Time Stamp Based Mining in Multiple Asynchronous Text Sequences

Suresh Kumar*, Mrs. D. Saravanapriya**

*Department of computer science, PA College of engineering and technology,
Pollachi, Coimbatore District, Tamilnadu, India
sureshkumar.apacet@gmail.com

**Assistant professor, Department of information technology, PA College of engineering and technology,
Pollachi, Coimbatore District, Tamilnadu, India
sdspriyapacet@gmail.com

Abstract—Text sequences are ubiquitous, multiple text sequence are often related to each other by sharing common topics. The interactions among these sequences provide more information to derive more meaningful topics. Discovering valuable knowledge from a text sequence involves extracting topics from the sequence with both semantic and temporal information. The method is relied on a fundamental assumption that different sequences are always synchronous in time. The documents from different sequences on the same topic have different time stamp and there is no guarantee that the articles covering the same topic are indexed by the same time stamps. The key idea is to introduce a generative topic model for utilizing correlation between the semantic and temporal information in the sequences. Topic model is mainly focused on extracting a set of common topics from given sequences using their original time stamps. It performs topic extraction and time synchronization alternatively to optimize a unified objective function. A local optimum is guaranteed with the proposed method.

Index Terms-Asynchronous sequences, temporal text mining, Topic model.

I. INTRODUCTION

Text mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics. Text mining is alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from Text mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics.

Text mining is alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. Text analysis involves information retrieval, lexical analysis to word frequencydistributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics.

Addressing the problem of topic detection is the main focus of in text mining. The goal of the task is to identify a collection of news articles about a topic. It is viewed as d as natural text streams with publication dates as time stamps. It would be very useful it can discover, extract, and summarize the evolutionary theme patterns automatically. The algorithm contains several interesting applications that can make it easier for people to understand the information contained in large knowledge domains, including exploring topic dynamicsand indicating the role that words play in the semantic content of documents.

Application domains, encounter a stream of text, in each text document has some meaningful time stamp. An event covered in news articles generally has an underlying temporal and Evolutionary structure consisting of themes characterizing the beginning, progression, and impact of the event, among others. It is classification of document into topics and actions into activities.

Parameter estimation in these models discovers a low-dimensional set of multinomial word distributions called topics in textual documents. Mixtures of these topics give high likelihood to the training data, and the highest probability words in each topic provide keywords that briefly summarize the themes in the text collection. The topic models have also been applied to images, biological findings and other non-textual multi-dimensional discrete data.

The topics addressed by a paper are also one of the first pieces of information a person tries to extract reading a scientific abstract. Papers are relevant to their interests, search areas are rising or falling in popularity, and the papers are related to one another. A statistical method is provided for automatically extracting a representation of documents that provides a first-order approximation to the kind of knowledge available to domain experts. The method discovers a set of topics expressed by documents, providing quantitative measures that can be used to identify the content of those documents, track changes in content over time, and express the similarity between documents.

A novel problem of text mining referred to as Baseline Text Mining. The task of comparative text mining is to discover any latent common themes across all collections as well as summarize the similarity and divergences of these collections along each common theme. The task of comparative text mining involves discovering the different common themes across all the collections and for each discovered theme, characterize the common and unique term.

The rest of the paper is organized as follows: related work is discussed in Section II; formalize our problem and propose a generative model with a unified objective function in Section III; how to optimize the objective function in Section IV; extensions of our model and algorithm are discussed in Section V; and conclude our work in Section VI.

II. RELATED WORK

The authors introducedpresent asynchronous distributed learning algorithms for two well known unsupervised learning frameworks is Latent DirichletAllocation and Hierarchical Dirichlet Processes the work contain some distinction. It will contain purely asynchronous communication. It is not applicable to the collapsed sampler for LDA. Another method is mining correlated busty topic patternsfrom coordinated text streams can reveal interesting latent associations or events behind these streams. It is effectively discover quite meaningful topic patterns. Using mutual reinforcement across streams discover correlated busty Topic patterns methods. It is applicable text stream only.

The approach that is to use state space models on the natural parameters of the multinomial distributionsthat represent the topics. Variational approximations based on Kalman filters and nonparametric wavelet regressions are developed to carry out approximate posterior inference over the latent topics. In addition to giving quantitative, predictive models of a sequential corpus, dynamic topic models provide a qualitative window into the contents of a large document collection. The main problem contains Factorial Hidden Markov models for topic intensity tracking with exponentialorder statistics for implicit data association. At the problem is data association and intensity tracking of multiple topics over time. The approach detects correct topic intensities even with 30% topic noise.

Another model capture arbiter nested a possibly sparse correlation between topics and then using a directed acyclic graph. The leaves of the DAG represent individual words in the vocabulary, each interior node represents correlation among its children may be words or other interior node. It proposed to correlation among children, may be words or other interior nodes. A correlation of textual documents parameter estimation in these model distributions called topics. Mixtures of this topic give likelihood to the training data and the highest probability words in each topic provide keywords that briefly summarize the themes in the text collection.

The authors introduced a probabilistic model to incorporate contentand time information in a unified framework. This model gives new representations of both news articles and news events. This algorithm is easy to understand and implement. A disadvantage is to find better representations of the contents of news articles very difficult.

Symbols	Description
WF	Total number of words in a particular sequence
WS	Number of words in a all sequence
SW	Number of sequence in Database which a particular word is found
TE	Topic Extraction
IF	Inverse frequency
ITE	Inverse topic frequency
SD	Sequence in Database

Table.1 Symbol and Their Meaning

The novel problem of mining spatiotemporal theme patterns from weblogs and propose a novel probabilistic approach to model the subtopic themes and spatiotemporal theme patterns simultaneously. The proposed model discovers spatiotemporal theme patterns by extracting common themes from weblogs. The proposed probabilistic model is general and can be used for spatiotemporal text mining on any domain with time and location information. The probabilistic model is generally applicable not to any text collections with time and location information, but also for other text mining problems.

Themodel is Markov assumptions or discretization of timeeach topic is associated with a continuous distribution over timestamps, and for each generated document, the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp. The meaning of a particular topic can be relied upon as constant, but the topics occurrence and correlations change significantly over time. More improved topics, better timestamp prediction, and interpretable trends.

Text stream is important to know that the hot burst events detection problem, it is different from TDT. It focuses, detecting a set of burst features for a burst event. Propose a new novel parameter free probabilistic approach, called feature-pivot clustering. Main technique is to fully utilize the time information to determine a set of burst features may occur in different time windows. Detect burst events based on the feature distributions. There is no need to tune or estimate any parameters.

III.PROBLEM STATEMENT

The asynchronies among multiple sequences, i.e., documents from different sequences on the same topic have different time stamps, is actually very common in practice.The main symbols used throughout the paper are listed in Table 1.

For instance, in news feeds, there is no guarantee that news articles covering the same topic are indexed by the same time stamps. There can be hours of delay for news agencies, days for newspapers, and even weeks for periodicals, because some sources try to provide first-hand flashes shortly after the incidents, while others provide more comprehensive reviews afterward.

Another example is research paper archives, the latest research topics are closely followed by newsletters and communications within weeks or months, and then the full versions may appear in conference proceedings, which are usually published annually and at last in journals, which may sometimes take more than a year to appear after submission.

IV. PROPOSED ALGORITHEM

Formally address this problem and put forward a novel algorithm based on the generative topic model.

Our algorithm consists of two alternate steps:

- The first step extracts common topics from multiple sequences based on the adjusted time stamps provided by the second step.
- The second step adjusts the time stamps of the documents according to the time distribution of the topics discovered by the first step.

V. DISCUSSIONS AND EXTENSIONS

Perform these two steps alternately and after iterations a monotonic convergence of our objective function can be guaranteed. The effectiveness and advantage of our approach were justified through extensive empirical studies on two real data sets consisting of six research paper repositories and two news article feeds, respectively.

The current time stamps of all sequences are synchronous and the common topics are extracted and TE is calculated for each of them. TE/ITF can be calculated for each word using the four values such as number of words in a document, frequency of a word in a document, the number of total documents, and the number of documents where the word appears. Instead of extracting words from an e-text, two-word phrases were extracted and TEITF is calculated for each of them.

$$TE = WS/WF^{-}1 \qquad (1)$$

Inverse frequency (IF) is essential. In this percentage denoting the number of times a word appears in a document [10]. It is mathematically expressed as WS/WF, where WS is the number of times a word appears in a document and WF is the total number of words in the same document. Inverse document frequency (IDF) takes into account that many words occur many times in many documents. IDF is mathematically expressed as SD/SW, where SD is the total number of sequence in database and SW is the number of document in which a particular word is found. As SD/SW increases so do the significance of the given word.

$$ITE = SD/SW^{-1} \qquad (2)$$

A. *Time Synchronization*

The timestamps (IF) are adjusted to synchronize the sequences. Once the common topics are extracted, the documents are matched to the topic. Topic related content retrieval from the various unstructured document based upon Time Synchronization

$$IF = TE/ITE \quad (3)$$

$$H(1:1,1:a) = \sum_{m,t=1} \max 1 \leq s\,a \sum_{w} Q(W,s)C(W,d) \qquad (4)$$

In this (3) and (4) equation gives the global optimum to our objective function in [6].

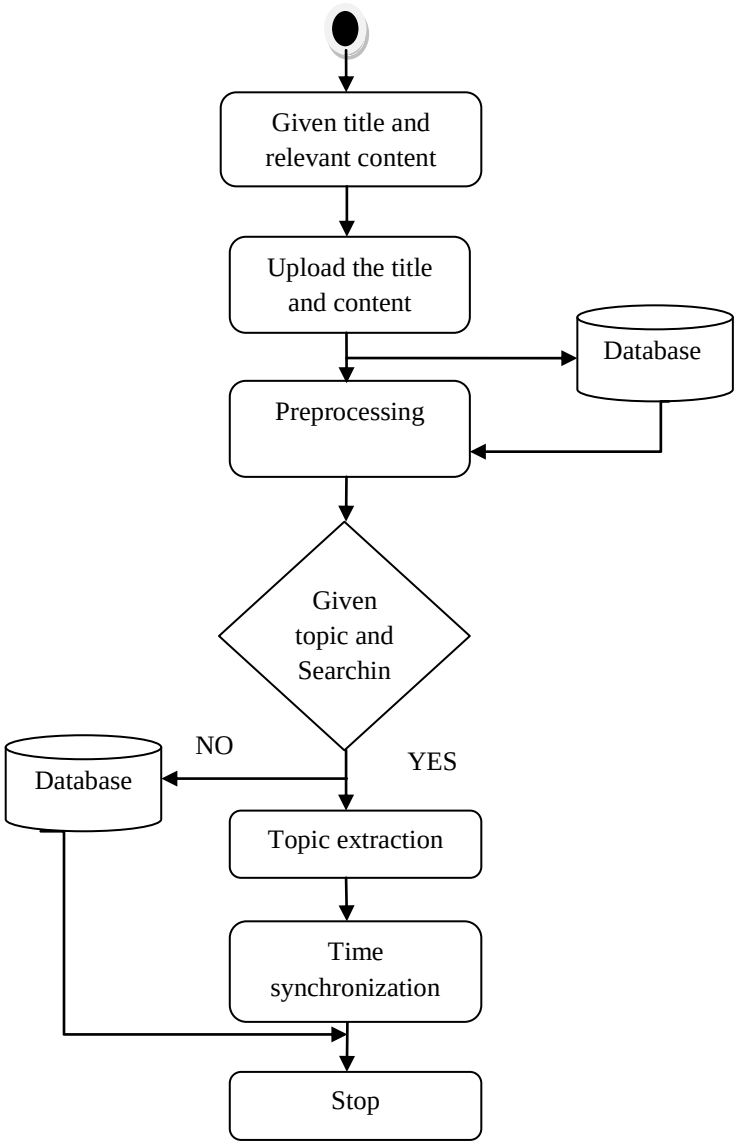

Fig.1.Query based browsing

B. *The local search stretchy*

A quantitative estimation of the asynchronism among sequences is available and it is unnecessary to search the entire time dimension is adjusting the time stamps of documents. It gives the opportunity to reduce the complexity of time synchronization step without causing substantial performance loss, by setting an upper bound for the difference between the time stamps of documents before and after adjustment in each iteration.

C. *The Baseline Method and Implementation*

The standard PLSA method as the topic extraction step of our algorithm. Yet in the experiments, we introduced two additional techniques as used in and this modified version of the PLSA algorithm was used as a baseline method for topic extraction.

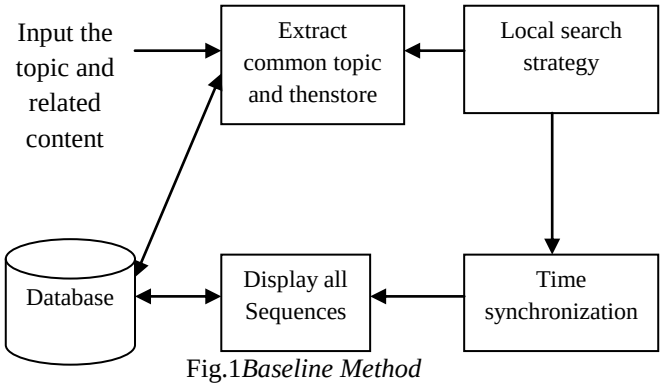

Fig.1*Baseline Method*

The first technique is to introduce a background topic into our generative model so that background noise can be removed and find more burst and meaningful topics.

D. *Algorithm*

STEP 1: Input to the document or topic.
STEP 2: Give the topic related content.
STEP 3: Using preprocessing state.
STEP 4: Get topic related content,allsequence already synchronous and extract common topic using topic extraction.
STEP 5: Give the searching string and then pick the searching related content.
STEP 6: Extractcommon topic content to be displayed.
STEP 7: Once the common topics are extract, match documents in all sequences' and then display synchronize the Sequences.
STEP 8: Get document content from unstructured Text sequence.

VI. RESULTS AND DISCUSSION

The proposed work web search result personalization is focused. From the Table.1 it is clear that the user can get the efficient results based on their domain. Queries based Browsing provides increased proficiency as in Fig.1.

Domain	**Number of queries**		**Response efficiency**
Common Login (%)		**Query based Login (%)**	
Science	50	52	56
Education	40	68	72
Research	42	45	62
Software	38	67	70
Total	170	232	260

Table.2 Query based browsing
Based on the user query the related links are displayed. Normally the query is in the form of keywords as in Fig.2.

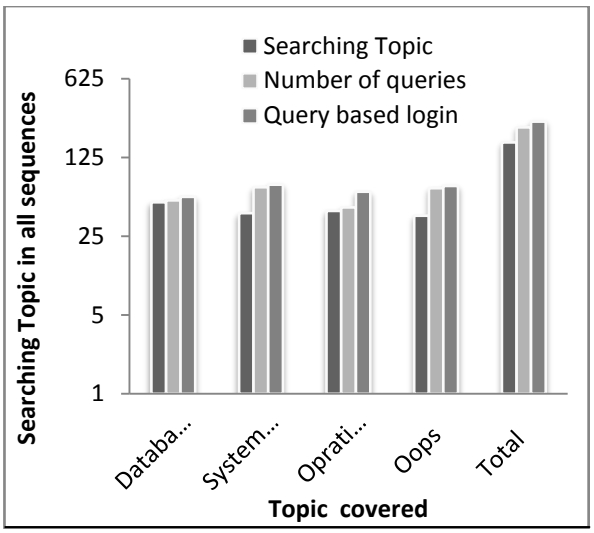

Fig.2 Query based browsing

The keywords which match with the sequence are indexed as in Fig.3. The lexical meaning of the keyword is analyzed through Word Net. Tree Tagger is for annotating text with part of speech and lemma information. Preprocessing or Tokenization is the process of breaking a stream of text up into words phrases, symbols or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing.

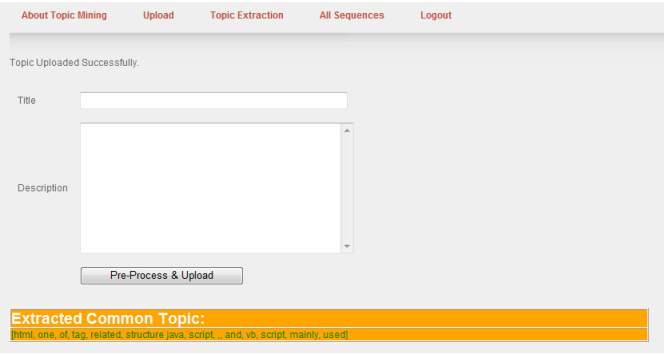

Fig.3 Home page

The related documents are captured and display. The similarity between the topic sequences is calculated. The documents with more similarity measures are clustered.

VII. CONCLUSIONS

A novel method was introduced to deal with and it automatically discovers and fix potential asynchronism among sequences and consequentially extract better common topics.

The proposed method is used by utilizing correlation between the semantic and temporal information in the sequences. It performs topic extraction and time synchronization alternatively to optimize a unified objective function. A local optimum is guaranteed. Preventing duplications in text sequences considering similarities according to temporal analysis is a constrain proceed further.

1) The method is able to find meaningful and discriminative topics from asynchronous text sequences;
2) The performance of our method is robust and stable against different parameter settings and random initialization.

Acknowledgment

The authors are heartily thankful to their supervisor, Mrs. D. Saravanpriya Department of information technology, P.A college of engineering and technology, whose sprit to work, guidance and support from the initial to the fine level enabled them to develop an understanding of the subject. Above all and the most needed, she provided them unflinching encouragement and support in every possible ways.

Finally, the authors would like to thank everybody who were important and helped them out in every process to the successful realization of the thesis.

REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko, '*On-Line New Event Detection and Tracking*', Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 37- 45, 1998.

[2] A. Asuncion, P. Smyth, and W. Welling, '*Asynchronous Distributed Learning of Topic Models'*, Proc. Neural Information Processing Systems, pp. 81-88, 2008.

[3] D. J. Berndt and J. Clifford, '*Using Dynamic Time Warping to Find Patterns in Time Series'*, Proc. Knowledge Discovery in Databases (KDD) Workshop, pp. 359-370, 1994.

[4] D. M. Bleiand J. D. Lafferty, '*Correlated Topic Models'*, Proc. Neural Information Processing Systems, 2005.

[5] D. M. Blei and J. D. Lafferty, '*Dynamic Topic Models'*, Proc. Int'l Conf. Machine Learning (ICML), pp. 113-120, 2006.

[6] D. M. Blei, A. Y. Ng., and M. I Jordan., '*Latent Dirichlet Allocation'*, Proc. Neural Information Processing Systems, pp. 601-608, 2001.

[7] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, '*Parameter Free Bursty Events Detection in Text Streams'*, Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 181-192, 2005.

[8] T. Hofmann, '*Probabilistic Latent Semantic Indexing'*, Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 50-57, 1999.

[9] A. Krause, J. Leskovec, and C. Guestrin, '*Data Association for Topic Intensity Tracking'*, Proc. Int'l Conf. Machine Learning (ICML), pp. 497-504 , 2006.

[10] W. Li and A. McCallum, '*Pachinko Allocation: Dag-Structured Mixture Models of Topic Correlations'*, Proc. Int'l Conf. Machine Learning (ICML), pp. 577-584, 2006.

[11] Z. Li, B. Wang,M. Li, andW.Y. MaY, '*A Probabilistic Model for Retrospective News Event Detection*', Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 106-113, 2005.

[12] Mei.Q, Liu.C, Su.H, and Zhai.C (2006), '*A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs'*, Proc. Int'l Conf. World Wide Web (WWW), pp. 533-542.

[13] X. Wang. and A. McCallum, '*Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends'*, Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 424- 433, 2006.

[14] X. Wang, C. X. Zhai,X. Hu and R. Sproat, '*Mining Correlated Bursty Topic Patterns from Coordinated Text Streams'*, Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 784-793, 2007.

[15] C. Zhai, A. Velivelliand B. Yu, '*A Cross-Collection Mixture Model for Comparative Text Mining'*, Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 743-748, 2004.