LPTA: A Probabilistic Model for Latent Periodic Topic Analysis

Zhijun Yin[1], Liangliang Cao[2], Jiawei Han[1], Chengxiang Zhai[1], Thomas Huang[2]
[1]Department of Computer Science
[2]Department of ECE and Beckman Institute
University of Illinois at Urbana-Champaign
zyin3@illinois.edu, cao4@ifp.uiuc.edu, hanj@cs.uiuc.edu, czhai@cs.uiuc.edu, huang@ifp.uiuc.edu

ABSTRACT

This paper studies the problem of latent periodic topic analysis from timestamped documents. The examples of timestamped documents include news articles, sales records, financial reports, TV programs, and more recently, posts from social media websites such as Flickr, Twitter, and Facebook. In this paper, we are interested in discovering latent periodic topics embedded in the timestamped documents. Different from detecting periodic patterns in traditional time series database, we discover the topics of coherent semantics and periodic characteristics where a topic is represented by a distribution of words. We propose a model called LPTA (Latent Periodic Topic Analysis) that exploits the periodicity of the terms as well as term co-occurrences. To show the effectiveness of our model, we collect several representative datasets including seminar, DBLP and Flickr. The results show that our model can discover the latent periodic topics effectively and leverage the information from both text and time well.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining

General Terms

Algorithm

Keywords

Periodic topics, topic modeling

1. INTRODUCTION

Periodic phenomena exist ubiquitously in our lives, and lots of natural and social topics have periodic recurring patterns. Hurricanes strike over the similar seasons every year. Many music and film festivals are held during similar periods annually. Sales offered by different brands culminate during Thanksgiving and Christmas every year. TV programs usually follow weekly schedules. Publicly traded companies are required to disclose information on an ongoing basis by submitting both annual reports and quarterly reports. Due to the prevalent existence of periodic topics, periodicity analysis is important in real world. Based on the discovered periodic patterns, people can not only analyze natural phenomena and human behavior, but also predict the future trends and help decision making.

Nowadays with the development of the Web, many text data exist with time information, e.g., news articles associated with their publishing dates, tagged photos annotated with their taken dates in Flickr[1] and published tweets along with their upload times in Twitter[2]. A lot of useful information is embedded in these text data, and it is interesting to discover topics that are periodic and characterize their temporal patterns.

Due to the importance of periodicity analysis, many research works have been proposed in periodicity detection for time series database [18, 9, 28, 7, 24]. Some studies follow the similar strategies to analyze the time distribution of a single tag or query to detect periodic patterns [23, 6, 20]. However, most of the existing studies are limited to time series database and cannot be applied on text data directly. First, a single word is not enough to describe a topic, and more words are needed to summarize a topic comprehensively. Second, analyzing the periodicity of single terms only is not sufficient to discover periodic topics. For example, the words like "music", "festival" and "chicago" may not have periodic patterns separately, but there may be periodic topics if these words are considered together. Third, there are synonyms and polysemy words due to the language diversity, which makes the problem even more challenging.

In this paper, we propose a model called LPTA (Latent Periodic Topic Analysis) to handle the above difficulties. Instead of analyzing periodicity based on the occurrence of single terms or patterns, our model exploits the periodicity of the terms as well as term co-occurrences, and in the end discovers the periodic topics where a topic is represented by a distribution of words. Our method can be viewed as a variant of latent topic models, where a document is generated by several latent topics which correspond to the semantic concepts of interests. Popular latent topic models include Probabilistic Latent Semantic Analysis (PLSA) [11], Latent Dirichlet Allocation (LDA) [4], and many variants of them (see Section 6 for a detailed review of these models). Unlike these traditional models, LPTA focuses on the periodic

KDD '11 San Diego, USA

[1] http://www.flickr.com
[2] http://twitter.com

Table 1: Notations used in the paper.

	Description
V	Vocabulary (word set), w is a word in V
D	Document collection
d	A document d that consists of words and timestamp
$\mathbf{w}_d$	The text of document d
t_d	The timestamp of document d
Z	The topic set, z is a topic in Z
θ	The word distribution set for Z

property in the time domain. The goal of learning LPTA is not only to find a latent topic space to fit the data corpus, but also detect whether a topic is periodic or not.

The contributions of this paper are summarized as follows.

1. We introduce the problem of latent periodic topic analysis that has not been studied before.
2. We propose the LPTA model to discover periodic topics by exploring both the periodic properties and the co-occurrence structures of the terms.
3. We perform extensive experiments on several representative datasets to demonstrate the effectiveness of our method.

The rest of the paper is organized as follows. We formulate the problem of latent periodic topic analysis in Section 2. We propose our LPTA model in Section 3. We analyze the complexity of the algorithm and illustrate its connection to other models in Section 4. We compare the performance of different methods in Section 5. We summarize the related work in Section 6 and conclude the paper in Section 7.

2. PROBLEM FORMULATION

In this section, we define the problem of latent periodic topic analysis. The notations used in this paper are listed in Table 1.

Definition 1. A **topic** is a semantically coherent theme, which is represented by a multinomial distribution of words. Formally, each topic z is represented by a word distribution $\theta_z = \{p(w|z)\}_{w \in V}$ s.t. $\sum_{w \in V} p(w|z) = 1$.

Definition 2. A **periodic topic** is a topic repeating in regular intervals. Formally, the conditional probability of time t given topic z, i.e., $p(t|z)$, follows periodic patterns in terms of periodic interval T. In order words, the timestamp distribution for each topic has bursts every interval T. Periodic interval T can be defined by users, such as 1 week(weekly), 1 month(monthly), 1 year(annually), etc.

Definition 3. A **timestamped document** is a text document associated with a timestamp. A timestamped document can be a news article along with its release date. It can also be a tweet associated with its publishing time in Twitter. Another example is a tagged photo uploaded to Flickr where the tags are considered as text and the time when the photo was taken is considered as its timestamp.

Given the definitions of timestamped document and periodic topic, we define the problem of latent periodic topic analysis as follows.

Definition 4. Given a collection of timestamped documents D, periodic interval T and the number of topics K, we would like to discover K periodic topics repeating every interval T, i.e., $\theta = \{\theta_z\}_{z \in Z}$ where Z is the topic set, along with their time distributions $\{p(t|z)\}_{z \in Z}$.

Here we give an example of latent periodic topic analysis.

***Example** 1.* Given a collection of photos related to *music festival* along with tags and timestamps in Flickr, the desired periodic topics are annual music festivals such as *South By Southwest* every March, *Coachella* every April, *Lollapalooza* every August, etc. As shown in Figure 1, the topic related to *Coachella* festival occurs in April every year. The top words in the topic are *coachella(0.1106), music(0.0915), indio(0.0719), california(0.0594)* and *concert(0.0357)* where the numbers in the parentheses are the weights of the corresponding words in $p(w|z)$ given z is the topic of *Coachella* festival.

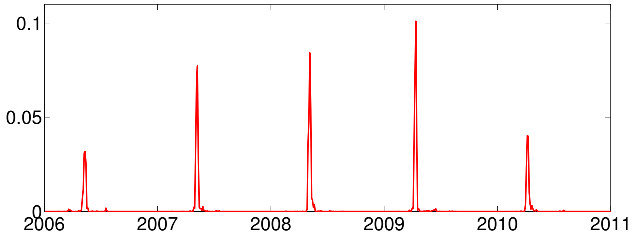


Figure 1: The distribution of the timestamps for the topic related to *Coachella* festival.

In the following sections, we present our model for latent periodic topic analysis.

3. LATENT PERIODIC TOPIC ANALYSIS

In this section, we propose our LPTA (Latent Periodic Topic Analysis) model. First, we introduce the general idea of our model. Second, we present the detail of our periodic topic generative process. Third, we introduce how to estimate the parameters.

3.1 General Idea

In general, the temporal patterns of topics can be classified into three types: periodic topics, background topics, and bursty topics. A periodic topic is one repeating in regular intervals; a background topic is one covered uniformly over the entire period; a bursty topic is a transient topic that is intensively covered only in a certain time period. We assume that a word is generated by a mixture of these topics and infer the most likely time domain behaviors. We will discuss how to model three kinds of topics and then study how to infer the mixture model. To encode the periodic topics, we take both the temporal structure and term co-occurrence into consideration. The words occurring around the same time in each period are likely to be clustered. If two words co-occur often in the same documents, they are more likely to belong to the same topic. In order to capture this property, we assume the timestamps of each periodic topic follow similar patterns in each period. Specifically, we model the distribution of timestamps for each periodic topic as a mixture of Gaussian distributions where the interval between the consecutive components is periodic interval T. In addition to periodic topics, the document collection may contain background words. In order to alleviate the problem of background noises, we model the background topics as well in our model. In particular, the timestamps of the background topics are generated by a uniform distribution. Other than periodic topics and background topics, we employ bursty topics to model patterns with bursting behavior in a short period but not regularly. The timestamps of the bursty topics are generated from a Gaussian distribution. Therefore, the document collection is modeled as a mixture of background topics, bursty topics and periodic topics. By fitting such a mixture model to timestamped text data, we can discover periodic topics along with their time distributions.

3.2 LPTA Framework

Let us denote the topic set as Z and the word distribution set as θ, i.e., $\{\theta_z\}_{z\in Z}$ where $\theta_z = \{p(w|z)\}_{w\in V}$ s.t. $\sum_{w\in V} p(w|z) = 1$. ϕ is the multinomial distributions for topics conditioned on documents, i.e., $\{\phi_d\}_{d\in D}$ where $\phi_d = \{p(z|d)\}_{z\in Z}$ s.t. $\sum_{z\in Z} p(z|d) = 1$. μ and σ are the collections of the means and standard deviations of timestamps for bursty topics and periodic topics. μ_z and σ_z are the mean and standard deviation of timestamps for topic z respectively. The generative procedure of latent periodic topic analysis model is described as follows.

To generate each word in document d from collection D:

1. Sample a topic z from multinomial ϕ_d.
   - **(a)** If z is a background topic, sample time t from a uniform distribution $[t_{start}, t_{end}]$, where t_{start} and t_{end} are the start time and end time of the document collection.
   - **(b)** If z is a bursty topic, sample t from $N(\mu_z, \sigma_z^2)$.
   - **(c)** If z is a periodic topic, sample period k of document d from a uniform distribution. Sample time t from $N(\mu_z + kT, \sigma_z^2)$, where T is periodic interval.
2. Sample a word w from multinomial θ_z.

Given the data collection $\{(\mathbf{w}_d, t_d)\}_{d\in D}$ where $\mathbf{w}_d$ is the word set in document d and t_d is the timestamp of document d, the log-likelihood of the collection given $\Psi = \{\theta, \phi, \mu, \sigma\}$ is as follows.

$$\begin{aligned} L(\Psi; D) &= \log p(D|\Psi) \\ &= \log \prod_{d\in D} p(\mathbf{w}_d, t_d|\Psi) \end{aligned} \tag{1}$$

$$p(\mathbf{w}_d, t_d|\Psi) = \sum_d \sum_w n(d,w) \log \sum_z p(t_d|z)p(w|z)p(z|d) \tag{2}$$

where $n(d, w)$ is the count of word w in document d.

If topic z is a background topic, $p(t|z)$ is modeled as a uniform distribution:

$$p(t|z) = \frac{1}{t_{end} - t_{start}} \tag{3}$$

If topic z is a bursty topic, $p(t|z)$ is modeled as a Gaussian distribution:

$$p(t|z) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z)^2}{\sigma_z^2}} \tag{4}$$

If topic z is a periodic topic, $p(t|z)$ is modeled as a mixture of Gaussian distributions:

$$p(t|z) = \sum_k p(t|z,k)p(k) \tag{5}$$

where k is the period id, $p(t|z,k) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z-kT)^2}{\sigma_z^2}}$ and $p(k)$ is uniform in terms of k.

3.3 Parameter Estimation

In order to estimate parameters Ψ in Equation 1, we use maximum likelihood estimation. Specifically, we use Expectation Maximization(EM) algorithm to solve the problem, which iteratively computes a local maximum of likelihood. We introduce the probability of the hidden variable $p(z|d, \Psi)$, which is the probability that document d belongs to topic z given Ψ. In the E-step, it computes the expectation of the complete likelihood $Q(\Psi|\Psi^{(t)})$, where $\Psi^{(t)}$ is the value of Ψ estimated in iteration t. In the M-step, it finds the estimation $\Psi^{(t+1)}$ that maximizes the expectation of the complete likelihood.

In the **E-step**, $p(z|d, w)$ is updated according to Bayes formula as in Equation 6.

$$p(z|d,w) = \frac{p(t_d|z)p(w|z)p(z|d)}{\sum_z p(t_d|z)p(w|z)p(z|d)} \tag{6}$$

In the **M-step**, $p(w|z)$ and $p(z|d)$ are updated as follows.

$$p(w|z) = \frac{\sum_d n(d,w)p(z|d,w)}{\sum_d \sum_{w'} n(d,w')p(z|d,w')} \tag{7}$$

$$p(z|d) = \frac{\sum_w n(d,w)p(z|d,w)}{\sum_w \sum_{z'} n(d,w)p(z'|d,w)} \tag{8}$$

If topic z is bursty topic, μ_z and σ_z are updated accordingly as follows.

$$\mu_z = \frac{\sum_d \sum_w n(d,w)p(z|d,w)t_d}{\sum_d \sum_w n(d,w)p(z|d,w)} \tag{9}$$

$$\sigma_z = (\frac{\sum_d \sum_w n(d,w)p(z|d,w)(t_d - \mu_z)^2}{\sum_d \sum_w n(d,w)p(z|d,w)})^{1/2} \tag{10}$$

If topic z is a periodic topic, we partition the time line into intervals of length T and assume that each document is only related to its corresponding interval. In other words, $p(t_d|z, k)$ is set as 0 if document d is not in the k-th interval. μ_z and σ_z for periodic topic z can be updated according to the following steps.

$$\mu_z = \frac{\sum_d \sum_w n(d,w)p(z|d,w)(t_d - I_d T)}{\sum_d \sum_w n(d,w)p(z|d,w)} \tag{11}$$

$$\sigma_z = (\frac{\sum_d \sum_w n(d,w)p(z|d,w)(t_d - \mu_z - I_d T)^2}{\sum_d \sum_w n(d,w)p(z|d,w)})^{1/2} \tag{12}$$

where I_d is the corresponding interval of document d.

4. DISCUSSION

4.1 Complexity Analysis

We analyze the complexity of parameter estimation process in Section 3.3. In the E-step, it needs $O(K|W|)$ to calculate $p(z|d, w)$ in Equation 6 for all (z, d, w) triples, where K is the number of topics and $|W|$ is the total counts of the words in all the documents. In the M-step, it needs $O(K|W|)$ to update $p(w|z)$ according to Equation 7 for all (w, z) pairs and $O(K|W|)$ to update $p(z|d)$ according to Equation 8 for all (z, d) pairs. It needs $O(|W|)$ to update μ_z in Equation 9 and $O(|W|)$ to update σ_z in Equation 10 for each bursty topic z. Similarly, it needs $O(|W|)$ to update μ_z in Equation 11 and $O(|W|)$ to update σ_z in Equation 12 for each periodic topic z. Therefore, the complexity of the LPTA model is $O(iterK|W|)$, where $iter$ is the number of the iterations in the EM algorithm.

4.2 Parameter Setting

In LPTA, we have two types of parameters, i.e., the number of topics K and the length of periodic interval T. Users can specify the value of K according to their needs. For example, if topics of finer granularity are to be discovered, K can be set to a relatively large number, whereas if topics of coarser granularity are desired, K can be set to a relatively small value. When the parameters are unknown, Schwarz's Bayesian information criterion (BIC) provides an efficient way to select the parameters. The BIC measure includes two parts: the log-likelihood and the model complexity. The first part characterizes the fitness over the observations, while the second is determined by the number of parameters. In practice we can train models with different parameters, and compare their BIC values. The model with the lowest value will be selected as the final model. For periodic interval T, users can specify as 1 week (for weekly topics), 1 year (for annual topics), etc. Besides, instead of fixing the periodic interval as one value, we can also make a mixture of topics with different periodic intervals. In this way, we can discover the topics of different periodic intervals simultaneously. Specifically, a bursty topic can be considered as a periodic topic with only one period during the entire time span. We will study how to extract the periodic interval automatically in future work.

4.3 Connections to Other Models

Probabilistic Latent Semantic Analysis PLSA is a latent variable model for co-occurrence data which associates an unobserved topic variable with the occurrence of a word in a particular document [11]. PLSA does not consider the time information, and it can be considered as a special case of our LPTA model when all the topics are background topics.

Retrospective News Event Detection RED is a probabilistic model to incorporate both content and time information to detect retrospective news event [17]. Although RED models the time information into the framework, it can only detect bursty topics with unigram models. RED can be considered as a simplified version of our LPTA framework, which contains bursty topics only and uses a mixture of unigram models.

Topic Over Time TOT is an LDA-style generative model to extract the evolutionary topic patterns in timestamped documents [25]. In our model LPTA, we model background topics, bursty topics as well as periodic topics. Compared with TOT, LPTA focuses on *recurring periodic* topic patterns instead of the evolution of the topics.

5. EXPERIMENT

In this section, we demonstrate the evaluation results of our method. First, we introduce the datasets used in the experiment. Second, we compare our method with other methods on these datasets qualitatively. Third, we use multiple measures including accuracy and normalized mutual information to evaluate our method quantitatively.

5.1 Datasets

In this paper, we evaluate our ideas on several representative datasets from real life to social media.

- *Seminar* We collected the weekly seminar announcements for one semester from six research groups in computer science department at University of Illinois at Urbana-Champaign[3]. The research groups include AIIS (Artificial Intelligence and Information Systems), DAIS (Database and Information Systems), Graphics, HCI, Theory and UPCRC (Universal Parallel Computing Research Center). The seminar time is considered as the document timestamp. We would like to discover weekly topics, so we set periodic interval T as 1 week. The resulting dataset has 61 documents and 901 unique words.
- *DBLP* Digital Bibliography Project (DBLP)[4] is a computer science bibliography. We collected the paper titles of several different conferences from 2003 to 2007. The conferences include WWW, SIGMOD, SIGIR, KDD, VLDB and NIPS. The timestamps of the documents are determined according to the conference programs. We would like to discover annual topics, so we set periodic interval T as 1 year. The resulting dataset has 4070 documents and 2132 unique words.
- *Flickr* Flickr is an online photo sharing website. We crawled images through Flickr API[5]. The tags of a photo are considered as document text, while the time when the photo was taken is considered as document timestamp. Specifically, we crawled the photos for several music festivals from 2006 to 2010 including SXSW (South by Southwest), Coachella, Bonnaroo, Lollapalooza and ACL (Austin City Limits). We would like to discover annual topics, so we set periodic interval T as 1 year. The resulting dataset has 84244 documents and 7524 unique words.

5.2 Qualitative Evaluation

5.2.1 Topics Discovered by LPTA

We set the number of periodic topics as 6 in both Seminar and DBLP datasets and 5 in Flickr dataset according to our construction of these datasets. We evaluate the change of the number of topics in quantitative evaluation in Section 5.3. We list selected topics discovered by LPTA in different datasets in Table 2. In Seminar dataset, LPTA can effectively discover the topics for different research groups and their corresponding seminar time. For example, Topic 1 is *DAIS* at 16:00 every Tuesday, where *data*, *text* and *mining* are the popular words. Topic 2 is *AIIS* at 14:00 every Friday, which focuses on *machine learning*. In DBLP dataset, LPTA can identify six periodic topics, i.e., six annual conferences. For example, Topic 1 is *KDD* in August, which focuses on *data mining*. Topic 2 is *SIGIR*. The terms like *retrieval*, *search*, *relevance* and *evaluation* are the core topics in *SIGIR*. In Flickr dataset, LPTA can clearly detect the music festivals as well as their durations. For example, Topic 1 is about *ACL*, which is held around late September in *zilker park*, *austin*, *texas*. Since the dates that *ACL* took place were not fixed every year, i.e., Sep 15-17 in 2006, Sep

[3] http://cs.illinois.edu/

[4] http://www.informatik.uni-trier.de/~ley/db/

[5] http://www.flickr.com/services/api/

14-16 in 2007, Sep 26-28 in 2008, Oct 2-4 in 2009 and Oct 8-10 in 2010, the standard deviation of the timestamps is 10d13h20m. Topic 2 is *Bonnaroo* in *manchester*, *tennessee*. Since the dates of *Bonnaroo* did not vary too much every year, the standard deviation of the timestamps for *Bonnaroo* is only 2d14h21m.

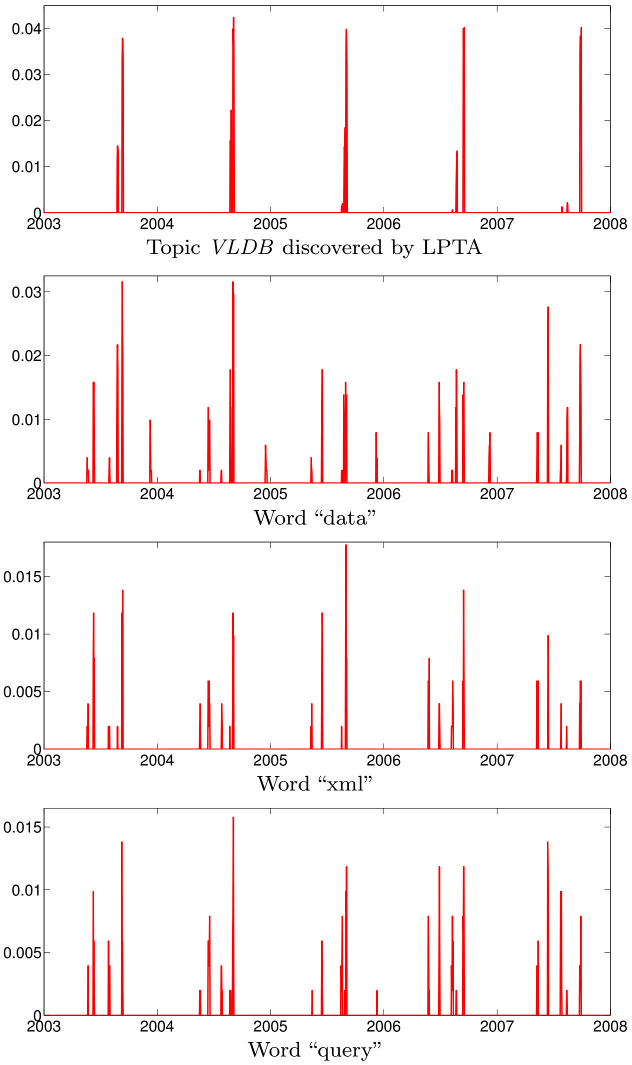
5.2.2 LPTA vs. Periodicity Detection

In order to see whether pooling together related words is better than analyzing periodicity at single word level, we make a comparison between LPTA and periodicity detection method. We attempt to detect periodic words by periodicity detection algorithm. Fourier decomposition represents a sequence as a linear combination of the complex sinusoids. To identify the power content of each frequency, the power spectral density PSD (or power spectrum) of a sequence is used to indicate the signal power at each frequency in the spectrum [23]. A well known estimator of the PSD is the periodogram, which is a vector comprised of the squared magnitude of the Fourier coefficients. We use AUTOPERIOD [24], a two-tier approach by considering the information in both the autocorrelation and the periodogram, to detect periods for each word. Unfortunately, the method fails to detect meaningful periodic words because the time series are sparse and few words have apparent periodic patterns in the datasets. Most of the words do not occur periodically without considering the topics.

Compared with single word representation, LPTA uses multiple words to describe a topic. For example, in DBLP dataset, LPTA discovers topic *VLDB* with the word distribution *data 0.0530, xml 0.0208, query 0.0196, queries 0.0176, efficient 0.0151, mining 0.0142, database 0.0136, based 0.0128, streams 0.0112, databases 0.0111.* We can see that a single word may not be enough to represent such a topic and multiple words can represent a topic better. LPTA can not only provide a more comprehensive description of the topic, but also discover the periodic topic when its consisting words do not have periodic patterns separately. In LPTA, we can plot the time distributions of the discovered topics based on $p(d|z)$ and document timestamps, where $p(d|z)$ can be obtained from $p(z|d)$ according to Bayes' theorem. In Figure 2, we plot the time distribution of topic *VLDB* in DBLP dataset as well as the time distributions of word *data*, *xml* and *query* which are the top popular words in the topic. We can see that topic *VLDB* has the clear periodic patterns while *data*, *xml* and *query* do not occur periodically. It shows that LPTA can discover the periodic topics effectively even if its consisting words do not have periodic patterns by themselves.

5.2.3 LPTA vs. Topic Models

In order to see whether traditional topic models can detect meaningful topics, we compare the results of topic modeling methods including PLSA and LDA with the one of LPTA. We set the number of topics as 6 in both Seminar and DBLP datasets and 5 in Flickr dataset for both PLSA and LDA. We list several selected topics by using PLSA and LDA in Table 3. Since the words in computer science areas are closely related, PLSA and LDA cannot identify the topics of different research areas in Seminar dataset. In DBLP dataset, the conferences are from the areas including database, data mining, information retrieval, Web and machine learning. All these topics are similar, so both PLSA and LDA cannot discover the meaningful topic clusters. In Flickr dataset, PLSA mixes several music festivals together. For example, both *southbysouthwest* and *coachella* appear in Topic 1, and in Topic 2 *lollapalooza* and *austincitylimits* are merged together. We find that LDA performs better than PLSA in this dataset. LDA can discover several festivals although it mixes *coachella* and *bonnaroo* in Topic 1. Compared with the result of LPTA in Table 2, we can see that LPTA can discover the meaningful topics of better quality.



Topic *VLDB* discovered by LPTA

Word "data"

Word "xml"

Word "query"

Figure 2: Time distribution of topic *VLDB* discovered by LPTA and time distributions of the words in the topic.

5.2.4 Integration of Text and Time Information

To demonstrate the effectiveness of LPTA model for combining the information of both text and time, we study the following two specific cases in DBLP dataset.

SIGMOD vs. VLDB SIGMOD and VLDB are two reputed conferences in database area, and the concentrated topics in these two conferences are similar. Therefore, it is difficult to differentiate these two conferences based on text only. However, SIGMOD is usually held in June, while VLDB is usually held in September. In LPTA, we discover the periodic topics by considering the information from both

Table 2: Selected periodic topics discovered by using LPTA. The date and the duration in the parentheses are the mean and standard deviation of the timestamps for the corresponding periodic topic.

Seminar		DBLP		Flickr	
Topic 1 (DAIS)	Topic 2 (AIIS)	Topic 1 (KDD)	Topic 2(SIGIR)	Topic 1 (ACL)	Topic 2 (Bonnaroo)
Tue 16:00 (0h0m0s)	Fri 14:00 (0h0m0s)	Aug 23 (10d3h11m)	Aug 3 (9d6h56m)	Sep 29 (10d13h20m)	Jun 16 (2d14h21m)
model 0.0166	computer 0.0168	mining 0.0353	retrieval 0.0495	acl 0.0945	bonnaroo 0.1066
based 0.0158	learning 0.0158	data 0.0289	based 0.0197	austin 0.0827	music 0.0870
mining 0.0151	machine 0.0138	search 0.0233	web 0.0189	music 0.0763	manchester 0.0587
text 0.0143	science 0.0128	clustering 0.0208	text 0.0171	austincitylim. 0.0442	tennessee 0.0518
network 0.0135	algorithms 0.0128	based 0.0195	query 0.0164	limits 0.0441	live 0.0327
web 0.0119	language 0.0118	web 0.0168	search 0.0162	city 0.0441	concert 0.0275
problem 0.0111	work 0.0108	learning 0.0159	document 0.0149	texas 0.0426	arts 0.0175
data 0.0111	problems 0.0108	networks 0.0114	language 0.0118	concert 0.0283	performance 0.0174
query 0.0111	models 0.0108	analysis 0.0105	relevance 0.0111	live 0.0212	backstagegall. 0.0113
latent 0.0095	prediction 0.0108	large 0.0104	evaluation 0.0111	zilker 0.0173	rock 0.0111

Table 3: Selected topics discovered for different datasets by using PLSA and LDA.

Seminar				DBLP				Flickr			
PLSA		LDA		PLSA		LDA		PLSA		LDA	
Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2
data	memory	problem	systems	web	search	web	system	sxsw	lollapaloo.	music	lollapaloo.
latent	computer	algorithm	computer	data	text	mining	database	austin	music	coachella	music
visualizati.	data	network	science	xml	databases	semantic	distributed	music	chicago	bonnaroo	chicago
intel	mining	graph	algorithms	queries	relational	detection	user	texas	concert	california	live
talk	parallel	time	time	mining	user	automatic	adaptive	southbyso.	acl	manchester	concert
analysis	science	networks	agent	semantic	analysis	services	content	live	grantpark	indio	grantpark
computer	pattern	influence	visualizati.	search	ranking	applicatic.	relevance	atx	live	tennessee	august
systems	programm.	online	data	streams	structure	graph	performan.	coachella	austincity.	arts	photos
machine	hardware	work	engineering	managem.	support	extraction	feedback	downtown	august	art	summer
visual	algorithms	question	function	adaptive	evaluation	patterns	image	livemusic	austin	palmsprin.	performan.

text and time, so we can easily identify these two topics. We set the number of periodic topics as 2 and show the topics in Table 4. As we can see from Table 4, Topic 1 is *SIGMOD* on Jun 17 with the standard deviation 7d11h6m and Topic 2 is *VLDB* on Sep 11 with the standard deviation 9d5h29m. Although the popular words in both of the topics are *data*, *query* and *xml*, these two topics can be clustered because the timestamps form two clusters.

SIGMOD vs. CVPR SIGMOD and CVPR are held in June, so it is difficult to differentiate these two if we rely on time information only. However, SIGMOD is a database conference while CVPR is a computer vision conference. Therefore, in this case, text information will help identify these two topics even though the timestamps of these two topics overlap with each other. We set the number of periodic topics as 2 and show the topics in Table 4. As we can see from Table 4, Topic 1 is *SIGMOD* with its focus on *data*, *query* and *xml*, and Topic 2 is *CVPR* focusing on *image*, *recognition*, *tracking*, *detection* and *segmentation*.

Table 4: Periodic topics for SIGMOD vs. VLDB and SIGMOD vs. CVPR datasets by using LPTA.

SIGMOD vs. VLDB		SIGMOD vs. CVPR	
Topic 1	Topic 2	Topic 1	Topic 2
(SIGMOD)	(VLDB)	(SIGMOD)	(CVPR)
Jun 17	Sep 11	Jun 20	Jun 21
(7d11h6m)	(9d5h29m)	(7d15h42m)	(3d4h37m)
data	data	data	image
query	xml	query	based
xml	query	xml	tracking
database	queries	database	recognition
processing	efficient	processing	learning
efficient	database	efficient	object
databases	based	based	shape
queries	databases	databases	detection
web	system	queries	motion
system	processing	queries	motion

5.2.5 Periodic vs. Bursty Topics

To demonstrate the effectiveness of LPTA model for balancing periodic and bursty topics, we study the following case in Flickr dataset. Instead of pooling the photos related to music festivals all together, we keep the photos related to SXSW and ACL festivals from 2006 to 2010 and those related to Coachella and Lollapalooza in 2009 only. In this way, we simulate the dataset with 2 periodic topics and 2 bursty topics. We set the number of periodic topics as 2 and the number of bursty topics as 2 in LPTA and show the topics in Table 5. From Table 5, we can see that the words recurring during similar periods every year like *sxsw* and *acl* fit into two corresponding periodic topics (i.e., Topic 1 and Topic 2), while the words that occur only in one period like *lollapalooza*, *chicago*, *coachella* and *indio* fit into two corresponding bursty topics (i.e., Topic 3 and Topic 4). LPTA can differentiate between the bursty topics and periodic topics in this dataset. The mean dates for periodic topics *SXSW* and *ACL* are Mar 18 and Sep 28 every year, and the mean dates for bursty topics *Lollapalooza* and *Coachella* are Aug 8 2009 and Apr 17 2009, respectively.

5.2.6 Summary

From the above qualitative evaluation, we can see that compared with periodicity detection for every single word, LPTA can not only provide a more comprehensive description of a topic, but also discover the periodic topic even when its consisting words do not have periodic patterns separately. Compared with topic modeling methods including PLSA and LDA, LPTA can discover the periodic topics with more meaningful semantics. Besides, LPTA can identify the mean date and its standard deviation for each periodic topic

Table 5: Topics discovered for periodic vs. bursty dataset by using LPTA.

Bursty topics		Periodic topics	
Topic 1 (Lollapalooza) Aug 8 2009 (1d0h12m)	Topic 2 (Coachella) Apr 17 2009 (10d20h23m)	Topic 3 (SXSW) Mar 18 (6d8h33m)	Topic 4 (ACL) Sep 28 (14d7h22m)
lollapalooza	coachella	sxsw	acl
chicago	indio	austin	austin
concert	music	texas	music
music	california	music	austincityli.
grantpark	concert	southbysouth.	city
august	live	live	limits
live	desert	concert	texas
illinois	art	atx	concert
performance	musicfestival	downtown	live
lolla	livemusic	gig	zilker

effectively. From the SIGMOD vs. VLDB and SIGMOD vs. CVPR datasets in DBLP, we can see that it is difficult to discover meaningful topics without the combination of text and time information and LPTA achieves good balance between these two. With regards to the tradeoff between periodic topics and bursty topics, from periodic vs. bursty dataset in Flickr, we can see that the words will fit into the corresponding periodic or bursty topics if they have periodic or bursty patterns.

5.3 Quantitative Evaluation

5.3.1 Evaluation Metric

To evaluate the results quantitatively, we provide some evaluation metrics to compare the results. The latent topics discovered by the topic modeling approaches can be regarded as clusters. Based on the estimated conditional probability of topic z given document d, i.e., $p(z|d)$, we can infer the cluster label for document d. Therefore, accuracy (AC) and normalized mutual information (NMI) can be used to measure the clustering performance [5]. Given document d, its label l_d in the dataset and the topic z_d for document d obtained from the topic modeling approach, accuracy is defined as follows.

$$AC = \frac{\sum_d \delta(l_d, map(z_d))}{|D|}$$

where $|D|$ is the number of all the documents and $\delta(x, y)$ is the delta function that is one if $x = y$ and is zero otherwise, and $map(z_d)$ is the permutation mapping function that maps the topic z_d of document d to the corresponding label in the dataset. The best mapping between the topics and document labels can be found by Kuhn-Munkres algorithm [14].

We denote L as the set of document labels obtained from the dataset and Z as the topics obtained from the topic modeling approaches. The mutual information metric $MI(L, Z)$ between L and Z is defined as follows.

$$MI(L, Z) = \sum_{l \in L, z \in Z} p(l, z) \log \frac{p(l, z)}{p(l)p(z)}$$

where $p(l)$ and $p(z)$ are the probabilities that a document arbitrarily selected from the dataset has label l or belongs to topic z, and $p(l, z)$ is the joint probability that the arbitrarily selected document has label l and belongs to topic z. The normalized mutual information NMI is defined as follows.

$$NMI(L, Z) = \frac{MI(L, Z)}{\max(H(L), H(Z))}$$

where $H(L)$ and $H(Z)$ are the entropies of L and Z. Specifically, $NMI = 1$ if L and Z are identical, and $NMI = 0$ if L and Z are independent.

5.3.2 Performance Evaluations and Comparisons

In Table 6, we list the comparison of accuracy and normalized mutual information by using different methods in different datasets. We vary the number of topics from 2 to 10. From Table 6, we can see that LPTA performs significantly better than PLSA and LDA. In average, LDA performs better than PLSA, but is not as good as LPTA. It demonstrates that LPTA makes good use of the text and time information. Accuracy and NMI of PLSA and LDA in Flickr dataset are higher compared with other datasets. The reason is that the topical clusters are relatively apparent in Flickr while the clusters are not clear in both Seminar and DBLP datasets. In Seminar and DBLP datasets, the words are related to computer science, it is difficult to differentiate subjects in various research areas. Especially in DBLP dataset, the conferences from database, data mining, information retrieval and machine learning are closely related to each other, it is difficult to cluster them without considering the periodic patterns, which explains why accuracy and NMI in DBLP dataset are extremely low. However, LPTA is stable and has relatively high values in both accuracy and NMI in all the datasets, because LPTA leverages both the topical clusters and periodic patterns.

6. RELATED WORK

In this section we discuss related work to our study, including temporal topic mining, event detection and tracking and periodic pattern mining.

Temporal topic mining Topic modeling is a classic problem in text mining. The representative algorithms include PLSA [11] and LDA [4]. Besides modeling the text itself, many other methods have been proposed to mine topics from document associated with timestamps. Wang et al. [25] used an LDA-style topic model to capture both the topic structure and the changes over time. Mei et al. [19] partitioned the timeline into buckets and proposed a probabilistic approach to model the subtopic themes and spatiotemporal theme patterns simultaneously in weblogs. Wang et al. mined correlated bursty topic patterns from coordinated text streams in [26]. Blei and Lafferty [3] employed state space models on the natural parameters of multinomial distributions of topics and design a dynamic topic model to model the time evolution of stream. Iwata et al. [12] proposed an online topic model for sequentially analyzing the time evolution of topics in document collections, in which current topic-specific distributions over words are assumed to be generated based on the multiscale word distributions of the previous epoch. Stochastic EM algorithm was used in the online inference process. In [30], Zhang et al. discovered different evolving patterns of clusters, including emergence, disappearance, evolution within a corpus and across different corpora. The problem was formulated as a series of hierarchical Dirichlet processes by adding time dependencies to the adjacent epochs, and a cascaded Gibbs sampling scheme is used to infer the model. All the existing studies on tem-

Table 6: Accuracy and Normalized Mutual Information in different datasets by using different methods.

K	Seminar						DBLP						Flickr					
	Accuracy(%)			NMI(%)			Accuracy(%)			NMI(%)			Accuracy(%)			NMI(%)		
	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA
2	31.1	31.8	**37.7**	11.7	12.3	**34.7**	24.2	25.4	**38.3**	1.9	2.8	**23.9**	45.7	48.9	**49.7**	22.4	28.3	**37.2**
3	37.0	38.0	**51.0**	19.0	19.9	**53.0**	26.8	26.8	**51.1**	3.6	3.8	**45.7**	57.7	59.9	**63.1**	35.9	42.1	**54.9**
4	39.4	41.3	**65.4**	23.6	24.0	**70.7**	26.5	27.7	**61.5**	3.8	4.5	**56.7**	63.7	70.6	**74.8**	42.2	53.8	**67.4**
5	40.1	42.1	**78.5**	25.7	26.6	**82.4**	27.1	28.7	**66.1**	4.5	5.6	**63.0**	69.2	74.8	**85.7**	48.6	59.9	**79.2**
6	43.0	41.9	**90.4**	30.6	28.9	**92.3**	26.6	27.8	**67.8**	4.7	5.7	**65.9**	67.6	78.5	**90.2**	47.9	60.2	**82.1**
7	40.8	39.5	**94.5**	30.5	29.7	**94.2**	24.0	26.2	**65.9**	4.3	5.8	**63.8**	67.2	71.5	**89.6**	46.5	54.3	**80.2**
8	39.0	40.0	**91.9**	30.4	31.0	**91.7**	22.3	23.9	**66.7**	4.4	5.6	**63.1**	66.0	69.8	**86.5**	45.7	53.1	**77.6**
9	35.3	36.9	**90.0**	30.5	30.8	**88.8**	20.8	22.3	**65.1**	4.4	5.6	**60.8**	64.2	64.5	**83.7**	44.3	50.6	**74.7**
10	34.9	33.9	**88.1**	31.7	30.2	**86.8**	19.6	20.6	**63.6**	4.5	5.5	**58.2**	63.1	67.7	**81.4**	43.5	51.4	**73.1**
Avg	37.9	38.4	**76.4**	26.0	26.0	**77.2**	24.2	25.5	**60.7**	4.0	5.0	**55.7**	62.7	67.3	**78.3**	41.9	50.4	**69.6**

poral topic mining focus on the evolutionary pattern of the topics, but they do not study the periodic topics. Instead of studying the evolution of the topics, we focus on periodic topic patterns in this paper.

Event detection and tracking In [1], Allan et al. introduced the problems of event detection and tracking within a stream of broadcast news stories. To extract meaningful structure from document streams that arrive continuously over time is a fundamental problem in text mining [13]. Kleinberg developed a formal approach for modeling the stream using an infinite-state automaton to identify the bursts efficiently. Fung et al. [8] proposed Time Driven Documents-partition framework to construct a feature-based event hierarchy for a text corpus based on a given query. In [17], Li et al. proposed a probabilistic model to incorporate both content and time information in a unified framework to detect the retrospective news events. In [10], He et al. used concepts from physics to model bursts as intervals of increasing momentum, which provided a new view of bursty patterns. Besides traditional text documents like news articles and research publications, event detection is also studied in those new social media like Twitter and Flickr [2, 21]. Becker et al. [2] explored a variety of techniques for learning multi-feature similarity metrics for social media documents to detect events. In [21], Sakaki et al. proposed an algorithm to monitor tweets to detect real time events such as earthquakes and typhoons. In [15], Leskovec et al. proposed a meme-tracking approach to provide a coherent representation of the news cycle, i.e., daily rhythms in the news media. Yang et al. [27] studied temporal patterns associated with online content and how the content's popularity grows and fades over time. These studies of event detection and tracking focus on mining temporal bursts instead of analyzing the topics that have periodic patterns.

Periodic pattern mining Some studies focus on searching periodic patterns in time-series databases [9, 28, 29, 7]. Besides traditional time-series databases, some other studies detect periodic events on other datasets such as video [22] and moving objects [16]. Compared with the above studies, our paper focuses on the latent periodic topic analysis on the text dataset. Some work studies periodic analysis in text domain [23, 6, 20]. In [6], Chen et al. analyzed spatial temporal distributions of tag usage to detect events from photos on Flickr and extracted the periodic tags by checking the standard deviation of the distances between every two adjacent entries in the timeline for each tag and clustered the tags into events. In [20], Murata et al. classified queries based on the number of search intentions and their temporal features, and performed the Discrete Fourier Transform(DFT) on the ratios of each search intention to detect the periodic changes. However, these studies analyze the distributions of single terms only. In this paper, we model the latent periodic topic analysis in a more systematic way where each topic is represented by a word distribution. We analyze the periodic patterns from the perspective of topics instead of single words, and we discover the periodic bursts and the corresponding topics together instead of making the process into two separate stages.

7. CONCLUSION AND FUTURE WORK

In this paper, we introduce the problem of latent periodic topic analysis on timestamped documents. We propose a model called LPTA (Latent Periodic Topic Analysis) that exploits both the periodicity of the terms and term co-occurrences. To test our approach, we collect several representative datasets including seminar, DBLP and Flickr. Evaluation results show that our LPTA model works well for discovering the latent periodic topics by combining the information from topical clusters and periodic patterns.

Periodicity analysis is an important task for web mining and social media mining. In the future we will focus on how to extend our current work to handle the increasing amount of web documents and complex structure of social media. We are especially interested in three scenarios:

- Effectively analyzing large scale data. Although we have tested our model in quite a few datasets, these datasets are relatively small compared with web-scale information resources. We are interested in designing scalable algorithms that can also handle the potentially noisy data in real life.

- Automatically determining the optimal number of topics in real life. In our current model, the number of topics is given as a parameter. In the future, we plan to use Bayesian information criterion to select the optimal number of topics or employ Dirichlet process for model selection.

- Incorporating the social networks into periodicity detection. In our current scheme, document are treated isolatedly and we do not consider whether these documents come from the same user or users who are close friends. In social media websites such as Flickr and Twitter, the social network plays an important role and incorporates rich information. In the future we would like to combine such network structure for analysis.

8. REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR*, pages 37–45, 1998.

[2] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM*, pages 291–300, 2010.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, pages 113–120, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *CIKM*, pages 911–920, 2008.

[6] L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *CIKM*, pages 523–532, 2009.

[7] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid. Periodicity detection in time series databases. *IEEE Trans. Knowl. Data Eng.*, 17(7):875–887, 2005.

[8] G. P. C. Fung, J. X. Yu, H. Liu, and P. S. Yu. Time-dependent event hierarchy construction. In *KDD*, pages 300–309, 2007.

[9] J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. In *ICDE*, pages 106–115, 1999.

[10] D. He and D. S. Parker. Topic dynamics: an alternative model of bursts in streams of topics. In *KDD*, pages 443–452, 2010.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[12] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Online multiscale dynamic topic models. In *KDD*, pages 663–672, 2010.

[13] J. M. Kleinberg. Bursty and hierarchical structure in streams. In *KDD*, pages 91–101, 2002.

[14] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[15] J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[16] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. In *KDD*, pages 1099–1108, 2010.

[17] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *SIGIR*, pages 106–113, 2005.

[18] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering frequent episodes in sequences. In *KDD*, pages 210–215, 1995.

[19] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

[20] M. Murata, H. Toda, Y. Matsuura, R. Kataoka, and T. Mochizuki. Detecting periodic changes in search intentions in a search engine. In *CIKM*, pages 1525–1528, 2010.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[22] E. P. Vivek, E. Pogalin, and A. W. M. Smeulders. Periodic event detection and recognition in video. In *ICASSP*, pages 3537–3540, 2009.

[23] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *SIGMOD Conference*, pages 131–142, 2004.

[24] M. Vlachos, P. S. Yu, and V. Castelli. On periodicity detection and structural periodic similarity. In *SDM*, 2005.

[25] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, pages 424–433, 2006.

[26] X. Wang, C. Zhai, X. Hu, and R. Sproat. Mining correlated bursty topic patterns from coordinated text streams. In *KDD*, pages 784–793, 2007.

[27] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, 2011.

[28] J. Yang, W. Wang, and P. S. Yu. Mining asynchronous periodic patterns in time series data. *IEEE Trans. Knowl. Data Eng.*, 15(3):613–628, 2003.

[29] J. Yang, W. Wang, and P. S. Yu. Mining surprising periodic patterns. *Data Min. Knowl. Discov.*, 9(2):189–216, 2004.

[30] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *KDD*, pages 1079–1088, 2010.