IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Automatic Inference Multiple Topics and Trend Analysis using DSTM

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. This work focuses on analysis of time series documents like collection of news articles, series of scientific papers, posts or tweets on social media sites, etc. Topics are evolved over time and are correlated. The system finds temporal topic evolution and topic hierarchy. Along with the topic modeling, topic trend forecasting is also done. The document data is pre-processed using machine learning techniques and important words are extracted. These words are used for topic modeling. The topic modeling is performed using Latent Dirichlet Allocation and Gibbs Sampler. The important topic words are treated as topic dictionary. The performance in terms of accuracy will be compared with the existing approaches.

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.

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.

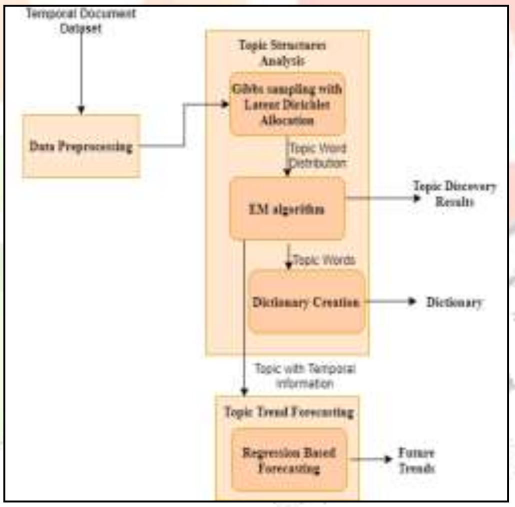

Temporal Document Dataset
Data Preprocessing
Topic Structures Analysis
Gibbs sampling with Latent Dirichlet Allocation
Topic Word Distribution
EM algorithm
Topic Discovery Results
Topic Words
Dictionary Creation
Dictionary
Topic with Temporal information
Topic Trend Forecasting
Regression Based Forecasting
Future Trends

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.

C. Algorithms

```
Algorithm 1: DSTM algorithm:
Input: : Dataset with Dt documents and Temporal information T
Output: TL: Topic List,
TG: Topic Dictionary
FTT: Forecasted Topic Trends
Processing:
1.  Preprocess data and Extract words
2.  For each subtopic k = 1,2..,K
        Draw a topic-word distribution
3.  For each document d = 1,2..,Dt,
4.  Draw a super-topic distribution
5.  For each super-topic s = 1,2..,S,
6.  For each word i = 1,2..,$n^t_d$
7.  Draw a super-topic-word assignment
8.  Draw a subtopic-word assignment
9.  Draw a word-observation
10. Allot super-topic and subtopic to each document
11. TL: Label documents
12. TG: Generate topic dictionary
13. Study Temporal relationship among topic
14. FTT: Forecast topic Trends
```

D. Mathematical Modeling:

The System S can be defined in set theory form as,
S = {I, O,F} where,
I = {D, M}, Set of inputs
D = Text document with temporal information
M = Top number of trends count

O= {TL, TG, FTT } Set of outputs
TL = Topic List,
TG = Topic Dictionary
FTT = Forecasted Topic Trends

F ={F1,F2,F3,F4,F5,F6,F7,F8,F9,F10,F11,F12 } Set of Functions
F1 = Upload Temporal Data
F2 = Data Preprocessing
F3 = Stop-word Removal
F4 = Lemmatization
F5 = Keyword Extraction
F6 = Topic Assignment
F7 = Gibbs sampling with Latent Dirichlet Allocation
F8 = EM algorithm
F9 = Topic Structures Analysis
F10 = Dictionary Creation
F11 = Topic Trend Forecasting
F12 = Regression Based Forecasting

V. RESULT AND ANALYSIS

The system is implemented using jdk1.7 on Windows 10 operating system. For Implementation core I5 system with 4gb ram is used.

A. Dataset:

Drone[10] dataset is is used for testing. Drone is the dataset containing articles from journals for surveying recent developments in unmanned aerial vehicles (UAV). Following table 1 gives the summarize information of dataset.

TABLE 1 : DATASET DESCRIPTION

#	Drone Dataset
Document Span	2009–2016
Number of documents	1,035
Vocabulary Words	3,442
Number of Tokens	68,305

B. Performance Measures:

The system performance is measured in terms of :

1. Time:
   The time required for processing topic extraction, dictionary creation and topic forecasting.

2. perplexity (PPL): From the dataset, 90% of data is given as training and remaining 10% data is treated as test dataset. The mean value of perplexity of 10 fold data is calculated using following equation.

$$PPL = \exp \frac{\sum_{d=1}^{D} \sum_{w_d^{test}} \log p(w_{d,i} \mid M)}{\sum_{w_d^{test}} n_d^{test}} \qquad (1)$$

   Where,
   $w_{d,i}$ = the ith word in the dth document in test dataset
   n_d^{test} = number of words in the dth document in test dataset

3. Precision: This is for future trend forecasting. This can be calculated as:

$$\text{Precision} = \frac{Ot \cap Rt}{Rt} \qquad (2)$$

   Where,
   Ot = Original number of documents at time epoch t
   Rt = Retrieved predicted document at time epoch t

Results:

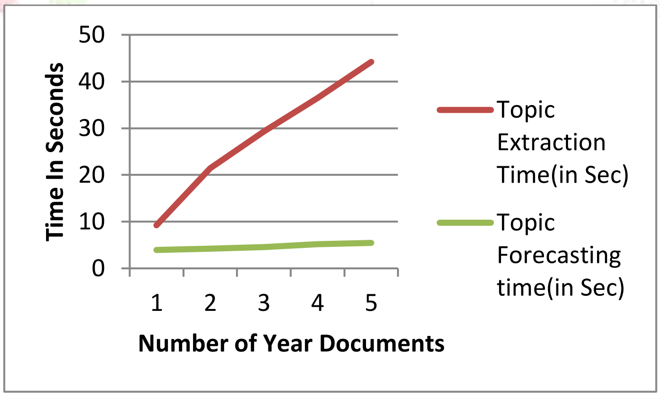
1. Time and Accuracy:

   Following table contains the results for topic and subtopic discovery for static and dynamic analysis. For static analysis a single document is input to the system. System finds the topics and subtopics in 9.23 sec. The time series data is input to the system. System extracts the static model from each document set for a given timestamp and finds the relational changes among the models with respect to the time.

   For the given time stamp based information system can predict topics for further timestamps. The number of topics to be predicted is input to the system. The following table contains the number of topic prediction time and its accuracy is enlisted. The accuracy is measured by comparing the predicted topics with actually occurrence topics.

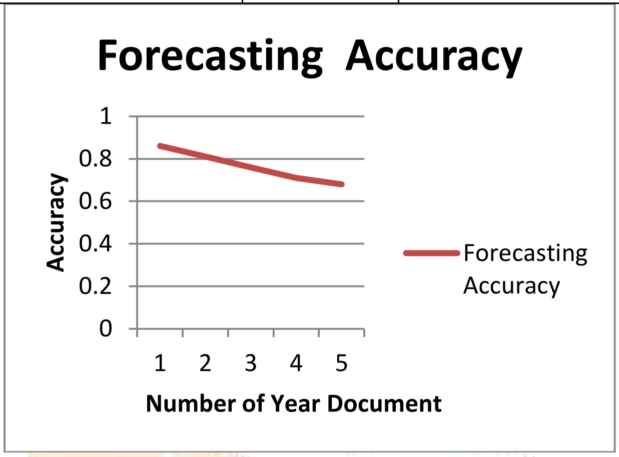
Number of Year Documents	Topic Extraction Time(in Sec)	Topic Forecasting time(in Sec)	Forecasting Accuracy
1	9.23	3.94	0.86
2	21.43	4.21	0.81
3	29.32	4.54	0.76
4	36.43	5.2	0.71
5	44.23	5.45	0.68

Following table contains the comparative study of topic extraction time ans topic forecasting time. The time requried for calculating the topics from actual dataset requires higher time as comapred to the time series based forecasting. The forecasting is Auto Regressive Integrated Moving Average model while topic-subtopic discovery executes the Gibbs sampling with Latent Dirichlet Allocation followed by EM algorithm.



The following graph includes the analysis of accuracy of topic forecasting. As we increase the forecasting count the accuracy decreases. The accuracy of prediction is high for early forecasting.

	Topic Discovery	Sub topic super topic discovery	Time series analysis	Time Series Forecasting
Dynamic topic models	Y		Y	
Correlated topic models	Y	Y		
Topic discovery and future trend forecasting for texts[16]	Y		Y	Y
Dynamic and Static Topic Model for Analyzing Time-Series Document Collections[1]	Y	Y	Y	
Proposed System	Y	Y	Y	Y



System Comparison:

Following table contains the comparative study among multiple approaches. The proposed approach is combination of all topics such as topic discovery, topic and subtopic discovery, time series analysis and time series forecasting.
The topic discovery and topic subtopic discovery is static analysis model where as time series analysis is the dynamic analysis of documents having time stamp information. The time series forecasting is the dynamic analysis of data and helps to predict the future occurrence of values.

VI. CONCLUSIONS

The proposed system works on finding co-occurrence patterns on based on its time information. The dataset is divided in number of time cubes based on the time information. Apriori algorithm is used to find co-occurrence patterns in each cube and from the overall dataset. Density of each cube is checked to avoid over estimation problem. For efficiency improvement, parallel processing is introduced. In parallel processing multiple cubes are processed simultaneously. During system execution, following points are noticed:.

- Time and memory are not depends on the dataset size. It depends on the number of itemset found based on a given minimum support value.
- Parallel processing requires less time as compared to the single thread application for all minimum support values.

In Future the system can be implemented on continuous streaming data.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Valerio Perrone, Paul A. Jenkins, Dario Spano, and Yee Whye Teh., "Poisson random fields for dynamic feature models", arXiv:1611.07460, 2016.

[10] Chun Fui Liew, Danielle DeLatte, Naoya Takeishi, and Takehisa Yairi., "Recent developments in aerial robotics: A survey and prototypes overview", arXiv:1711.10085,2017