman filters and nonparametric wavelet regression technique are used to find approximate posterior inference over the latent topics.	Topic evolution over time is calculated. Time factor in analysis is introduced.
David M. Blei, et. al, [4]	It proposed a topic modeling technique based on latent Dirichlet allocation (LDA). This topic modeling technique finds the correlation among topics and generates correlated topic model (CTM).	LDA which is very reliable algorithm for topic extraction is discussed.
Tomoharu Iwata, et. al. [5]	It proposed a technique to perform sequential analysis of dynamic topics based on evolving time scale. This technique is called as Multiscale Dynamic Topic Model (MDTM).	Major work is done on topic relation by considering small as well as long time dependency. Using expectation maximization (EM) algorithm. The topic model is updated as per the dynamic document update.

Paper	Description	Analysis
Dahua Lin ,et.,al.[6]	It proposed Segmented Topic Model (STM). The document is segmented based on the paragraph. Also This technique uses collapsed Gibbs sampling and marginalized posterior of a two-parameter Poisson-Dirichlet process. This technique simultaneously processes the topic distribution and segment topic distribution under same latent space.	Document division on the basis of paragraph technique is tried. But it has certain drawbacks such as it is assume that Each paragraph is a one sentence. Sometimes which cannot be the case
Amr Ahmed et, al.[7]	It addresses the problem of modeling time-varying document collections. This technique works on finding infinite dynamic topic models (iDTM). It finds topic evolved over time. Trend over time is analyzed. Gibbs sampler techniques is used for this.	This system is useful to generate summarized information from topics and generate bird eye view information of each topic. Also it is proved that Gibbs sampler is useful technique.
Hurtado, Jose, et. al,[8]	It proposes a system that mainly deals with structured document analysis, topic discovery and trend analysis. NLP is firstly used to find important words in documents. Also association mining rules are used to find frequent patterns in documents. These frequent patterns are refined and based on that topic community is identified using temporal frequency analysis and based on that topic forecasting is done.	In this paper it is clear that NLP is useful to find important words as well as association mining rules are used to find frequent patterns. By considering temporal frequency analysis topic forecasting is done.
Jun Song, et. al.[9],	It proposed The hierarchical topic evolution model (HETM) technique. The timestamp information is used to measure the dependencies among multiple documents. (HTEM) captures the relation between evolving topics using a nested distance-dependent Chinese restaurant process. Used Gibbs sampler for document information extraction. Also this technique organizes the time series documents in hierarchical manner.	Hierarchical arrangement of documents is considered. The topics near at the root note of the hierarchy are more abstract while topics at leaf nodes define more specific topics.
X.wang , Andrew McCallum. 2006 [10]	It proposed LDA- Style topic model. This model not only works for low dimenstional structured data but also analyze that how structure changes over time. Word co-occurrences and document's timestamp is considered for topic finalization.	This paper deals with very particular datasets such as personnel email (9 - months) , research papers (17 years of NIPS research papers) , presidential state-of-the-union addresses (Over last 200 years). Word co-occurrences discussed for the very first time
Thomas L. Griffiths and Mark Steyvers. 2004[11]	It proposed a system to analyze the abstracts from PNAS website having various articles. For this MCMC algorithm is used. Number of topics are fetched that represents particular corpus. Also hot topics are also extracted.	Small datasets is used as an input to the system. This dataset contains corpus of abstracts from PNAS from 1991 to 2001. This system is dedicated one.

Wei Li and Andrew McCallum. 2006 [12]	this system proposed PAM (pachinko allocation model) over LDA. This is because topic correlation is not captured in LDA. Whereas Pachinko Allocation Model calculates and captures arbitrary , sparse as well as nested topic correlation using DAG (Direct Acyclic Graph). DAG has leaves which are nothing but the words and their nodes a nothing but the correlation amongst children..	PAM shows improved performance in document classification.

Paper	Description	Analysis
Hida, Rem, et.al.[1]	It proposed study on static and dynamic relationship among topics. Used structured document as an input. The document contains time information and the text content such as news article, research papers, etc. For each time epoch i.e. defined time span the relationship among text document is identified. The relationship is identified based on the word collection and topic information. The relationship among documents at different time epoch is also studied called as dynamic topic analysis.	For modeling topic information ensemble approach is used. This technique uses Latent Dirichlet Allocation with Gibbs Sampler and Expectation maximization(EM) algorithm is used. The System considers frequently occurred words as a candidate words for topic modeling and do not contribute in topic dictionary creation and topic trend forecasting. This approached is much better and selected for implementation purpose.

III. ANALYSIS AND PROBLEM FORMULATION

Lot of work has been done in text mining and topic modeling. Some of the existing techniques only find the topic from group of document, some focuses on topic modeling and topic specific dictionary creation. The topics are evolved over time. Hence new topics are proposed to study temporal relationship among topic and detection of topic trends at each time epoch. Based on the topic trend information some techniques contribute to forecast topics that may occur in future.

All the techniques have their specific aim and uses data mining technique such as Gibbes sampler, Latent Dirichlet Allocation, clustering of documents using EM algorithm. There is need to collects summarized information in the form of :

- Topics from a given dataset
- Generate topic Dictionary
- Trending top k topics
- Topic trend forecasting

By collectedly applying multiple techniques on the temporal document set.

IV. PROPOSED METHODOLOGY

A. Architecture

Following fig.1 describes the architecture of the system. The structured document set is input to the system. The structured document contains text document and timestamp information. Such document set is called as temporal document set. After processing the document set system generates topic list and relationship among topics, topic dictionary and topic trend that may occur in future as an output. The system mainly uses Gibbs sampling with Latent Dirichlet Allocation algorithm, Expectation maximization algorithm and Regression algorithm as an ensemble approach for topic discovery and trend forecasting.

The input document set is initially partitioned as per the timestamp information. The group of document within defined time span is given input to the LDA and gibbs sampler. This technique fined the probable topic words. The topic word distribution is given to the EM algorithm. This algorithm returns the topic information as per the timestamp. Based on the topic discovery the dictionary creation function returns the generated dictionary for different topics. The time specific topic discovery information is then given to the topic trend forecasting module. This module includes the regression based forecasting algorithm. This algorithm is useful for trend forecasting.

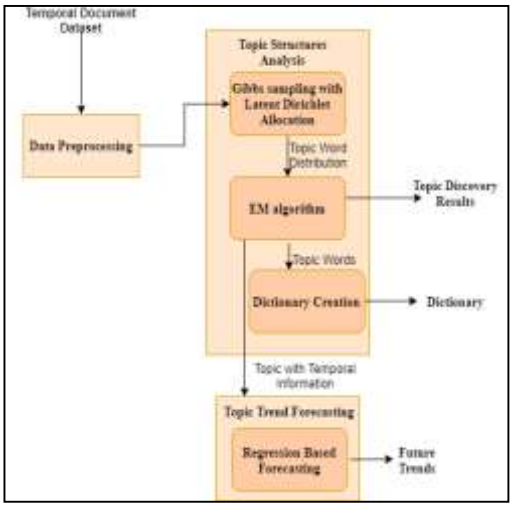

Figure 1 : System Architecture

B. System Working:

The data processing is mainly described in following 4 sections:

1. **Data Preprocessing:**
   In data preprocessing language processing techniques are used and extracted nouns, adjectives, and adverbs in the statements. The extracted words are then lemmatizing that converts the words in original form. This word set is used for further processing for topic modeling.
2. **Topic Assignment:**
   In this section super-topic and subtopics are extracted from the documents based on the words collection. For topic extraction Gibbs sampling with Latent Dirichlet Allocation is used.
   For time dependant analysis of documents, an EM algorithm is applied. This algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters.
3. **Dictionary Creation:**
   Based on the Latent Dirichlet Allocation and NLP technique a filtered topic specific dictionary is created.
4. **Topic Trend Forecasting:**
   In this section, topic trends forecasting is performed. Ensemble forecasting approach is proposed to predict the popularity of research topics in the future. For forecasting regression models in the WEKA tool is used.

V. CONCLUSIONS

Lot of work has been done in text mining and topic modeling. Some of the existing techniques only find the topic from group of document, some focuses on topic modeling and topic specific dictionary creation, some finds the temporal topic information. Each technique has its own advantages and limitations. Based on the analysis of existing system, a new system is proposed. The proposed system works on topic discovery form temporal text document data. The system works on time series documents and finds topic trends based on temporal information. For analyzing trends it uses Gibbes sampling with Latent Dirichlet Allocation and EM Clustering algorithm. The extracted topic words are grouped together in topic dictionary. Based on the current topic information future topic trends are extracted using regression based forecasting.

VI. REFERENCES

[1] Yang Hida, Rem & Takeishi, Naoya & Yairi, Takehisa & Hori, Koichi,."Dynamic and Static Topic Model for Analyzing Time-Series Document Collections",May,2018

[2] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. 2003. An introduction to MCMC for machine learning. Machine Learning, 50(1):5–43

[3] David M. Blei and John D. Lafferty,"Dynamic topic models", In Proceedings of the 23rd International Conference on Machine Learning, 2006, pages 113–120.

[4] John D. Lafferty and David M. Blei. 2006. Correlated topic models. In Advances in Neural Information Processing Systems, volume 18, pages 147–154.

[5] Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda, "Online multiscale dynamic topic models", in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pages 663–672.

[6] Lan Du, Wray Buntine, and Huidong Jin., "A segmented topic model based on the two-parameter Poisson-Dirichlet process", Machine Learning,2010, 81(1):5–19.

[7] Amr Ahmed and Eric P. Xing., "Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream", In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, 2010, pages 20–29.

[8] Hurtado, Jose & Agarwal, Ankur & Zhu, Xingquan,"Topic discovery and future trend forecasting for texts", in Journal of Big Data, in researchgate,2016 PP:10.1186/s40537-016-0039-2.

[9] Jun Song, Yu Huang, Xiang Qi, Yuheng Li, Feng Li, Kun Fu, and Tinglei Huang, "Discovering hierarchical topic evolution in time-stamped documents", Journal of the Association for Information Science and Technology, 2016, 67(4):915–927.

[10] Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 424–433

[11] Thomas L. Griffiths and Mark Steyvers. 2004. " Finding scientific topics." In Proceedings of the National Academy of Sciences of the United States of America, volume 101, pages 5228–5235.

[12] Wei Li and Andrew McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd International Conference on Machine Learning, pages 577–584