International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Topic Mining for Time Stamped Text Document Using PLSA Method

[1*] Deepali Panditrao Shelke, [2]Prof.NitinM.Shahane
[1*]*Department of Computer Engineering, PG Student, K.K.wagh Institute of Engineering Education & Research, Nasik, India,deepalishelke86@gmail.com*
[2]*Department of Computer Engineering, Faculty of, K.K.wagh Institute of Engineering Education & Research, Nasik, India,nm_shahane@yahoo.co.in*

Abstract- In the real-world applications, time stamped document or text sequences, are begin anywhere. Various text sequences are related to each other by sharing common topics. The correlation between these sequences provides more meaningful and complete clues for topic mining than those from each particular sequence. so it shows the interrelationship between the different documents with the existence of asynchronism.. i.e., documents from different sequences about the same topic may have different time stamps. Our algorithm consists of to extract the co-related common topics from the given documents using PLSA method. The advantage and efficiency of our approach were justified through empirical studies on real data set consisting of six research paper repositories.

Keywords: Asynchronous text sequences, word distribution, time distribution, Topic mining, Temporal text mining

I. INTRODUCTION

In the today's world the text sequences are being generated in different forms such as news streams, emails, research paper repositories etc.,To extract the valuable knowledge from time stamped text sequences that are sharing common topics with semantic as well as temporal information. In different sequences there are data Stored in different timestamp, we combine the all sequences and produce more informative data than individual sequence. To extract the common topics from different sequences using PLSA method.

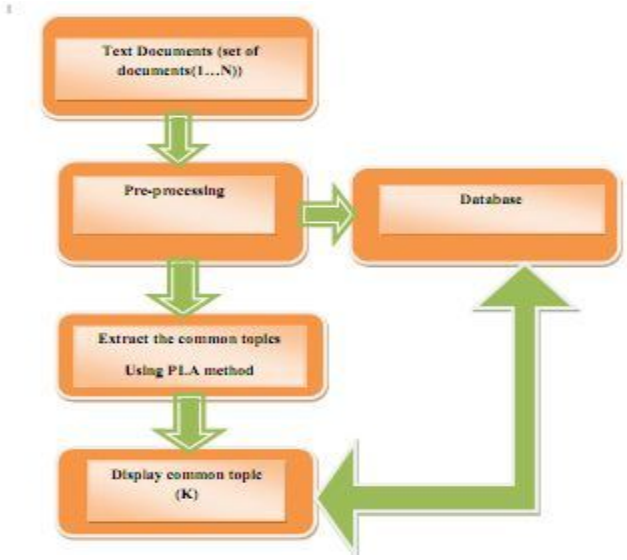
In the different text sequences information are correlated with each other by sharing common topics. Such as two repositories are IEEE and ACM .IEEE having the topic is data mining from year 2005 to 2010, Database related journal and ACM has similar topic from 2011 to 2015, which are share common topic by two sequences. These are the different sequences share same topic but in different timestamp so, we consider as timestamp in synchronous way, then extract the information related to data mining that is more informative than individuals.

II. LITERATURE SURVEY

The amount of electronically available information is rapidly-growing which threatens to overwhelm attention of human, raising new challenges for information retrieval technology. Tradionoal query driven approach may not help to find out "what it happened?" Browsing without any helps to find out information on minimal scale. So, it is required that a new tool should automatically organize, search, index and should browse from large collection. Many approaches have studied mine topic considering its time frame In[3], a dynamic topic model is developed which captures the evolution of

topics in a sequentially organized corpus of documents. The approach used is to use state space models on the natural parameters of the multinomial distributions that represent the topics.
In the other study [4], a new problem is interrogated and named as hot bursty events detection in a text stream, where a text stream can be a sequence of documents which are ordered chronologically, and a hot bursty event is a minimal set of bursty features that occur together in certain time windows with strong support of documents in the text stream. This work focused on detecting a set of bursty features for a bursty event. A new novel model is explored as parameter as probability approach, which is named as feature-pivot clustering. They have utilized the time information to the fullest to find a set of bursty features which may occur in different time. For this there is no need to tune or estimate any parameters.
In another paper, TDT (topic detection and tracking) is contained which attempts to cluster documents as events using clustering techniques. This study having facility to the problem of extracting important topics and events from a stream of news articles, that process the type of document stream analysis questions considered here. Much of the emphasis in the TDT study was on techniques for the on-line version of the problem, in which events must be detected in real-time; but there was also a retrospective version in which the whole stream could be analyzed [2].
Temporal Text Mining (TTM) is concerned with discovering temporal patterns in text information collected. Most of text information having time stamps, temporal text Mining has many applications in various domains, like summarizing events in news articles and revealing research trends in scientific literature. In many application domains, a stream of text is encountered where each text document has some meaningful timestamp [1],[2], [5]. Text mining work in this area is following intuitive premise that the appearance of a topic in a document stream is signaled by a "the burst of activity". In this features are rising accurately in frequency as the topic is emerged. Documents can be naturally organized by topic, but in many settings we also experience that their arrival over time plays a important role. News articles and E-mails are two such clear examples of this kind of documents so, the strong temporal ordering of the content is necessary.
Another approach combines an extension of Factorial Hidden Markov models for topic intensity tracking with exponential order statistics for implicit data association. And they have demonstrated the use of a switching Kalman Filter to track content evolution of the topic over time. This approach is general in the sense that it can be combined with a variety of learning techniques. It can be applied in supervised and unsupervised settings.[6] Mining subtopics from weblogs and analyzing their spatiotemporal patterns have applications in multiple domains. In the work done by[6], 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 simultaneous is defined.
The Event Detection and Tracking problems are part of a broader initiative called Topic Detection and Tracking (TDT). New event detection is an abstract document classification task that has reasonable solutions using a single pass clustering approach. An approach is presented in [2] which is evaluation methodology based on miss and false alarm rates. These were used to measure detection error in a cross-validation approach. Overall system performance is specified using a bootstrap method that produced performance distributions for the TDT corpus. Here, the emphasis was given on time as well as event instead of just the 'topic'. So, it is expected that multiple correlated sequences will facilitate topic mining by generating topics with higher quality. However, the asynchronism among sequences brings new challenges to conventional topic mining methods. Therfore a method should be able to handle this asynchronism as well. [1][2][3][4][6].

III. SYSTEM ARCHITECTURE



IV. METHODOLOGY

PLSA

LSA- aims to discover something about the meaning behind the words, about the topics in the documents. Difference between topics and words is that words are observable, but topics are latent. Latent Semantic Analysis is method to perform the automatic indexing and information retrieval that achieve to defeat these problems by mapping documents as well as terms to a representation is called latent semantic space. The LSA has implemented to address automated document indexing. It has two concepts.

- .Dimensionality Reduction

A set of documents that is usually represented in the form of a document-term matrix (this gives the number of times each word appears in each document) Represent each document by a high-dimensional vector in the space of words so dimension reduction means it represents the high dimension vector space into lower dimension vector space.Dimensionality reduction of word-document co-occurrence matrix. Keep the K – largest singular values which show the dimensions with the greatest variance between words and documents discarding the lowest dimensions is supposed to be equivalent to reducing the "noise" Terms and documents are converted to points in a K dimensional latent space.

- Construction of Latent Space

It constructs the clusters according to the document –term matrix. This gives the each word and its term frequency in one cluster. The problem addressing in the LSA is that synonymy such as buy-purchase and Polysemy such as book (verb) - book (noun) LSA may classify documents together even if they don't have common words, so to overcome these problems it required PLSA.Probabilistic Latent semantic analysis is the statistical method for the analysis of co-occurrence data. PLSA is a technique from the classification of topic models. Queries can be more reliably estimated in the reduced latent space representation than in the original representation.

Topic Extraction

1. Pick a document with probability p(d).
2. Given the document d, pick a common topic z with probability (z/d).

3. Pick a word w with probability (z/w).

4.Pick a probability of topic z, p(z).

The probability of word in document d is

$$P(w,d) = \sum p(z)\, p(z/d) p(z/w)$$

Log-likelihood function over all sequence is

$$£ = \sum \sum c(w,d) \log p(w,d),$$

Where c(w,d) is the number of occurrences of w in document d.

Table 1: Symbols and Their meanings

Symbols	Description
d	document
w	word
z	topic
K	Number of topics
N	Number of documents

IV. ALGORITHM

Step-1 Input to the set of text documents.(1….N) how many topics required which is given by user.

Step-2 Using preprocessing state produces vocabulary.

Step- 3Using Probabilistic Latent Semantic Analysis technique, to extract co-related topics in one Cluster according to their meanings.

In PLSA requires following steps:

3.1 To initialize Data structures.

3.2 Calculate the probability of word in document d is

$$P(w,d) = \sum p(z) p(z/d) p(z/w)$$

Where, p(z) - Probability of topic z.

p(z/d) - Probability that topic z present in document d.

p(z/w) –Probability that topic z associated with word w.

3.3 Calculate Log-likelihood function over all sequence is

$$£ = \sum \sum c(w,d) \log p(w,d)$$

Step-4 To perform estimation maximization (EM) method.

Step-5 To perform normalization concept until it gives monotonic output.

Step-6 To produce PLSA position of Words.

Step- 7 Display the common topics (K).

CONCLUSIONS

Topic mining is a field of text mining. The problem of mining common topics from different text sequences using novel method Multiple text sequences are often related to each other as they share the common topics among them. The proposed method will help us in dealing with the problem of asynchronism as real world data cannot have common topics sharing the same time distribution over different sequences. In practice, documents from different sequences on the same topic have different time stamps. To deal with this problem, first t he common topics can be extracted from the set of documents and then the time stamps can be adjusted to make them synchronous. Hence the problem of mining common topics from multiple asynchronous text sequences can be tackled with the help of proposed method by using a self refinement process by utilizing co -relation between the semantic and temporal information in the sequences. The method can be applied on real world data sets to evaluate its effectiveness.

REFERENCES

[1] 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, pp.156 -169.
[2] 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
[3] D.M. Blei and J.D. Lafferty, "Dynamic Topic Models," Proc. Int'l Conf. Machine Learning (ICML), pp. 113-120, 2006.
[4] 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.
[5] J.M. Kleinberg, "Bursty and Hierarchical Structure in Streams," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 91-101, 2002.
[6] A. Krause, J. Leskovec, and C. Guestrin, "Data Association for Topic Intensity Tracking," Proc. Int'l Conf. Machine Learning (ICML), pp. 497-504, 2006.
[7] 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.
[8] Q. Mei, C. Liu, H. Su, and C. Zhai, "A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs," Proc. Int'l Conf. World Wide Web (WWW), pp. 533-542, 2006.
[9] Q. Mei and C. Zhai, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining," Proc. ACM 0SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 198-207, 2005.
[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. 424433, 2006.
[11] 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.
[12].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.
[13] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. Ann.Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 50-57, 1999.
[14] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," Proc. Neural Information Processing Systems, pp. 601-608, 2001.

