Semantic Approach to Discover Topic over Mail Data

D. A. Kiran Kumar [α] & M. Saidi Reddy [σ]

Abstract - Text sequences or Time stamped texts, are ever-present in real-world applications. Multiple text sequences are frequently connected to each other by distributing common topics. The correspondence between these sequences provides more significant and comprehensive clues for topic mining than those from every individual sequence. However, it is non retrieval to explore the equivalence with the existence of asynchronism among multiple sequences, i.e., documents from different sequences about the same topic may have different time stamps. In this paper, we properly addressed the problem and suggested a new algorithm based on the generative topic model. The proposed algorithm consists of two alternate steps: the first step retrieves common data from multiple sequences based on the arranged time stamps provided by the second step; the second step arranges the time stamps of the documents according to the time distribution of the topics found by the first step. We accomplish these two steps simultaneously and after number retrievals a monotonic convergence of our objective function can be extracted. The effectiveness and advantage of our approach were justified through extensive practical studies on two real data sets consisting of six research paper repositories and two news article feeds, respectively.

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, 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.

However, this assumption is too strong to hold in all cases. Rather, asynchronism among multiple sequences, i.e, documents from different sequences on the same topic have different time stamps, is actually very common in practice. For instance, in news feeds, there is no guarantee that news articles covering the same topic are indexed by the same time stamps. There can be hours of delay for news agencies, days for newspapers, and even weeks for periodicals, because some sources try to provide first-hand flashes shortly after the incidents, while others provide more comprehensive reviews afterward. Another example is research paper archives, where the latest research topics are closely followed by newsletters and communications within weeks or months, then the full versions may appear in conference proceedings, which are usually published annually and at last in journals, which may sometimes take more than a year to appear after submission.

To visualize it, we have the relative frequency of the occurrences of two terms warehouse and mining, respectively, in the titles of all research papers published in SIGMOD (ACM International Conference on Management of Data), a database-related conference, and TKDE (IEEE Transactions on Knowledge and Data Engineering) from 1992 to 2006, a database-related journal. The first term identifies the topic data warehouse and the second data mining, which are two common topics shared by the two sequences. As shown in Fig. 1a, the bursts of both terms in SIGMOD are significantly earlier than those in TKDE, which suggests the presence of asynchronism between these two sequences. Thus, in this paper, we do not assume that given text sequences are always synchronous. Instead, we deal with text sequences that share common topics yet are temporally asynchronous. We apparently expect that multiple correlated sequences can facilitate topic mining by generating topics with higher quality. However, the asynchronism among sequences brings new challenges to conventional topic mining methods. If we overlook the asynchronism and apply the conventional topic mining methods directly, we are very likely to fail in identifying data mining and/or data warehouse as common topics

Author α : (M.Tech), CSE Dept, MRCET, Hyderabad.
E-mail : kiran.dingari@gmail.com
Author σ : Ph.D., CSE Dept, MRCET, Hyderabad.
E-mail : msreddy33@gmail.com

of the two sequences, since the bursts of the topics do not coincide (therefore, the relative frequency of the topical words becomes too low as compared to other words). As a contrast, after adjusting the time stamps of documents in the two sequences using our proposed method, the relative frequency of both warehouse and mining are boosted over a certain range of time, relatively. Thus, we are more likely to discover both topics from the synchronized sequences. It proves that fixing asynchronism can significantly benefit the topic discovery process. However, as desirable as it is to detect the temporal asynchronism among different sequences and to eventually synchronize them, the task is difficult without knowing the topics to which the documents belong beforehand. A native solution is to use coarse granularity of the time stamps of sequences so that the asynchronism among sequences can be smoothed out. This is obviously dissatisfactory as it may lead to unbearable loss in the temporal information of common topics and different topics would inevitably be mixed up.

A second way, shifting or scaling the time dimension manually and empirically, may not work either because the time difference of topics among different sequences can vary largely or irregularly, of which we can never have enough prior knowledge. In this paper, we target the problem of mining common topics from multiple asynchronous text sequences and propose an effective method to solve it. 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. We show that our 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. System Development

a) Titles Pre-Processing (Dataset1)

- Collection of titles from research journals
  i. E.g.: DEXA, ICDE, IS, SINMOD, TKDE, and VLDB.
- Timestamp for each title words.
- Timestamp is the Volume, SNO, Month, and Year of publication.
  i. Ex: Title (d1) = “Topic Mining over Asynchronous Text Sequences” Timestamp (t1) = “24, 1, JANUARY, 2012”.
- Titles Dataset is prepared for all the documents collected.

b) Stemming: Eliminating Stop Words

- The Dataset1 consists of large number of titles.
- The words in all the titles may consists of stop words such as “an”, “the” etc,.
- For eliminating stop words from the titles, there is a need of stop words list.
- The stop words list in the proposed approach is “TMG” list.

c) Word Pre-Processing

- Title consists of sequence of words (w).
- Each word is assigned with the timestamp as title sequence (z).
- The order of sequence of words produces meaningful and suitable topical words for the context that they have.

d) Unique Topical Words

- The dataset1 consists of all the titles.
- Unique words of all the titles becomes topical words
- Each topical word sequence is equal the title.
- The module identifies all the topical words.

e) Dataset2: Document Pre-Processing

- Research Articles can be in pdf format or word format in general.
- The Proposed method requires the documents in the text format.
- Hence, pdf or doc files have to be converted to text format.

f) Frequency of Unique Topical Words

- When a pdf is given, the pdf document may consist of unique topical words (dataset1).

- Unique Topical Words Frequency for the given pdf document will be calculated.
- If a unique topical word is present, it means that the document is relevant to the topical word and impact will be based on its frequency that the unique topical word occurred.

g) Finding High Frequency Unique Words

- The document may contain more number of topical words with some frequency.
- Frequently Occurred Unique Topical Words are most relevant to the document data.
- The module identifies the high listed unique topical words.

III. Related Work

a) Dynamic Topic Models

While traditional time series modeling has focused on continuous data, topic models are designed for categorical data. Our approach is to use state space models on the natural parameter space of the underlying topic multinomial's, as well as on the natural parameters for the logistic normal distributions used for modeling the document-specific topic proportions. First, we review the underlying statistical assumptions of a static topic model, such as latent Dirichlet allocation (LDA) (Blei et al., 2003). Let fi1:K be K topics, each of which is a distribution over a fixed vocabulary. In a static topic model, each document is assumed drawn from the following generative process:

1. Choose topic proportions fi from a distribution over the (K − 1) simplex, such as a Dirichlet.
2. For each word: (a) Choose a topic assignment Z fi Mult (fi). (b) Choose a word W _ Mult (fiz).

b) Data Sets

The first data set used in our experiment is six research paper repositories extracted from DBLP, 2 namely, DEXA, ICDE, Information Systems (journal), SIGMOD, TKDE (journal), and VLDB. These repositories mainly consist of research papers on database technology. Each repository is considered as a single text sequence where each document is represented by the title of the paper and time stamped by its publication year. The second data set is two news articles feeds, which consist of the full texts of daily news reports published on the websites of International Herald Tribune3 and People's Daily Online, 4 respectively, from 1 April 2007 to 31 May 2007. Each document is time stamped by its publication date. Text sequences are preprocessed by TMG5 for stemming and removing stop words. Words that appear too many (appear in over 15 percent of the documents) or too few (appear in less than 0.5 percent of the documents) times are also removed. After preprocessing, the literature repositories have a vocabulary of 1,686 words and news feeds of 3,358 words.

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 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.

d) Hierarchical Structure and E-mail Streams

Extracting hierarchical structure. From an algorithm to compute an optimal state sequence, one can then define the basic representation of a set of bursts, according to a hierarchical structure. For a set of messages generating a sequence of positive inter-arrival gaps x = (x1; x2; : : : ; xn), suppose that an optimal state sequence q = (qi1 ; qi2 ; : : : ; qin) in Afis; has been determined.

Following the discussion of the previous section, we can formally define a burst of intensity j to be a maximal interval over which q is in a state of index j or higher. More precisely, it is an interval [t; t0] so that it; : : : ; it0 fi j but it □1 and it0+1 are less than j (or undaunted if t □ 1 < 0 or t0 + 1 > n).

It follows that bursts exhibit a natural nested structure: a burst of intensity j may contain one or more sub-intervals that are bursts of intensity j + 1; these in turn may contain sub-intervals that are bursts of intensity j+2; and so forth. This relationship can be represented by a rooted tree □, as follows. There is a node corresponding to each burst; and node v is a child of node u if node u represents a burst Bu of intensity j (for some value of j), and node v represents a burst By of intensity j + 1 such that By fi Bu. Note that the root of □ corresponds to the single burst of intensity 0, which is equal to the whole interval [0; n]. Thus, the tree □ captures hierarchical structure that is implicit in the underlying stream.

The transformation from an optimal state sequence, to a set of nested bursts, to a tree. Hierarchy in an e-mail stream. Let us now return to one of the initial motivations for this model, and consider a stream of e-mail messages. What does the hierarchical structure of bursts look like in this setting?

I applied the algorithm to my own collection of saved e-mail, consisting of messages sent and received between June 9, 1997 and August 23, 2001. (The cut-ofi dates are chosen here so as to roughly cover four academic years.) First, here is a brief summary of this collection. Every piece of mail I sent or received during this period of time, using my cs.cornell.edu e-mail address, can be viewed as belonging to one of two categories: rest, messages consisting of one or more large files, such as drafts of papers mailed between co-authors (essentially, E-mail as file transfer); and second, all other messages. The collection I am considering here consists simply of all messages belonging to the second, much larger category; thus, to a rough approximation, it is all the mail I sent and received during this period, unaltered by content but excluding long files. It contains 34344 messages in UNIX mailbox format, totaling 41.7 megabytes of ASCII text, excluding message headers.1

e) Enumerating Bursts

Given a framework for identifying bursts, it becomes possible to perform a type of enumeration: for every word w that appears in the collection, one computes all the bursts in the stream of messages containing w. Combined with a method for computing a weight associated with each burst, and for then ranking by weight, this essentially provides a way to and the terms that exhibit the most prominent rising and falling pattern over a limited period of time. This can be applied to e-mail, and it can be done very efficiently even on the scale of the e-mail corpus from the previous section; roughly speaking, it can be performed in a single pass over an inverted index for the collection, and it produces a set of bursts that correspond to natural episodes of the type suggested earlier. In the present section, however, I focus primarily on a different setting for this technique: extracting bursts in term usage from the titles of conference papers. Two distinct sources of data will be used here: the titles of all papers from the database conferences SIGMOD and VLDB for the years 1975-2001; and the titles of all papers from the theory conferences STOC and FOCS for the years 1969-2001. The first issue that must be addressed concerns the underlying model: unlike e-mail messages, which arrive continuously over time, conference papers appear in large batches essentially, twenty to sixty new papers appear together every half year. As a result, the automaton A_S; is not appropriate, since it is fundamentally based on analyzing the distribution of inter-arrival gaps. Instead, one needs to model a related kind of phenomenon: documents arrive in discrete batches; in each new batch of documents, some are relevant (in the present case, their titles contain a particular word w) and some are irrelevant. The idea is thus to find an automaton model that generates batched arrivals, with particular fractions of relevant documents. A sequence of batched arrivals could be considered bursty if the fraction of relevant documents alternates between reasonably long periods in which the fraction is small and other periods in which it is large.

IV. Conclusion

In this paper, we tackle the problem of mining common topics from multiple asynchronous text sequences. We propose a novel method which can automatically discover and fix potential asynchronism among sequences and consequentially extract better common topics. The key idea of our method is to introduce a self-refinement process by utilizing correlation between the semantic and temporal information in the sequences. It performs topic extraction and time synchronization alternately to optimize a unified objective function. A local optimum is guaranteed by our algorithm. We justified the effectiveness of our method on two real-world data sets, with comparison to a baseline method. Empirical results suggest that 1) our method is able to find meaningful and discriminative topics from asynchronous text sequences; 2) our method significantly outperforms the baseline method, evaluated both in quality and in quantity; 3) the performance of our method is robust and stable against different parameter settings and random initialization.

References Références Referencias

1. D.M. Blei and J.D. Lafferty, “Dynamic Topic Models,” Proc. Int’l Conf. Machine Learning (ICML), pp. 113-120, 2006.
2. 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.
3. 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.
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. 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.
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. 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. 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.
12. 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.
13. Y. Yang, T. Pierce, and J.G. Carbonell, “A Study of Retrospective and On-Line Event Detection,” Proc. Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 28-36, 1998.
14. T. Hofmann, “Probabilistic Latent Semantic Indexing,” Proc. Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 50-57, 1999.

This page is intentionally left blank