REVIEW ARTICLE OPEN ACCESS

Text Mining: (Asynchronous Sequences)

Sheema Khan(Student ME CSE), Zafar Ul Hasan, (ME (CSE), MBA(Mktg), LMISTE)
Department of Computer Science and Engineering, Everest Educational Societies College of Engineering & Technology, Aurangabad, 431001 India.
Intellisense Research & Development Cunsultancy (IRD), Aurangabad, 431001 India.

Abstract
In this paper we tried to correlate text sequences those provides common topics for semantic clues. We propose a two step method for asynchronous text mining. Step one check for the common topics in the sequences and isolates these with their timestamps. Step two takes the topic and tries to give the timestamp of the text document. After multiple repetitions of step two, we could give optimum result.
Keywords: *Correlation, Asynchronous sequences, Timestamps*

I. Introduction

If we see today's scenario, we would find that lot of text material is generated in this decade. The sources of text generation are different like Internet, Application packages, Research investigations reports, Communication, Entertainment, etc. We can be lost in searching knowledge from these text sources. Hence it is required to give some method for discovering knowledge from this enormous text generated from decades. In this paper we tried formulate a method to extract knowledge from the available source of text sequences. This method of discovering knowledge from text can be achieved in two steps. In steps one, we try to determine the distribution of word intensity by knowing the meaning of the topic which is to be mined and in steps two, the distribution can by using time distribution method.

In reality many topics and text sequences are correlated. A semantic topic and the comprehensive topic are discovered by the interaction of multiple sequences, rather than a single or individual stream. According to recent work, over different sequences same time distribution are share by the common topic, usually different sequences are synchronous in time. But multiple sequences which contain synchronisms, is actually very common in practice. For example this is not sure that the article of any news feeds having the same topic indexed the same timestamps, because for any news agencies there is delay of hours, and days for news papers, and also weeks for periodicals.

We proposed an effective method to solve the problem of mining common topics from multiple asynchronous text sequences. We formally define the framework which is a principled, and the problem based on which a unified objective function can be derived. To optimize the objective function an algorithm is define, this algorithm helps by exploiting the mutual impact between topic discovery and time synchronization.

The key point is to use the correlation between sequences which is temporal and meaningful to build up a mutual and strong process. First we extract common topic using their timestamp from a given set of sequences. Second we update the timestamp of documents by checking them to most closely connected topic, on the basis of extracted topic and their word distribution. This step reduces the asynchronism among sequences. According to new timestamp the common topic are refined after synchronization. To maximize the unified objective function these steps are repeated alternatively and these function are provably converges monotonically.

The main points of our work are to

Mining common topic from multiple text sequences.

An objective function of problem which introduce the principled and probabilistic framework to formalize our problem

To maximize the objective function we develop optimization algorithm which is guaranteed optimum.

II. Literature Survey

Yuekui Yang et al. [1] parsed every web page as a Dom-Tree. They proposed some rules in tree aiming at extracting the relationship among different paragraphs and then presented a new topic-specific web crawler which calculated the unvisited URLs prediction score based on the web page hierarchy and the text semantic similarity. They calculated the text similarity using vector space model (VSM) which considered the query or paragraph as a vector in which the terms are independent and contacted different paragraphs in a web page according to their hierarchy in a Dom-Tree. Xiaofeng Liao et al. [2] considered the problem of modeling the topics in a sequence of images with known timestamp.

Detecting and tracking of temporal data is an important task in multiple applications, such as finding hot research point from scientific literature, news article series analysis, email surveillance, search query log mining, etc.

Besides collections of text document they also considered mining temporal topic trends from image data set. Chenguha Lin et al. [3] proposed joint sentiment-topic (JST) model based on latent Dirichlet allocation [7], which detects sentiment and topic simultaneously from text. JST is equivalent to Reverse-JST without a hierarchical priority. Neustein [4] showed how sequence package analysis is informed by algorithms that can work with, rather than be hindered by, less than perfect natural speech for intelligent mining of doctor-patient recordings and blogs. Watts et al. [5] throws light on organization's knowledge gained through technical conference. They worked out that there are processes where the knowledge gains are limited to the experiences and communication skills of the individuals attending the conference.

Many conference proceedings are published and provided to attendees in electronic format, such as on CD-ROM and/or published on the internet, such as IEEE conference proceedings. These proceedings provide a rich repository that can be mined. They compiled reflected hot topics as defined by the researchers in the field and delineate the technical approaches being applied. As per their work R&D profiling can more fully exploited by recorded conference proceedings' research to enhance corporate knowledge. They illustrated in their paper the potential in profiling conference proceedings through use of WebQL information retrieval and TechOasis (Vantage Point) text mining software by showing how tracking research patterns and changes over a sequence of conferences can illuminate R&D trends map dominant issues and spotlight key research organizations. Walenz et al. [6] described sequencer system for the temporal analysis of named entities in news articles between media reported stories and user generated content. They explored the evolution of social contexts with time that can provide unique insights into human social dynamics.

Wenfeng Li et al. [7] used the user's web browsing history that can be mined out. They presented an innovative method to extract user's interests from his/her web browsing history. They applied an algorithm to extract useful texts from the web pages in user's browsed URL sequence [10]. Unlike other works that need a lot of training data to train a model to adopt supervised information, they directly introduced raw supervised information to the procedure of LLDA-TF. In their paper Fotiadis et al. [8] presented a methodology for biosequence classification, which employs sequential pattern mining and optimization algorithms. In first stage, sequential pattern mining algorithm is applied to a set of biological sequences and the sequential patterns are extracted. Then, the score of each pattern with respect to each sequence is calculated using a scoring function and the score of each class under consideration is estimated. The scores of the patterns and classes are updated, multiplied by a weight. In the second stage optimization technique was employed to calculate the weight values to achieve the optimal classification accuracy. The methodology is applied in the protein class and fold prediction problem. Extensive evaluation is carried out, using a dataset obtained from the Protein Data Bank. Itoh [9] gave a contextual analysis processing technique, consisting in determining the context understanding together with coherences in sentences, of concepts and phenomena related to each others that must be able to simultaneously interpret accurately a sequence of multiple semantic representations.

By applying a semantic analysis of co-occurrence expressions, based on the use of phrases having an absolute evaluation polarity, he developed a system achieving analysis capable of detecting the role relations between words, the relationship of meaning in a sentence, identifying transitions in the topic, anaphora, endophora, and analyzing even idiomatic expressions and textual emoticons. Our system evaluated correctly ldquopositiverdquo or ldquonegativerdquo nuance for 75.0% of those. Subasic et al. [11] proposed a method and visualization tool for mapping and interacting stories published in web pages and articles. In contrast to existing approaches, their method concentrated on relational information and on local patterns rather than on the occurrence of individual concepts and global models. They also presented an evaluation framework.

Sekiya et al. [13] proposed that a word sequence can be used to identify context. Both contexts identified by word sequences and word sets related to the contexts are shown concretely. They used the confabulation model and five statistical measures as relations. Comparing the measures they found that cogency and mutual information were the most effective. Creamer et al. [14] in their paper analyzed the relationship between asset return, volatility and the centrality indicators of a corporate news network conducting a longitudinal network analysis. They built a sequence of daily corporate news network using companies of the STOXX 50 index as nodes, the weights of the edges the sum of the number of news items with the same topic by every pair of companies identified by the topic model methodology. They performed the Granger causality test and the Brownian distance covariance test of independence among several measures of centrality, return and volatility. They found that the average

eigenvector centrality of the corporate news networks at different points of time has an impact on return and volatility of the STOXX 50 index. Likewise, return and volatility of the STOXX 50 index also had an effect on average eigenvector centrality.

Perez et al. [15] proposed architecture for the integration of a corporate warehouse of structured data with a warehouse of text-rich XML documents. Yanpeng Li et al. [16] present a feature coupling generalization (FCG) framework for generating new features from unlabeled data. It selected two special types of features example-distinguishing features (EDFs) and class-distinguishing features (CDFs) from original feature set, and then generalizes EDFs into higher-level features based on their coupling degrees with CDFs in unlabeled data. It gave EDFs with extreme sparsity in labeled data that can be enriched by their co-occurrences with CDFs in unlabeled data so that the performance of these low-frequency features can be greatly boosted and new information from unlabeled can be incorporated. Navigli et al. [17] presented a method, called structural semantic interconnections (SSI), which creates structural specifications of the possible senses for each word in a context and selects the best hypothesis according to a grammar G, describing relations between sense specifications. They created sense specifications from several available lexical resources that they integrated in part manually, in part with the help of automatic procedures. The SSI algorithm applied to different semantic disambiguation problems.

III. Proposed Model

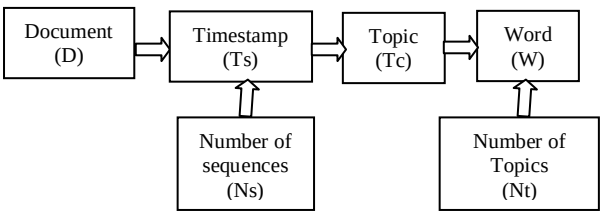

Fig a Proposed generative model for asynchronous text mining.

The process as per the fig a is as follows.

1. Select a document D from its source like website, data warehouse, etc.
2. A timestamp T for that related document is selected. This means for every document, only one timestamp is associated.
3. The next step will be to select common topic Tc and then to select the topic T and get the word.

Conventional methods on topic mining try to maximize the likelihood function L by adjusting the probabilities of topic and word assuming probability of timestamp is known. However, in our work, we need to consider the potential asynchronism among different sequences. Thus, besides finding optimal probabilities of topic and words we also need to decide probability of timestamp to further maximize the likelihood function. In other words, we want to assign the document with timestamp T to a new timestamp by determining its relevance to respective topics, so that we can obtain larger L, or equivalently, topics with better quality. By the term asynchronism, we refer to the time distortion among different sequences. The relative temporal order within each individual sequence is still considered meaningful and generally correct. Therefore, during each synchronization step, we preserve the relative temporal order of documents in each individual sequences with earlier timestamp before adjustment will never be assigned to later timestamp after adjustment as compared to its successors. This constraint aims to protect local temporal information within each individual sequence while fixing the asynchronism among different sequences.

3.1 Algorithm

In this section, we show how to solve our objective function through an alternate (constrained) optimization scheme. Our algorithm has two steps. The first one assumes that the current timestamps of the sequences are synchronous and extract common topics from them. The second step synchronizes the timestamps of all documents by matching them to most related topics, respectively. Then, we go back to first step and iterate until convergence.

3.2 Topic Extraction

We assume the current timestamps of all sequences are already synchronous and extract common topics from them. Our algorithm is summarized as below. K is the number of topics specified by users. The initial values of timestamps and objective function are counted from the original timestamps in the sequences.

3.3 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

For w=1 to T do compute objective function
End
End
Update timestamp
End
Until convergence

3.4 Constraint on Time Synchronization

We assumed asynchronism in given sequences. We assumed that timestamps are distorted and sequential information between documents is correct. 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.

We argue that the second option works better in practice since real-world data sets are not perfect. Although we assume that the sequential information 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 documents and assign them to the correct topics.

3.5 Convergence

Our objective function will converge to a local optimum after iterations. Notice that there is a trivial solution to the objective function, which is to assign all documents to an arbitrary timestamp and our algorithm would terminate at this local optimum. This local optimum is apparently meaningless since it is equivalent to discard all temporal information of text sequences and treat them like a collection of documents. Nevertheless, this trivial solution only exists theoretically. In practice, our algorithm will not converge to this trivial solution, as long as we use the original timestamps of text sequences as initial value and have more numbers of topics.

3.6 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 timestamps of documents. 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 timestamps of documents before and after adjustment in each iteration. Specifically, given document D with time T, we now look for an optimal topic function within the neighborhood of topic.

IV. Conclusion

Our first aim is that to extract common topic from multiple sequences which are asynchronous. We propose a method which importantly extract common topic by fixing potential asynchronism among sequences. Self improvement process is introduced by utilizing correlation between the semantic and temporal information in sequences. To optimize a unified objective function by extracting common topic and time synchronization alternately. Most likely results are guaranteed by our algorithm. Two baseline methods are

From asynchronous text sequences we extract meaningful and discriminative topic.

Quality and quantity are maintained by our method.

Performance of our method is strong and healthy against random initialization and parameter setting.

References

[1]. Yuekui Yang, Yajun Du, Yufeng Hai and Zhaoqiong Gao "A Topic-Specific Web Crawler with Web Page Hierarchy Based on HTML Dom-Tree",IEEE Vol 1, 10.1109/APCIP.2009.110, 2009 , Page(s): 420 - 423

[2]. Xiaofeng Liao, Wang, Yongji and Liping Ding, "Discovering Temporal Patterns from Images using Extended PLSA", IEEE, 10.1109/ICMULT.2010.5630978, 2010 , Page(s): 1 – 7

[3]. Chenghua Lin, Yulan He, Everson, R. And Ruger, S., "Weakly Supervised Joint Sentiment-Topic Detection from Text" IEEE Vol 24, Issue 6, 10.1109/TKDE.2011.48, 2012 , Page(s): 1134 - 1145

[4]. Neustein, A., "SEQUENCE PACKAGE ANALYSIS: A New Natural Language Understanding Method for Intelligent Mining of Recordings of Doctor-Patient Interviews and Health-Related Blogs", IEEE, 10.1109/ITNG.2007.179, 2007 , Page(s): 431 – 438

[5]. Watts, R.J. and Porter, A.L., "Mining conference proceedings for corporate technology knowledge management", IEEE, 10.1109/PICMET.2005.1509711, 2005 , Page(s): 349 – 358
Walenz, B., Gandhi, R., Mahoney, W. And Quiming Zhu, "Exploring Social Contexts along the Time Dimension: Temporal Analysis of Named Entities", IEEE, 10.1109/socialcom.2010.80, 2010 , Page(s): 508 – 512

[6]. Wenfeng Li, Xiaojie Wang, Rile Hu and Jilei Tian, "User interest modeling by labeled LDA with topic features", IEEE, 10.1109/CCIS.2011.6045022, 2011 , Page(s): 6 – 11

[7]. Fotiadis, D.I., Exarchos, T.P. Tsipouras, M.G. and Papaloukas, C., "Biosequence Classification using Sequential Pattern Miningand Optimization", IEEE,

10.1109/ITAB.2007.4407423 , 2007 , Page(s): 58 – 61

[8]. Itoh M, "Contextual Analysis Processing Methods Able to Interpret Sentiments Evaluation Representations" IEEE,10.1109/ICSC.2009.98, 2009 , Page(s): 71 – 76

[9]. He Bai, jinlin Wang and Ye Li, "An Approach to Extracting Central urls on Catalog Page", IEEE, 10.1109/KAM.2008.71 , 2008 , Page(s): 388 – 392

[10]. Subasic, I. And Berendt, B., "Web Mining for Understanding Stories through Graph Visualisation, IEEE, 10.1109/ICDAM.2008.138

[11]. A. K. Santra, C. Josephine Christy, " Genetic Algorithm and Confusion Matrix for Document Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012, ISSN (Online): 1694-0814, , Page 322-328

[12]. Sekiya, H., Kondo, T., Hashimoto, M. And Takagi, T. , "Context representation using word sequences extracted from a news corpus", IEEE, 10.1109/NAFIPS.2005.1548639 , 2005 , Page(s): 783 - 786

[13]. Creamer, G.G., Ren, Y. And Nickerson, J.V., "Impact of Dynamic Corporate News Networks on Asset Return and Volatility", IEEE, 10.1109/socialcom.2013.121, 2013 , Page(s): 809 – 814

[14]. Perez, J.M., Berlanga, R., Aramburu, M.J.and; Pedersen, T.B., "R-Cubes: OLAP Cubes Contextualized with Documents", IEEE, 10.1109/ICDE.2007.369041, 2007 , Page(s): 1477 - 1478

[15]. 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

[16]. Navigli, R., and Velardi, Paola, "Structural semantic interconnections: a knowledge-based approach to word sense disambiguation", IEEE, Vol 27 , Issue: 7, 10.1109/TPAMI.2005.149, 2005 , Page(s): 1075 - 1086