IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Topic Quarrying Over Same Time Period Text Documents

Sujata B. Sanap[1], Prof. Vivek P. Kshirsagar[2]

Computer Science & Engineering Department, Government College of Engineering, Aurangabad, India[1,2]

Abstract: Texts are spread over in different database having different timestamps. Text share common topics and thus related with each other. They are correlated with respect to the content they have. The content of text may be related with each other with common topics but have different time stamp. The interaction between common topics may derive valuable information but they may not be arranged in indexed fashion as they differ in timestamp. The main goal of this paper is to extract common topic mining with the help of generative model using exact timestamp. It will perform two main operations alternatively. Common topic extraction with adjusted timestamp and next is adjusting the time stamp according to time distribution of the common topics generated previously. These steps will work alternatively and extract the information of common topics.

Keywords: Timestamp, Asynchronous Sequences, Topic mining, Text mining, Correlation.

I. INTRODUCTION

In the today's world the text sequences are being generated in different forms such as news streams, emails, social media, research paper repository, weather forecast etc. To extract the valuable knowledge from time stamped text sequences which share common topics with semantic as well as temporal information. In different sequences different data is stored at different timings. We will combine this data and sequences to produce informative knowledge. To extract common topics PLSA method is used [1].Addressing the problem of topic detection is the main focus in text mining. The goal of the tasks is to detect a collection of news articles about a topic. It is viewed as the natural text streams with publication dates a time stamps. It would be very useful to discover, extract and summarize the evolutionary theme patterns automatically. The algorithm contains several interesting applications that can make it easier for people to understand and the information contained in large knowledge domains including exploring topic dynamics and indicating the role that words play in the semantic content of documents. Application domains, encounter a stream of text document has 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.

Text mining, also known as text data mining or knowledge discovery from textual databases, refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from databases. Text mining can be visualized as consisting of two phases. Text refining that transforms free- form text documents into a chosen intermediate form, and knowledge distillation that deduces patters or knowledge from the intermediate form [2] .Addressing the a synchronism among multiple streams i.e.documents from different streams about the same topic have different timestamps, is actually very common in practice. For instance, in news streams, there is no guarantee that news article covering the same topic are indexed by the same timestamps. There can be hours of delay for news agencies, Days for newspapers and even weeks for periodicals [6]. This is because some news feeds try to provide first hand ashes shortly after the incidents, while others provide more comprehensive reviews afterwards. Thus different sequences are sharing same topic but in different timestamp. So one can consider timestamp in synchronous way and extract information related to data mining which is informative than individuals. [7]

This paper is addresses in different sections including 2.Liturature survey, 3.Proposed algorithm, 4Conclusion and 5.Acknowledgement.

II. LITERATURE SURVEY

In many real applications, text collections carry generic temporal information, thus can be considered as text sequences. To capture the temporal dynamics of topics, various methods have been proposed to discover topic over time in text sequences.[3].However ,these methods were designed to extract topics from single sequence. The asynchronous among multiple sequences i.e. Documents from different sequences on the same topic have different timestamp, is actually very common.It was assumed that given a document in the sequence, the timestamp of the document was generated conditionally independently from word. We introduced hyper-parameter that evolves over time in state transfer models in the sequence [8].

For each time slice, a hypermeter is assigned with a state by a probability, distribution, given the state on the former time slice.

The time dimension of the sequence was cut into time slices and topics were discovered from documents in each slice independently. [4]

DOI 10.17148/IJARCCE.2016.5225

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

TABLE I SYMBOLS AND THERE MEANING

Symbol	Description
D	document
T	timestamp
W	word
Z	topic
M	number of sequences
T	length of sequences
V	number of distinct words
K	number of topics

As a result, in multiple-Sequence cases, Topics in each sequence can only be estimated separately and potential correlation between topics in different sequences, both semantically and temporally. We also note that there is a whole literature on similarity measure between time series. Various similarity functions have been proposed, many of which addresses the asynchronous nature between times series. The main symbols used throughout the paper are listed in Table1. However, defining as asynchronous solves the problem in fact, most of the similarity measures deal with asynchronous implicitly 5 as shown on Figure 1

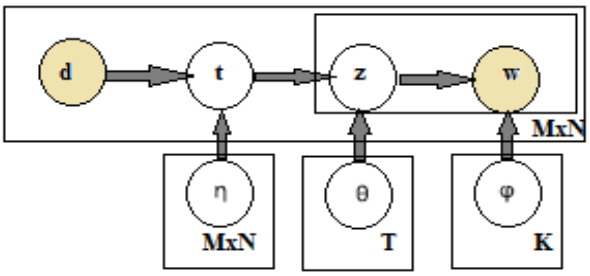

Fig.1 Illustration of Generative Model

We can define three definitions which will try to solve our main objective of common topic mining for asynchronous text as follows.

Definition 1(Text Sequence) – S is a sequence of N documents (d1,….,dn).Each document d is a collection of words over vocabulary V and indexed by a unique timestamp t € {1,….N}.
Here we are relating it to real world application which allows multiple documents in the same sequences to share common timestamp.

Definition 2 (Common Topic) – A common topic Ƶ over text sequences is defined by a word distribution over vocabulary V and a time distribution over time stamps (1,…,T).

Definition 3 (Asynchronism) – Given M Text sequences {Sm : 1 < m ≤ M},in which documents are indexed by timestamp{t : 1 ≤ t ≤ T}.Asynchronism means that the timestamp of the document sharing the same topic in different sequences are not properly aligned.[1][2]][9][10]

III.PROPOSED ALGORITHM

Formally addressed this problem and put forward a novel algorithm based on the generative topic model. Our algorithm consist 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 document according to the distribution of the topics discovered by first step.

The standard PLSA method is the extraction step of our algorithm. Yet in the experiment 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. The first technique is to introduce a background topic into our generative model so that background noise can be removed and find burst and meaningful topics. 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 adjusting in each iteration[1][4][5].

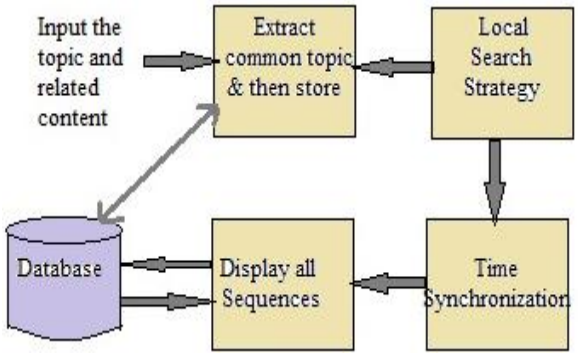

Fig.2 Baseline Method

A. Topic Extraction
We assume the current timestamp of all sequences are already synchronous and extract common topics from them. Our algorithm is summarized as K is the number of topics specified by user. The initial values of timestamps and objective function are counted from the original timestamps in the sequences. [11]

B. Algorithm
Topic mining with time synchronization
Input : K, Timestamp, Objective function
Output : Word, Topic, Timestamp
Repeat
Update word with timestamp and objective function
Initialize: Topic and word values with random numbers
Repeat
Update word and topic values until convergence
For m=1 to M do (M is no of steps)
For u=1 to T do initialize objective function
For v=2 to T do

DOI 10.17148/IJARCCE.2016.5225

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

For w=1 to T do Compute objective function
End
End
Update Timestamp
End
Until Convergence

Above algorithm can be explained as
STEP 1: Input to the topic or document.
STEP 2: Give the topic related content.
STEP 3: Use pre-processing state.
STEP 4: Get topic related content, all sequence already synchronous and extract common topic using topic extraction.
STEP 5: Give the searching string and then pick the searching related content.
STEP 6: Extract common topic to be displayed
STEP 7: Once the common topics are extracted, Match the document in all sequences and then display synchronized sequence.
STEP 8: Get document content from unstructured text sequence.

This assumption was based on observations from real-world applications like news stories published by different news agencies may vary in absolute timestamps, but their sequential information conforms to the order of the occurrences of the events [13].

We argue that the second option works better in practice since real –world data set are not perfect [15].Although we assume that that sequential format of the given sequences is correct in general, there will still be a small number of documents that do not conform to our assumption. Our iterative updating process and the relaxed constraint will help recover this kind of outlying.

C. The Local Search Strategy
In some real-world applications, we can have a quantitative estimation of the asynchronism among sequences so it is unnecessary to search the entire time dimension when adjusting the timestamp of documents [14].This gives us the opportunity to reduce the complexity of time synchronization step without causing substantial performance loss, by setting a upper bound for the difference between the timestamp of documents before and after adjusting in each iteration. Specifically, given document D with time T, we now look for an optimal topic function within the neighbourhood of topic.

The proposed method is used by utilizing correlation between semantic and temporal information in sequence. It performed topic extraction and time synchronization alternatively to optimize a unified objective function. A Local optimum is guaranteed. Preventing duplication in text sequences considering similarities according to temporal analysis is a constraint proceeds 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.

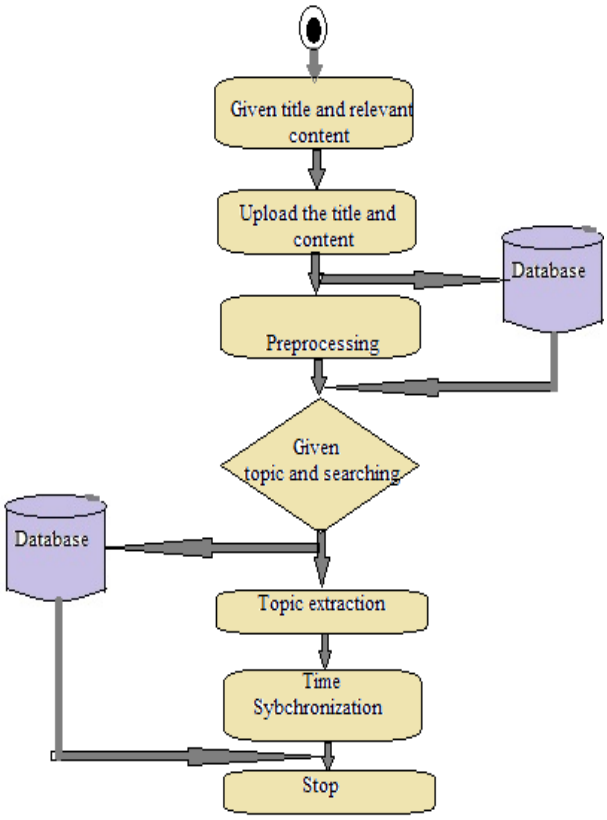

Fig.3 Query Based Browsing

IV. CONCLUSION

In real world, mining common topics having different timestamps is a tricky way. Thus this problem over multiple asynchronisms is tackled. In this we proposed a novel method which automatically discover the common topic and potential asynchronism among streams and consequently extract better result improving the informative discovery of knowledge. The base idea behind this method is to co-relate the topics sharing same information but spread over different timestamps with multiple sequences. It performs topic extraction and time synchronization alternatively to optimize a unified objective function. This effectiveness is proved on date sets with comparison to a baseline method .1.Our method is able to find meaningful and discriminative topics from asynchronous text streams.2.It significantly outperforms the baseline method ,evaluated in quality and quantity and 3.the performance of our method is more stable. As quality and quantity is maintained by our method so performance measure is also strong and healthy.

ACKNOWLEDGEMENT

The author are heartily thankful to their Guide, Prof. Vivek P. Kshirsagar ,Head of Computer Science and Engineering Department, Government College of Engineering, whose spirit 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, he provided them encouragement and support in every possible ways.

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

REFERENCES

[1]. T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 50-57, 1999.
[2]. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: An Overview. In Advances in Knowledge Discovery and Data Mining , U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., MIT Press, Cambridge, Mass., 1-36.
[3]. 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.
[4]. 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.
[5]. H. Sakoe, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-26, no.1, pp. 43-49, Feb. 1978
[6]. 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
[7]. Q. Mei and C. Zhai, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 198-207, 2005.
[8]. R.C. Swan and J. Allan, "Automatic Generation of Overview Timelines," Proc. Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 49-56, 2000.
[9]. 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.
[10]. T.L. Griffiths and M. Steyvers, "Finding Scientific Topics," Proc. Nat'l Academy of Sciences USA, vol. 101, no. Suppl 1, pp. 5228-5235, 2004.
[11]. 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.
[12]. Z. Li, B. Wang,M. Li, and W.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.
[13]. 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.
[14]. Yanpeng L, Xiaohua Hu, Hongfei Lin and Zhihao Yang, "A Framework for Semisupervised Feature Generation and Its Applications in Biomedical Literature Mining", IEEE, Vol: 8, Issue: 2, 10.1109/TCBB.2010.99, 2011, Page(s): 294 – 307 Applications in Biomedical Literature Mining", IEEE, Vol: 8, Issue: 2, 10.1109/TCBB.2010.99, 2011 , Page(s): 294 - 307
[15]. Navigli, R., and Velardi, Paola, "Structural semantic interconnections: a knowledge-based approach to word sense disambiguation", IEEE, Vol27, Issue: 7, 10.1109/TPAMI.2005.149, 2005, Page(s): 1075 – 1086

DOI 10.17148/IJARCCE.2016.5225