Topic Modeling over Asynchronous Text Sequences

[1]M. Divya, [2]Dr. S. Chitrakala
[1]Anna University, [2] AnnaUniversity,Chennai

***Abstract* - Topic models have been widely used to identify topics in text corpora. Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts. They provide a simple way to analyze large volumes of unlabelled text. A "topic" consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. One key challenge in topic modeling is to develop a fast solution for indexing high dimensional data which is crucial to build large scale application data. This survey highlights on the current probabilistic topic models. It also point out scope and challenges in topic modeling.**

1. INTRODUCTION

The traditional method of learning about the particular topic is to read a book or a survey paper related to the topic. To buy a book is a time consuming and is inconvenient .This traditional method is not been applicable in most of the cases because of the topics and the technologies emerge in the fast growing world. To know the depth knowledge about a particular topic among the collection of documents that has become increasingly important and popular. No one will spend time by writing a book on some particular topic. To have knowledge of such emerging topic one can resort to research papers. But the research papers are not understandable by non-researchers and few research papers cover all aspects of the topic, there came into existence of topic modeling.

Topic modeling is the way of extracting common topics that occurs among the collection of documents. A document may have depth about a particular topic or multiple topics with different proportions. By using mathematical framework topic model examines each and every documents and discovers based on statistical words in each documents.

Text sequences are in many forms namely news streams, web log articles, email, instant messages, and research papers. The first step is to extract topic from sequence with both semantic and temporal information that are described by two distribution namely word distribution describing the semantic of the topics and time distribution that describes the topic's intensity over time.

Multiple sequences that shared the common topic having correlating with each other provides meaningful and comprehensive topic than the individual stream. The recent work is explored in the multiple sequence of temporal correlation in order to get the semantic correlation of topics.

In the case of asynchronism among multiple sequences (i.e) document from different sequences among the same topic that have different timestamps is common in practice. It is no guarantee in the news feed that the news covering the same topic will not be indexed with the same timestamp. The newspaper which takes hourly basis to publish the news , in the periodic it takes weekly basis and even some source will take monthly basis to publish the news that discuss about the common topic. The another example in the research archives is that papers published in the researches will undergo the proceeding of newsletters that is within the week or a month and followed by the conference which are published annually and finally in the journals that takes more than a year to publish. The paper published will share a common topic but at different time stamp. To visualize it the occurrence of two terms warehouse and mining in the title of all research papers published in two different sequences namely SIGMOD and TKDE respectively. In the case of asynchronism if the conventional topic mining is used it fails to find that mining and warehouse as a common topic. For mining common topics from multiple text sequences an effective method is been proposed. It defines the problem by introducing a principled probabilistic framework and put forward an algorithm to optimize the objective function by exploiting the mutual impact between the topic discovered and time synchronization. The key idea is to utilize the semantic and temporal correlation among the sequences and to build up a reinforcement process. To start, by extracting a set of common topics from given sequences using their original time stamps. And to update the time stamp of the document based on the extracted time stamp and word distribution. Finally update the time stamps of documents in all sequences by assigning them to most relevant topic.

2. RELATED WORK

Topic Detection and Tracking (TDT) works with the clustering based technique that aims to find and track in news sequences [12], [13] then came into existence of probabilistic generative models such as PLSA [14], LDA [15] and their derivatives.

Hoffman et.al [3] has proposed PLSA is a statistical model for analyzing two modes and finding the co-occurrence of data. This model aims creating each document in corpus, where each word in a document is sampled from a mixture of multinomial distribution that can be interpreted as topics, and proportions corresponding to mixture weights are sampled from a separate multinomial distribution for each document. Problem with PLSI has mixture parameter of each document that has too many parameters.

Bhei et.Al [4] used each document in the LDA to be viewed as the mixture of topics. LDA is similar to PLSA expect that the topic distribution is Dirichlet Prior. The problem of PLSA can be overcome by LDA by increasing the number of estimates by using Dirichlet prior distribution. The complexity of LDA is more when compared to PLSI that helps to find exact inferences from generative model. SparseLDA is approximately 20 times faster than highly optimized traditional LDA and twice the speed of previously published fast sampling method. To efficiently cope with this problem, several approximate inference algorithms are derived such as Variational Inference, and various Markov Chain Monte Carlo algorithms, such as Gibbs Sampling. LDA is more expensive than PLSA and LDA does not have some of the features such as more complex model like finding relation between the topics.

Bhei et.Al [6] introduced hierarchical LDA which is the LDA's extension model is Hierarchical LDA and is introduced by Bhei in 2003. LDA model a flat topic structure instead HLDA models tree of topics. Hierarchical LDA uses non-parametric Bayesian approach to model hierarchies. Tree of topic is constructed hierarchically of nodes by an algorithm. In topic tree model each and every node is represented by random number and has got corresponding word-topic distribution assigned to it. The tree can be traversed from the root till its leaves while sampling topics along the path

Text sequences is that the text collection which carries the temporal information. The text sequences will be seen in many applications. In order to capture the dynamic topic various methods has been proposed over time in text sequences. Even from the single sequences also these methods can be used. The individual sequences [5], [9] can be modeled with the random variable can either be discrete or continuous. It is assumed that in the given sequences, the timestamp of the document is independent from word. In [1] the author introduced hyper parameter that evolves over time in state transfer in the sequence. For each time slice, a hyper parameter is assigned with a state by a probability distribution, given the state on the former time slice.(The timestamp of the document is independent from word)

.

In [7] for each slice the time dimension is cut into time slice and topics. As a result in multiple text sequences, topic in each sequence can only be estimated separately and potentially correlated between topics in different sequences both semantically, temporally cannot be fully explored. From the static text collection the semantic correlation [16], [17], [18] between different topics was considered. Similarly, Zhai et al.[19] explored common topic in multiple static text sequences.

Wang et al.[11] has proposed a very recent work in topic mining methods that aims to discover common (bursty) topic over multiple text sequences. Their approach is entirely different from this method because this method deals with asynchronism (i.e) common topic occurring over different sequences with different time stamp. But Wang proposed to find common topic from different sequences that are synchronized and coordinated. Based on this premise, documents with same time stamp are combined together over different sequences so that the word distributions of topics in individual sequences can be discovered, As a contrast, this method aim to find topics that are common in semantics, while having asynchronous time distributions in different sequences.

Asuncion et al. [20] studied a generalized asynchronous distributed learning scheme with applications in topic mining. However, in their work the term "asynchronous" means a set independent Gibbs sampler which communicates with each other in an asynchronous manner. Therefore, their problem setting is fundamentally different this method.

We also note that there is a whole literature on similarity measure between time series (sequences). Various similarity functions have been proposed, many of which addressed the asynchronous nature between time series [21], [22]. However, defining an asynchronism robust similarity measure alone does not necessarily solve our problem. In fact, most of the similarity measures deal with asynchronism implicitly, rather than fix the asynchronism explicitly. It introduces a generative topic model which incorporates both temporal and semantic information function, which is to maximize the likelihood subject to certain constraints.

3. PROPOSED SYSTEM

The proposed work is based on the topic modeling PLSI. PLSI is a statistical model for analyzing two modes and finding the co-occurrence of data. This model aims creating each document in corpus, where each word in a document is sampled from a mixture of multinomial distribution that can be interpreted as topics, and proportions corresponding to mixture weights are sampled from a separate multinomial distribution for each document.

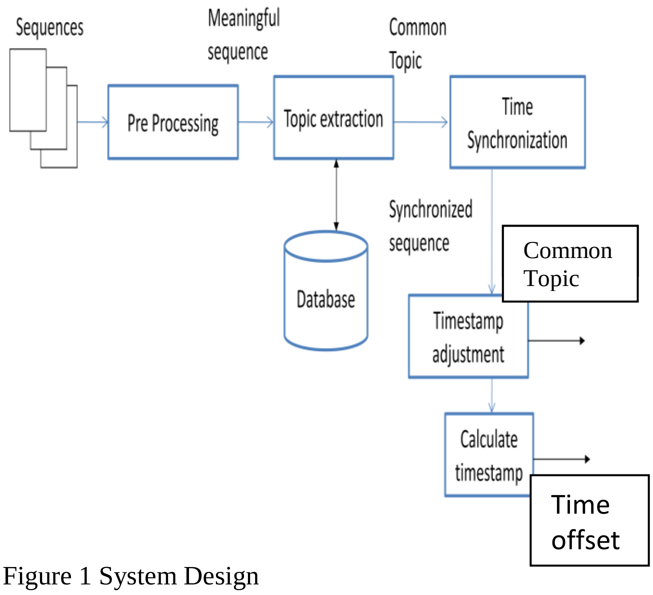


Figure 1 System Design

The text sequences which include the detail of title name, author name and time (in the form of date) is given as input. The text sequence is given to the first module as an input called pre-processing. Pre-Processing module will remove the stop word and will undergo stemming process that will lead to the meaningful sequences without the meaningful words like of, the, an etc. The meaningful sequence is given as input to the second module namely topic extraction which will extract the common topic from the meaningful text sequences and store it in the database. Based on the topic extracted in the topic extraction module along with the time stamp of the each document, synchronization is done and finally the topic is extracted based on the selected timestamp.

The Topic modeling over asynchronous text sequence will undergo the following modules

- Pre-Processing
- Topic Extraction
- Time Synchronization

Pre-processing is the first module in the topic modeling. The preprocessing has two process namely stop word removal and stemming. The dataset or the text corpus is given as input in which the stop word. The stop word will remove some common, short function words such as the, is, at, which, and on. So as a result the stop words output is given as input for stemming in which the root words will be removed like the words ending with 'ing', 'ive', 'ate', 'able'.

The input of the topic extraction is the processed sequence of the pre-processing and the output retrieved by this module is the common topic. Topic Extraction evaluates the common topics with no meaningful word from multiple text sequences. It then performs extracting of same topic from different sequences with different timestamps. The extraction is done by using an EM algorithm which will extract the top list words from the whole dataset.

The E-Step,

$$\Sigma_w \Sigma_t c(w,t) \log(\lambda B p(w \mid B) + (1 - \lambda \mathrm{B}) \quad \Sigma_z p(z|t) p(w \mid z)$$

and M-Step

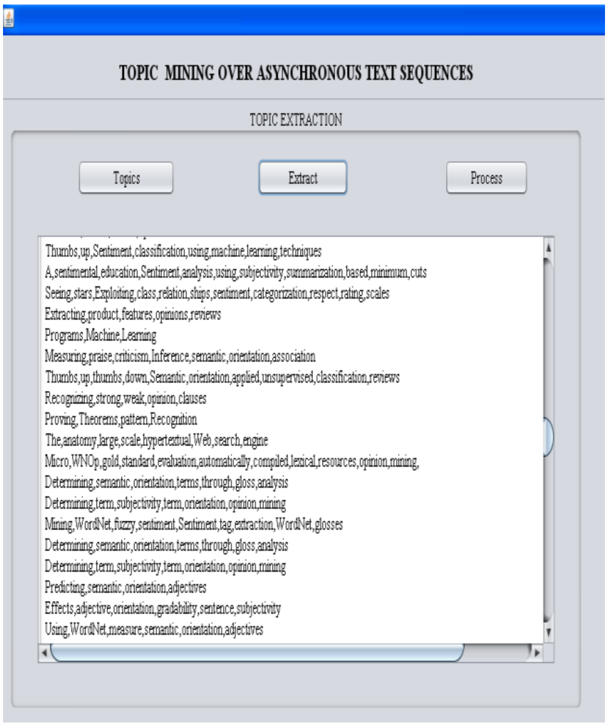
$$\mathbf{p(z \mid t) = c(w,t)p(z \mid w,t)} \;/\; \sum_z \sum_w c(w,t)p(z \mid w,t)$$

$$\mathbf{p(w \mid z)} = \sum_t c(w,t)p(z \mid w,t) \;/\; \sum_w \sum_t c(w,t)p(z \mid w,t)$$

The EM step is used to find the maximum likelihood function. These steps are repeated alternately and the objective function guarantees to converge to a local optimum. This process will collect the unique topic by using pair wise divergence.

$$\mathbf{KL(z1,z2)} = \sum_w p(w \mid z1) \log p(w|z1)/p(w \mid z2)$$

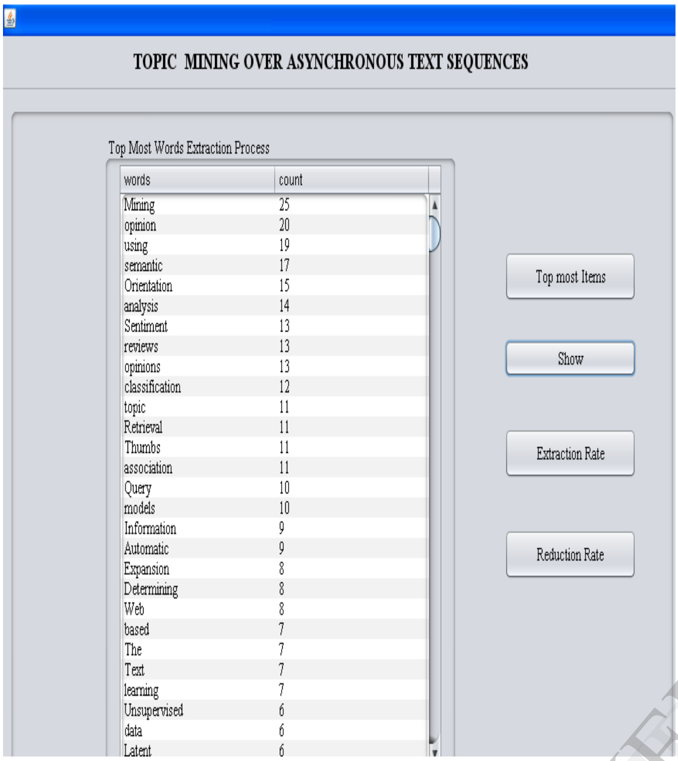
KL-divergence indicates that the two topics are more discriminative to each other and 0 divergence means that two topics are identical. This method extracts much more discriminative topics than those extracted by the previous method.



Thumbs,up,Sentiment,classification,using,machine,learning,techniques
A,sentimental,education,Sentiment,analysis,using,subjectivity,summarization,based,minimum,cuts
Seeing,stars,Exploiting,class,relation,ships,sentiment,categorization,respect,rating,scales
Extracting,product,features,opinions,reviews
Programs,Machine,Learning
Measuring,praise,criticism,Inference,semantic,orientation,association
Thumbs,up,thumbs,down,Semantic,orientation,applied,unsupervised,classification,reviews
Recognizing,strong,weak,opinion,clauses
Proving,Theorems,pattern,Recognition
The,anatomy,large,scale,hypertextual,Web,search,engine
Micro,WNOp,gold,standard,evaluation,automatically,compiled,lexical,resources,opinion,mining,
Determining,semantic,orientation,terms,through,gloss,analysis
Determining,term,subjectivity,term,orientation,opinion,mining
Mining,WordNet,fuzzy,sentiment,Sentiment,tag,extraction,WordNet,glosses
Determining,semantic,orientation,terms,through,gloss,analysis
Determining,term,subjectivity,term,orientation,opinion,mining
Predicting,semantic,orientation,adjectives
Effects,adjective,orientation,gradability,sentence,subjectivity
Using,WordNet,measure,semantic,orientation,adjectives

Figure 2 Topic Extracted by PLSI method

The input given to the time synchronization is the common topic extracted in the topic extraction module. The matching of the document is done based on the time stamp adjustment of the document that will extract the topic published during the period of time.



TOPIC MINING OVER ASYNCHRONOUS TEXT SEQUENCES

Top Most Words Extraction Process

words	count
Mining	25
opinion	20
using	19
semantic	17
Orientation	15
analysis	14
Sentiment	13
reviews	13
opinions	13
classification	12
topic	11
Retrieval	11
Thumbs	11
association	11
Query	10
models	10
Information	9
Automatic	9
Expansion	8
Determining	8
Web	8
based	7
The	7
Text	7
learning	7
Unsupervised	6
data	6
Latent	6

Figure 3 Maximum count of words extracted in dataset

Figure 3 and figure 4 shows the extracted common words from the dataset by using both topic extraction and time synchronization algorithm alternately and it finds the maximum likelihood of the words that shows the most research topic that is been published in the archives of both conferences and also journals. The collected topics will help the researches to concentrate in their focused area.

4. PARAMETERS USED FOR EXPERIMENTAL EVALUATION:

The standard parameters which are used for experimental evaluation are precision, recall and accuracy.

Precision is defined as number of retrieved relevant documents divided by total number of retrieved documents and the recall is the number of retrieved relevant document divided by total number of relevant documents in the database. Accuracy can be calculated as relevant document retrieved in top T returns divided by T. The formulas for calculation of these evaluation parameters can be given as following:

Precision:

Number of retrieved relevant documents

Total number of retrieved documents

Recall:

Number of retrieved relevant document

Total number of relevant documents

Accuracy :

Relevant documents retrieved in top T

T

Perplexity is a commonly used intrinsic evaluation metric in the topic model literature (Blei *et al.* 2003; Griffiths and Steyvers 2004). Perplexity originates from language modelling, and is calculated by estimating the probability of words in held out/test data based on training data.

$$\text{Perplexity}(test|train) = \exp - \left(\frac{\sum_{m=1}^{M} \log P(\mathbf{w}_m|train)}{\sum_{m=1}^{M} N_m} \right)$$

Where *M* is the number of documents in *test*; **w***m* are the words in document *m*; and
Nm is the number of words in document *m*.

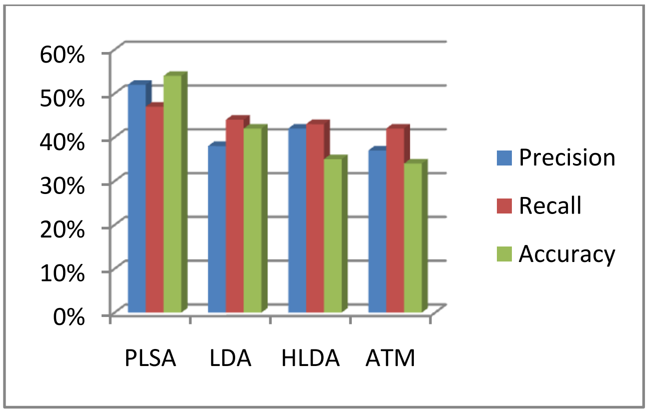


Fig. 2. Graphs showing the average recall, the average precision as in [9]

5. CONCLUSION:

The proposed method discovers and fixes potential asynchronism among sequences and consequentially extracts better common topics. It performs topic extraction and time synchronization alternatively to optimize a unified objective model. This method significantly outperforms both in quality and in quantity. So the performance of the method becomes more robust and stable. The framework is not only useful in practice but also valuable to machine learning because human being learn knowledge and even topics change over time and past knowledge is not discarded but used to solve new problems.

REFERENCES:

1. A. David, J. Li, L. Zhou, and F. Muhammad, "Knowledge discovery through directed probabilistic topic models: a survey," Frontiers of Computer Science in China, vol. 4, no. 2, pp. 280–301, Jun. 2010.
2. David M. Blei. Introduction to Probabilistic Topic Models. Communications of the ACM, 2011 pp.
3. Mark Steyvers, Tom Griffiths. Probabilistic Topic Models. In Landauer
4. Zhu, Jun, and Eric P Xing. "Conditional Topic Random Fields." Forbes. Ed. Johannes Fürnkranz& Thorsten Joachims.
5. Steven Abney and Marc Light. 1999. "Hiding a semantic hierarchy in a markov model". In Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing, pages 1–8.
6. T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating topics and syntax," In Advances in Neural Information Processing Systems 17, vol. 17, 2005, pp. 537–544.
7. A. Gruber, M. Rosen-Zvis, and Y. Weiss, "Hidden topic markov models," in Artificial Intelligence and Statistics, 2007.
8. T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.
9. Wang Wei, PayamBarnaghi, "Probabilistic Topic Models for Learning Terminological Ontologies" Member, IEEE, and AndrzejBargiela, Member, IEEE.
10. X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in ICDM 2007: Proceeding of the seventh IEEE International Conference On Data Mining, Ed., 2007, pp. 697–702.
11. X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 424–433.
12. D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proceedings of the 23rd international conference on Machine learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120.
13. D. Blei, T. Gri, M. Jordan, and J. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," 2003.
14. D. M. Blei and J. D. Lafferty. (2006) Correlated topic models.
15. T. Hofmann. (1999) Probabilistic latent semantic indexing.
16. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R.Harshman,"Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, pp. 391–407, 1990.
17. D. M. Blei and J. D. Mcauliffe, "Supervised topic models,", in Proceedings of th Neural Information Processing Systems – NIPS, 2007.
18. D. Mimno and A. McCallum, "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression," in Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI '08), 2008.
19. M. R. Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," ACM Trans. Inf. Syst., vol. 28, no. 1, pp. 1–38, Jan. 2010.
20. D. M. Blei and P. J. Moreno, "Topic segmentation with an aspect hidden markov model," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR '01. New York, NY, USA: ACM, 2001, pp. 343–348.
21. J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", Proceedings ICASSP-98, Seattle, May 1998.
22. Y. Wang, H. Bai, M. Stanton, W. Y. Chen, and E. Y. Chang, "PLDA: Parallel latent dirichlet allocation for Large-Scale applications," in Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management, ser. AAIM '09, vol. 5564. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 301–314.
23. W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in Proceedings of the 23rd international conference on Machine learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 577–586