|| Volume 2 || Issue 6 || December 2017|| ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

Time Stamp Based Topic Mining Over Text Sequences

Ms. Rohini M.Gujar

P.G Student, Computer Science & Engineering, C.S.M.S.S College of Engineering, Aurangabad, Maharashtra, India

rohinigujar14@gmail.com

Abstract— **Text are scattered and spread across in different documents with different timestamps shared messages, general topics, and other. They have a relationship with the content. The content of the message may be related to other documents of topics, but with different timestamps. Interactions between general topics may receive valuable information. However, it may not be prepared in an indexed format, because there is a difference in time. The main goal of this paper is to isolate common-topic mining with the help of the timestamp generator model, which will of course perform two major tasks. Extraction of general topics from text sequences documents by adjusting the time stamps. The timestamp is based on the time distribution of the general topic created previously. These steps will work or retrieve general topic information.**

Keywords: *Text sequences, Topic mining ,Topic model, word distribution, time distribution.*

I INTRODUCTION

In today's world, text sequences are created in a variety of ways, such as streams, news, email, social media, research, weather forecasts, etc., to gain valuable insights from timestamp sequences that share common topics. Meaningful is temporary in different sequences, different data is stored at different intervals. The semantic and temporal information are related to documents and both of the information described in two distribution like word distribution and time distribution [1] [2]. We will combine these data and sequences to create a better understanding. To isolate general topics, use the baseline method. Increasing the problem of detecting topics is an important part of text mining [3]. An example of text sequence is research paper repository. The goal of the paper is to examine a series of research paper about the topic. It was viewed as a natural text stream containing the publication date as time stamp. A discovering, identifying, and abstracting automatic theme is very useful. This algorithm has many interesting applications that can make it easier for people to understand and manipulate information contained in large knowledge domains, including exploring topic changes and identifying the role of words. Domain applications find text in document format with meaningful time stamps. Events covered in research papers are generally temporary and evolutionary, with topics that point to the beginning of the breakthrough and the impact of events among others. It categorizes documents as topics and actions into activities. Text mining, also known as knowledge discovery from its original database, refers to the process of extracting interesting and unintelligible patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from a database. Text mining can be seen as consisting of two phases.

Refining text that transforms text documents into free forms becomes a selective medium and refining knowledge that assumes a learning pattern or knowledge of the central form [4]. For example, in the research paper archives, there is no guarantee that research topic covering the same topic will be indexed with the same timestamps. There may be a delay time for news communication and newsletters, weeks or even months of journals. Full version of paper may be appearing in conference. The paper usually published in months, annually and last in journals and which may be sometimes taking more than a year to appear after submission. Another example of timestamp text sequence is news articles [5]. Different sequences are different for the same topic at different times.

II RELATED WORK

In many real applications, text collections are generally temporary, so they are treated as a sequence of messages. Various methods have been proposed to search for topics over time in sequence [6]. However, these methods are designed to separate threads from a single sequence. Asynchronous in several sequences, such as documents from different sequences on the same topic, have different time stamps, it is very common to assume that the document received in the order of time stamp of the document is created according to the terms of the word [7]. We introduced the hyper parameters that evolve indefinitely in the form of a transfer of status sequentially.

For each piece of time, the hypermeter is assigned to the state by the probability distribution to the state in the original time piece. The time sequence of the sequence is cut into pieces, the time and the topic are independently found in the documentation [8].So in many cases the sequence of topics in each sequence can be estimated separately and the

relationships that may occur between threads in different order are both semantic and temporal. We also know that there is a whole literature about measuring the similarity between time series. There are several similarity function functions, many of which refer to asynchronous features between sets of data [9]. The main symbols used throughout the paper are listed in Table 1. However, the asynchronous assignment solves the problem. In fact, most of the similarity measures deal with asynchronous defaults, as shown in Figure 1.

Table1: Symbol with their consecutive meanings

Symbol	Description
D	Documents
T	Timestamp
W	Word
Z	Topic
M	Number of Sequences
V	Number of Distinct Words
K	Number of Topics

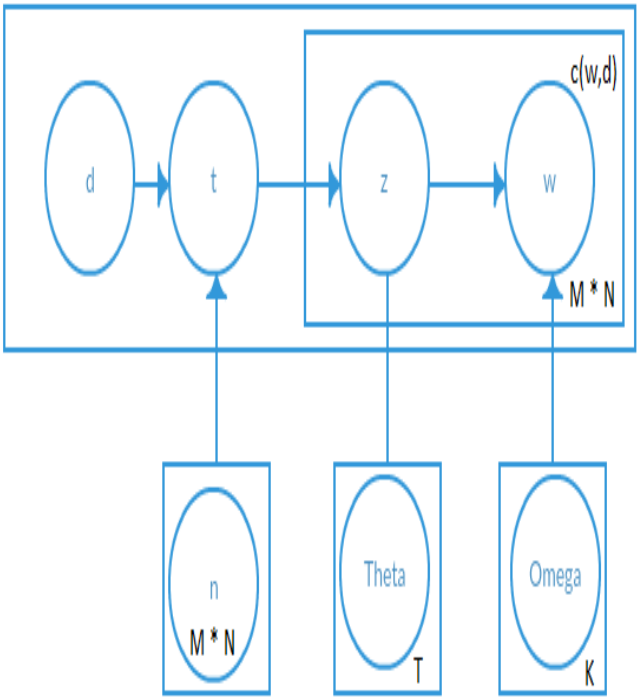

Figure 1: Generative Model

We can define three definitions that attempt to resolve the primary purpose of data mining for asynchronous messages:

Definition 1 (Text Sequence) - S is the sequence of documents N (d1,d2,…, dn). Each document d is a set of terminology V and is indexed by a unique timestamp t € {t1,t2, ...tn}. Here we deal with real applications, which allow multiple documents in the same order to share time.

Definition 2 (General Topics) - General Topics Ƶ the above sequence of messages is determined by the distribution of the term V and the time distribution through time stamps (1, ..., T).

Definition 3 (Asynchronism) - Specifies the order of the messages M (Sm: 1 <m ≤ M}, where the document is indexed with timestamp {t: 1 ≤ t ≤ T}. Same topic Different sequences are incorrectly aligned [9].

III PROPOSED ARCHITECTURE

The standard baseline method is the extraction step of our algorithm. However, in our experiments, we have introduced two additional techniques to use the modified baseline algorithm as the basis for extracting the topic [10]. The first technique is to introduce background topics in our empirical models so that they can remove noise in the background and search for meaningful and interesting topics. Quantitative estimates of sequence mismatches are available and no need to search for all dimensions. It gives the opportunity to simplify the synchronization process over time.

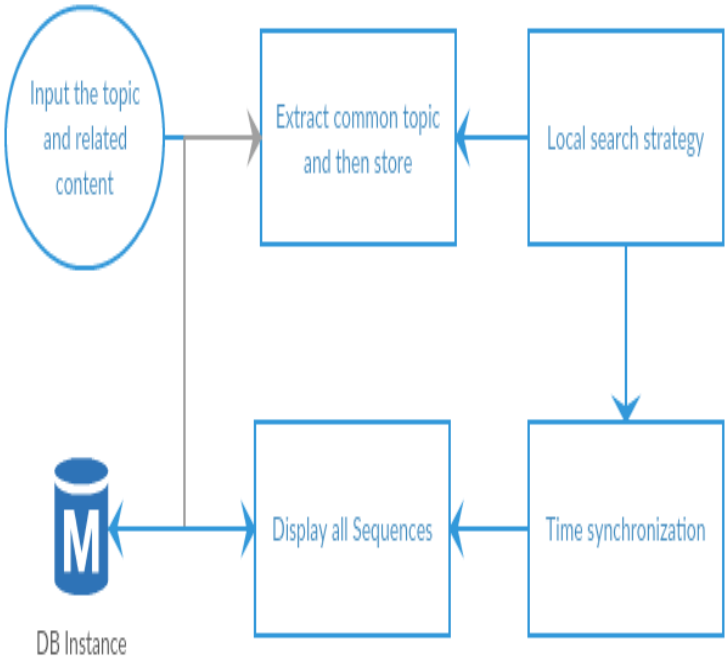

Figure 2: Baseline Architecture

Resolve this issue formally and propose a novel algorithm based on generative topic patterns. Our algorithm consists of two other steps. The first step separates common topics from several sequences based on the time set in the second step. The second step adjusts the time stamp of the document according to the distribution of the topic discovered by the first step.

A. Topic Extraction

The different elements present in sources of information are text analytics solution of topic extraction. This detection process is carried out by combining a number of complex natural language processing techniques that allow

obtaining the syntactic and semantic analyses of a text and using them to identify different types of significant elements. The initial values of timestamps and objective function are counted from the original timestamps in the sequences [11]. A document is seen as a mixture of topics. In this we will only cover analyses of texts. Words terms are the basic unit of data in a document. The set of all words in a corpus is called a vocabulary. Splitting up a document into words is referred to as tokenization. We assume that our current timestamp of sequences are already synchronous and extract common topics from them. Our algorithm is summarized as K is the number of topics specified by user.

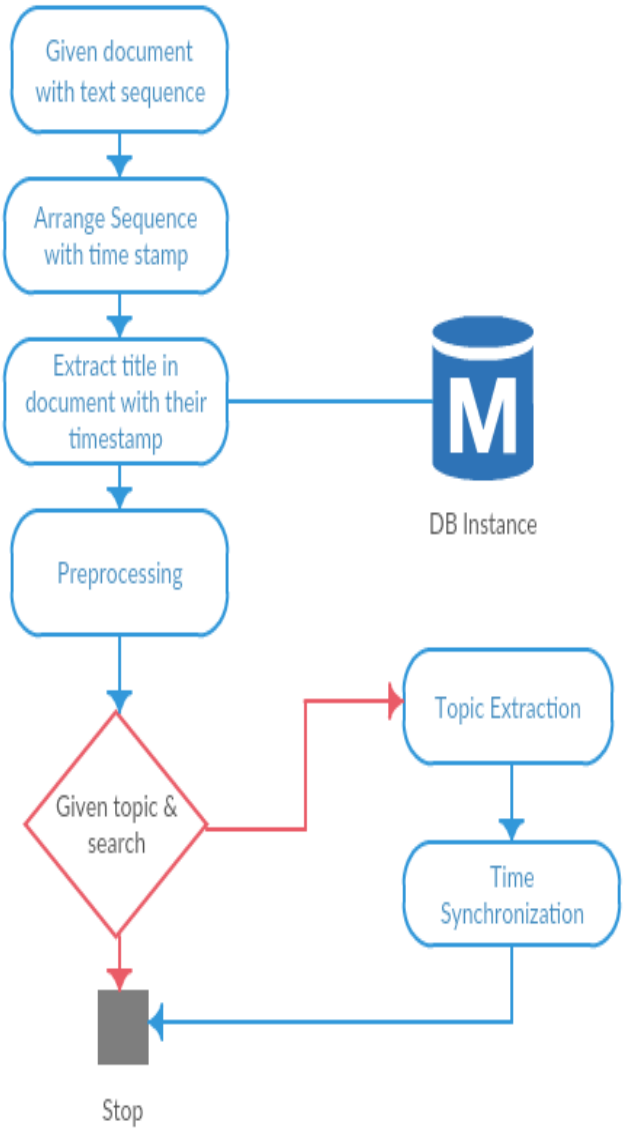

Figure 3: Proposed Architecture

B. Algorithm

Mining topics by time synchronization
Input: K, Timestamp, Objective function
Output: Word, Topic, Timestamp
Repeat
Update words using the timestamp and objective functions.
Beginner: Title values and words with random numbers.
Repeat
Update word and topic values until convergence
For m = 1 to M do (M no steps)
For u = 1 to T do initialize function function.
For v = 2 to t do for w = 1 to t
Calculate the objective function.
End
End
Update timestamp
End
To Convergence
Above algorithm can be explained as
Step 1: Enter a topic or document.
Step 2: Provide content related to the topic.
Step 3: Use pre-processing status.
Step 4: View all related headlines and pull general topics using topic headings.
Step 5: Search the search string and select the content that is relevant to the search.
Step 6: Retrieve general topics to display.
Step 7: When dividing the common topics, match the documents in all order and then show the synchronized order.
Step 8: Get document content from unstructured text sequence.

This assumption is based on observations from real-world applications, such as the research paper archives by various newsletters, may vary in exact timestamps, but sequential data is consistent with the sequence of events. We argue that the second option works better in practice because real world data sets are incomplete. Although we assume that the sequential pattern of the trunk typically, there are a few documents to do. Our reworked processes and relaxed restrictions will allow for the recovery of this type of data.

IV EXPERIMENTAL RESULTS

Dataset used in the operation are research papers extracted from the DBLP archives, including SIGMOD, TKDE (Journal), VLDB and World Wide Web Journal from https://hpi.de/naumann/projects/repeatability/datasets/dblp-dataset.html.

This archive mainly consists of research on database technology. Each repository is treated as a number of documents, each of which is represented by the title of the paper and the time stamped by the publication year. Separate number of documents from different sequence of titles with time stamps in the database .For each pair of papers, the corresponding work is to decide whether the two are written by the same author.
Preprocessing plays an important role in data mining techniques and applications. The necessary step is to separate the document into one word. This is called text segmentation (or word) or tokenization. Detect the punctuation, numbers, spaces, and stops words from documents. Converts word in to lowercase and

different token marks match the dictionary. Rarely used words in analysis can be tagged and deleted as stop words that are usually displayed in the stop list. Stop words being part of the natural language. Motivation to stop a word should be removed from the message, making the titles look heavy and unimportant for the analyst. Deleting a word stops the dimension of the word space. The most common words in a text document are compound words and proxies, which do not give the meaning of the document. These words are considered a stop word. Stop words: in, a, a, with, etc. Stop words being deleted from the document because those words are not being measured as keywords in the text mining applications.

Topics used to search or browse the source data. Search for common topics from different sequences and synchronize them. Based on this evidence, documents with the same timestamp are combined through different sequences so that the distribution of each topic's title can be found. Timestamp topic extraction is using how much time discuss in particular year. Once the common topics are extracted, we match topic in other sequences to their documents and adjust their time stamps to synchronize the sequences. In time synchronization consists of different step. The previous step is given the topic with topical word in different sequence. The topical word check in other sequences document. Time synchronization is selects document and extracts them in new sequence. Meaning, while having asynchronous time distribution in a different order and arrange synchronous manner. Based on the user query the related links are displayed. Normally the query is in the form of keywords as like a topic. Topic is search in documents at the same timestamp documents from database. Give the result shown in table 2 by extract the common topic in multiple text sequence. Draw the figure 4 given value of the table. Another table 3 is giving the total topic in particular year and arrange in one sequence for the user. Show the result of system in figure. Draw the figure 4 given value of the table.

Table 2: Title containing words in particular timestamp

YEAR	Word1	Word1+ Word2	Word1+ Word2+Word3
2008	15	10	7
2009	17	9	6
2010	14	11	5
2011	18	13	8
2012	20	10	5

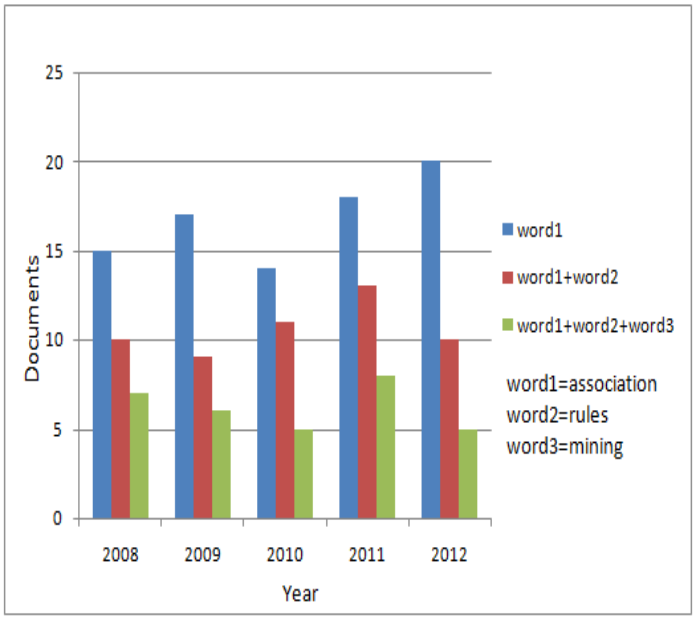

Figure 4 :Title containing words in particular timestamp titles.

Table 3: Total titles of word which present in given timestamp

Year	Count of title
2008	10
2009	13
2010	14
2011	11
2012	16

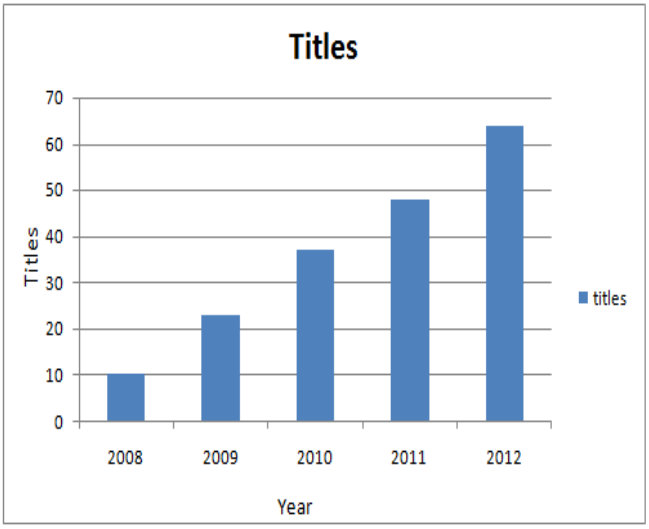

Figure 3: Total titles of word which present in given timestamp

V CONCLUSION

In real world mining, general topics with timestamps vary in different order. New approaches have been introduced to automate the management and discovery and resolution of asynchronism issues that may arise between better sequences and headlines. The proposed method is used in the relationship between semantic information and time in sequence. To handle this, the first two operations will be able to distinguish common topics from the document set and can be adjusted in the second step. The second step adjusts the time stamp of the document according to the time distribution of the topics discovered by the first step. It performs topic extraction and timing synchronization, or to optimize the full-function objective function. This method can search for meaningful and inappropriate topics from a sequence of matched text.

REFERENCES

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

[2] 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.

[3] X. Wang, C. 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.

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

[5] Z. Li, B. Wang, M. Li, and W.-Y. Ma, "*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.

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

[7] 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.

[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] Xiang Wang, Xiaoming Jin, Meng-En Chen, Kai Zhang, and Dou Shen "Topic Mining over Asynchronous Text Sequences", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[10] 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

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