Topic Identification Using Temporal and Concept Relationship Analysis on Text Streams

Jayalakshmi.M*1,Gunavathi.C*2
1*PG Scholar of Computer Science, K.S.Rangasamy College of Technology, Tiruchengode, India.*
Email: jayalakshmi.m88@gmail.com
2*Assistant professor (Academic), K S Rangasamy College of Technology, Tiruchengode, India.*
Email: sssguna@gmail.com

***Abstract*--Text documents are unstructured data elements. Similarity measures are used to analyze the document relationship. Document features are used in the classification process. Temporal mining methods are used to analyze time bounded data values. Classification technique is applied to assign labels for the transactions. Learning phase is carried out for transaction pattern identification. Testing process handles the pattern matching and label assignment task. Anomalous and normal transactions are identified using classification techniques. Documents from different sequences about the same topic may have different time stamps. Topic mining with time synchronization algorithm uses the generative topic model. The topic mining algorithm is divided into two steps. The first step extracts common topics from multiple sequences based on the adjusted time stamps. The second step adjusts the time stamps of the documents according to the time distribution of the topics. Unified objective function is used to correlate topic discovery and time synchronization. Topics and their word distributions are analyzed.The topic mining algorithm is enhanced with concept relationship analysis. Semantic and temporal correlations are used in the topic extraction process. Term weights are used to analyze the document similarity. Topics are ranked with the weight and similarity values.**

I. INTRODUCTION

More and more text sequences are being generated in various forms, such as news streams, weblog articles, emails, instant messages, research paper archives, web forum discussion threads, and so forth. To discover valuable knowledge from a text sequence, the first step is usually to extract topics from the sequence with both semantic and temporal information, which are described by two distributions, respectively: a word distribution describing the semantics of the topic and a time distribution describing the topic's intensity over time.

In many real-world applications, we are facing multiple text sequences that are correlated with each other by sharing common topics. Intuitively, the interactions among these sequences could provide clues to derive more meaningful and comprehensive topics than those found by using information from each individual stream solely. The intuition was confirmed by very recent work [2], which utilized the temporal correlation over multiple text sequences to explore the semantic correlation among common topics. The method proposed therein relied on a fundamental assumption that different sequences are always synchronous in time, or in their own term coordinated, which means that the common topics share the same time distribution over different sequences [1].

In this paper, we target the problem of mining common topics from multiple asynchronous text sequences and propose an effective method to solve it [9]. We formally define the problem by introducing a principled probabilistic framework, based on which a unified objective function can be derived. Then, we put forward an algorithm to optimize this objective function by exploiting the mutual impact between topic discovery and time synchronization.

The key idea of our approach is to utilize the semantic and temporal correlation among sequences and to build up a mutual reinforcement process. We start with extracting a set of common topics from given sequences using their original time stamps. Based on the extracted topics and their word distributions, we update the time stamps of documents in all sequences by assigning them to most relevant topics. This step reduces the asynchronism among sequences. Then after synchronization, we refine the common topics according to the new time stamps. These two steps are repeated alternately to maximize a unified objective function, which provably converges monotonically.

Besides theoretical justification, our method was also evaluated empirically on two sets of real-world text sequences. The first is a collection of six literature repositories consisting of research papers in the database literature from 1975 to 2006 and the second contains two news feeds of 61 days' news articles between 1 April and 31 May 2007. The method is able to detect and fix the underlying asynchronism among different sequences and effectively discover meaningful and highly discriminative common topics. To sum up, the main contributions of our work are:

- We address the problem of mining common topics from multiple asynchronous text

sequences. To the extent of our knowledge, this is the first attempt to solve this problem.

- We formalize our problem by introducing a principled probabilistic framework and propose an objective function for our problem.
- We develop a novel alternate optimization algorithm to maximize the objective function with a theoretically guaranteed (local) optimum.
- The effectiveness and advantage of our method are validated by an extensive empirical study on two real-world data sets.

II. RELATED WORK

Topic mining has been extensively studied in the literature, starting with the Topic Detection and Tracking (TDT) project, which aimed to find and track topics in news sequences1 with clustering-based techniques. Later on, probabilistic generative models were introduced into use, such as Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and their derivatives [5].

In many real applications, text collections carry generic temporal information and, thus, can be considered as text sequences. To capture the temporal dynamics of topics, various methods have been proposed to discover topics over time in text sequences [4], [6]. However, these methods were designed to extract topics from a single sequence. For example, in [8], which adopted the generative model, time stamps of individual documents were modeled with a random variable, either discrete or continuous. Then, it was assumed that given a document in the sequence, the time stamp of the document was generated conditionally independently from word. The authors introduced hyper-parameters that evolve over time in state transfer models in the sequence. For each time slice, a hyperparameter 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. 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, could not be fully explored. In [3], the semantic correlation between different topics in static text collections was considered. Similarly, Zhai et al. explored common topics in multiple static text collections.

Asuncion et al. [7] studied a generalized asynchronous distributed learning scheme with applications in topic mining. However, in their work the term "asynchronous" means set independent Gibbs samplers which communicate with each other in an asynchronous manner. Therefore, their problem setting is fundamentally different from ours.

We also note that there is a whole literature on similarity measure between time series. Various similarity functions have been proposed, many of which addressed the asynchronous nature between time series. 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, like what we do in this work.

III. PROBLEM FORMULATION AND OBJECTIVE FUNCTION

In this section, we formally define our problem of mining common topics from multiple asynchronous text sequences. We introduce a generative topic model which incorporates both temporal and semantic information in given text sequences. We derive our objective function, which is to maximize the likelihood estimation subject to certain constraints.

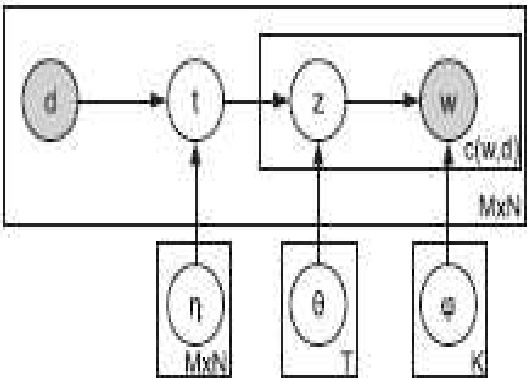

Fig. 1. An illustration of our generative model. Shaded nodes mean observable variables while white nodes mean unobservable variables. Arrow indicates the generation relationship.

The documents $\{d \in S_m : 1 \leq m \leq M\}$ are modeled by a discrete random variable d. The words are modeled by a discrete random variable w over vocabulary V. The time stamps are modeled by a discrete random variable t over {1, . . . , T}. At last, the common topics Z are encoded by a discrete random variable $z \in \{1, 2, \ldots, K\}$. Note that semantic information of a topic is encoded by the conditional distribution p(z|t) and its temporal information by p(z|t). The generating process is as follows (also see Fig. 1):

1. Pick a document d with probability p(d).

2. Given the document d, pick a time stamp t with probability p(t|d), where p(t = t|d) = 1 for some t. This means that a given document only has one time stamp.
3. Given the time stamp t, pick a common topic z with probability p(z|t) ~ Mult(θ).
4. Given the topic z, pick a word w with probability p(z|t) ~ Mult(φ).

According to the generative process, the probability of word w in document d is

p(w, d) = $\sum_{t,z}$ pp(d)p(t|d)p(z|t)p(z|t):

Consequentially, the log-likelihood function over all sequences is

$L = \sum_w \sum_d$ c(w, d) log p(w, d),

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

Conventional methods on topic mining try to maximize the likelihood function L by adjusting p(z|t) and p(z|t) while assuming p(t|d) is known. However, in our work, we need to consider the potential asynchronism among different sequences, i.e., p(t|d) is also to be determined. Thus, besides finding optimal p(z|t) and p(z|t), we also need to decide p(t|d) to further maximize L. In other words, we want to assign the document with time stamp t to a new time stamp g(t) by determining its relevance to respective topics, so that we can obtain larger L, or equivalently, topics with better quality.

Note that the mapping from t to g(t) is not arbitrary. 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, i.e., a document with earlier time stamp before adjustment will never be assigned to later time stamp 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.

IV. ALGORITHM

In this section, we show how to solve our objective function in (1) through an alternate optimization scheme. The outline of our algorithm is:

- Step 1. We assume that the current time stamps of the sequences are synchronous and extract common topics from them.
- Step 2. We synchronize the time stamps of all documents by matching them to most related topics, respectively. Then, we go back to Step 1 and iterate until convergence.

A.Topic Extraction

First, we assume the current time stamps of all sequences are already synchronous and extract common topics from them. In other words, now p(t|d) is fixed and we try to maximize the likelihood function by adjusting p(tjz) and p(z|t). Thus, we can rewrite the likelihood function as follows:

$$\sum_w \sum_d \text{c(w, d)} \log \sum_t \sum_z \text{p(d)p(t|d)p(z|t)p(w|z)}. \quad (1)$$

$$= \sum_w \sum_d \text{c(w, d)} \log \text{p(d)} \sum_t \text{p(t|d)} \sum_z \text{p(z|t)p(w|z)},$$

Since we have p(t = t|d) = 1 for some t, the above equation can be reduced to

$$\sum_w \sum_d \text{c(w, d,t)} \log \sum_t \text{p(z|t)p(w|z)}.$$

$$= \sum_w \sum_d \text{c(w, t)} \log \sum_t \text{p(z|t)p(w|z)} \quad (2)$$

Here, c(w, d, t) denotes the number of occurrences of word w in document d at time t, and p(d) is summed out because it can be considered as a constant in the formula.

Equation (2) can be solved by a well-established EM algorithm. The E-step writes

$$p(z|w,t) = \frac{p(z|t)p(w|z)}{\sum_z p(z|t)p(w|z)} \quad (3)$$

and the M-step writes

$$p(z|t) = \frac{\sum_w c(w,t)p(z|w,t)}{\sum_z \sum_w c(w,t)p(z|w,t)}$$

$$p(w|z) = \frac{\sum_t c(w,t)p(z|w,t)}{\sum_w \sum_t c(w,t)p(z|w,t)} \quad (4)$$

The E- and M-step repeat alternately and the objective function guarantees to converge to a local optimum.

B Time Synchronization

Algorithm 1: Topic mining with time synchronization

```
Input: F, p(t|d), c(w, d, t);
Output: p(w|z), p(z|t), p(t|d);
repeat
    Update c(w, t) with p(t|d) and c(w, d, t);
    Initialize p(z|t) and p(w|z) with random values;
    repeat
        Update p(z|t) and p(w|z) following Eq.(3) and (4);
    until Convergence;
    for m=1 to M do
        for j=1 to T do Initialize H(1 : 1, 1 : j);
        for i=2 to T do
            for j=1 to T do
                Compute H(1 : i, 1 : j) as shown in Eq.(7).
            end
        end
        Update p(t|d);
    end
until Convergence;
```

Once the common topics are extracted, we match documents in all sequences to these topics and adjust their time stamps to synchronize the sequences. Specifically, now p(z|t) and p(z|t) are assumed as known and we try to update p(t|d) to maximize our objective function. Given document d, we denote its current time stamp with t and its time stamp after adjustment with g(t).

The computational complexity of the topic extraction step (with the EM algorithm) is O(KV T) while the complexity of time synchronization step is approximately $O(VMT^3)$. Thus, the overall complexity of our algorithm is $O(V\ T(K\ )\ MT^2))$, where V is the size of vocabulary, T the number of different time stamps, K the number of topics, and M the number of sequences. If we take V, K and M as constants and only consider the length of sequence, which is T, the complexity of Algorithm 1 becomes $O(T^3)$. We will show in the next section how to reduce it to $O(T^2)$ with a local search strategy.

V. DISCUSSIONS AND EXTENSIONS

A. *The Constraint on Time Synchronization*

Recall that in our model we made a fundamental assumption about the asynchronism among the given sequences: we assume that the original time stamps as given are distorted, while the relative temporal order between documents is correct in general. This assumption is made based on observations from real-world applications. For example, news stories published by different news agencies may vary in absolute time stamps, but their relative temporal order conforms to the order of the occurrences of the events. Then, we translate this assumption into the formal constraint: $g(t_1) \leq g(t_2) \Leftrightarrow t_1 \leq t_2$. This constraint can be interpreted as a trade-off between two extreme cases: 1) strictly obey the original time stamps, which will harm the quality of the resultant topics due to the underlying asynchronism and 2) discard all temporal information given, which could result in topics without time distribution.

B. Convergence

Both of the two steps in our algorithm guarantee a monotonic improvement in our objective function in (1); the algorithm 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 a single time stamp 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 time stamps of text sequences as initial value and have K >1, where K is the number of topics. The adjusted time stamps of documents always converge to more than K different time points. Note that this is so even after relaxing the constraint by allowing two documents to swap their temporal order after multiple iterations, as discussed above. This is because our algorithm is essentially a mutual reinforcement process where we use both semantical and temporal information to identify common topics. The topic extraction step will prevent the algorithm from assigning all documents to a single time stamp, since in this case we may end up with topics with lower quality.

C. Cases Where Our Method May Not Work (Well)

Given the assumption we made, our model and our algorithm will not work well in the following cases: 1) there is no correlation between the semantical and temporal information of topics, i.e., the time distribution of any topic is random and 2) the temporal order of documents as given by their original time stamps varies greatly from the temporal order of underlying topics, e.g., Topic A appears before Topic B in one sequence, but after B in another. In either case, the better choice would be discarding the original temporal information and treating the text sequences as a collection of documents.

D. 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 time stamps 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 time stamps of documents before and after adjustment in each iteration. Specifically, given document d with time t, we now look for an optimal g(t) within the ε-neighborhood of t, where ε is the user-specified search range. Accordingly, becomes

$$\max_{g(t)} \sum_{w} \sum_{s=1}^{l} Q(w,s) \sum_{\{d:g(t)=s\}} c(w,d),$$
$$\text{s.t.} \quad \forall d, g(t) \in [t-\epsilon, t+\epsilon],$$
$$\forall d_1, d_2, g(t_1) \leq g(t_2) \Leftrightarrow t_1 \leq t_2.$$

This objective function can be solved with simple modifications. We can see that the complexity of the synchronization step has been reduced to $O(\varepsilon VMT^2)$; thus, the overall complexity is reduced from $O(T^3)$ to $O(T^2)$.

VI. TEMPORAL AND CONCEPT RELATIONS BASED TOPIC MINING

The topic mining scheme is developed to fetch the topics for the text streams using the termporal and concept analysis scheme. The data values are collected from the text streams. Different sources sent different messages. The text notes or news bits are collected in different time slot. All the messages are analyzed with time and content relationships. The content information is verified with statistical weight values. The term weight value is used in the text streams. The term weight is used for the similarity analysis.

The topics are extracted using the text content and message collected time information. The similarity measures are used to match the message contents. The topic mining algorithm is used to extract topics from the text stream contents. The topic mining algorithm is enhanced with concept relationship analysis. The concept relationships are used to extract semantic information. The temporal mining is applied to find out the time relationships. The time factors are involved to collect the topic flow details. Semantic and temporal correlations are used in the topic extraction process.

The topics are initially detected from the time and semantic relationships. The topics are ranked with reference to its associations and time sequences. Topics are ranked with the weight and similarity values. The topic weight is estimated from the text messages and semantic relationships. The similarity values are measured with the text contents. The ranked topics indicate that the topic arrival and similarity sequences.

VII. CONCLUSION

Multiple text sequences shares the common topics. The topic identification scheme is used to mine common topics from multiple asynchronous text sequences. Concept analysis and term weight factors are used to improve the topic detection process. The system produces the topics in ranked order. Topic extraction accuracy is improved by the system. Topics are produced with ranks in dynamic manner. Time distribution analysis is performed for temporal analysis. Term weight and semantic weight based text categorization process used in the system.

REFERENCES

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

[2] X. Wang, and R. Sproat, "Mining Correlated Bursty Topic Patterns from Coordinated Text Streams," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 784-793, 2007.

[3] W. Li and A. McCallum, "Pachinko Allocation: Dag-Structured Mixture Models of Topic Correlations," Proc. Int'l Conf. Machine Learning, pp. 577-584, 2006.

[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] D.M. Mimno, W. Li, and A. McCallum, "Mixtures of Hierarchical Topics with Pachinko Allocation," Proc. Int'l Conf. Machine Learning pp. 633-640, 2007.

[6] 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.

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

[8] 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.

[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.