A Probabilistic Model for Latent Periodic Topic Analysis

Zhijun Yin[1], Liangliang Cao[2], Jiawei Han[1],Chengxiang Zhai[1], Thomas Huang[1]

[1]University of Illinois at Urbana-Champaign

2IBM T.J. Watson Research Center

DAIS The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign

Large Scale Information Management

Periodic Phenomena Exist Ubiquitously

- Hurricanes
- Music and film festivals
- Product sales
- TV program
- Publicly traded company

Many Text Data Exist with Time Information

- News articles associated with their publishing dates
- Tagged photos annotated with their taken dates in Flickr
- Tweets published with their upload times in Twitter

Apply Periodicity Analysis on Text Data

- Periodicity detection for time series database
- Some studies follow the similar strategies to analyze the time distribution of a single tag or query to detect periodic patterns

Challenges

- A single word is not enough to describe a topic and more words are needed to summarize a topic comprehensively
- Analyzing the periodicity of single terms only is not sufficient to discover periodic topics
  - E.g., the words like "*music*", "*festival*" and "*chicago*" may not have periodic patterns if considered separately, but there may be periodic topics if these words are considered together
- Synonyms and polysemy words due to the language diversity

Latent Periodic Topic Analysis

Input:
Timestamped documents

ID	Text	Date
1	coachella, music, arts, festival, …	Apr 27 2008
2	sxsw, south by southwest, austin, …	Mar 14 2008
…	…	…

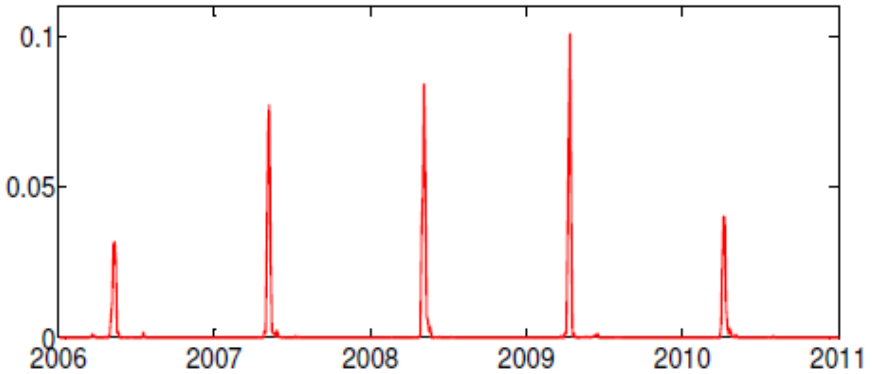
Periodic interval T
e.g., 1 year, etc.

Output:

1. Periodic topics: { p(w|z) }

Topic 1 (Coachella Festival)	…
coachella 0.1106	…
music 0.0915	…
indio 0.0719	…
california 0.0594	…
concert 0.0357	…
…	…

2. Time distribution of topics



The distribution of the timestamps for the topic related to Coachella festival

Problem Formulation

- Input:
  - A collection of timestamped documents D
  - The number of topics K
  - Periodic interval T
- Output:
  - K periodic topics $\theta = \{\theta_z\}_{z \in Z}$
    - $\theta_z = \{p(w|z)\}_{w \in V}$ p(w|z) is the probability of word w given topic z
  - The distribution of the timestamps for each topic

Related Work

- Periodicity Analysis in time-series database
- Topic models
  - PLSA and LDA
- Variants of topic models
  - Topic Over Time, etc.

General Idea of Periodic Topic Mining

- Term cooccurrence
  - If two words co-occur often in the same documents, they are more likely to belong to the same topic
- Temporal Structure
  - We assume that there are many consecutive periods across the time line. The words occurring around the same time in each period are likely to be clustered

Temporal Patterns of Topics

- Periodic Topics
  - A periodic topic is one repeating in regular intervals
- Background Topics
  - A background topic is one covered uniformly over the entire period
- Bursty Topics
  - A bursty topic is a transient topic that is intensively covered only in a certain time period

Temporal Patterns of Topics (Cont.)

- Periodic Topics
  - The distribution of timestamps for each periodic topic as a mixture of Gaussian distributions where the interval between the consecutive components is T
- Background Topics
  - The timestamps of the background topics are generated by a uniform distribution
- Bursty Topics
  - The timestamps of the bursty topics are generated from a Gaussian distribution
- The document collection is modeled as a mixture of background topics, bursty topics and periodic topics

Generative Process

- To generate each word in document d from collection D:
- (1) Sample a topic z from multinomial ϕ_d i.e., $\{p(z|d)\}_{z \in Z}$
  - (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 time 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 i.e., $\{p(w|z)\}_{w \in V}$

Log-likelihood of Document Collection

- Given the data collection $\{(w_d, t_d)\}_{d \in D}$ where 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

$$L(\psi; D) = \log p(D \mid \psi) = \log \prod_{d \in D} p(w_d, t_d \mid \psi)$$

$$\log p(w_d, t_d \mid \psi) = \sum_d \sum_w n(d, w) \log \sum_z p(t_d \mid z) p(w \mid z) p(z \mid d)$$

- If topic z is a background topic, $p(t \mid z) = \frac{1}{t_{end} - t_{start}}$
- If topic z is a bursty topic, $p(t \mid z) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z)^2}{\sigma_z^2}}$
- If topic z is a periodic topic, $p(t \mid z) = p(k) \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z - kT)^2}{\sigma_z^2}}$

Parameter Estimation

- EM (Expectation Maximization) algorithm
- E-step

$$p(z|d,w) = \frac{p(t_d|z)p(w|z)p(z|d)}{\sum_{z'} p(t_d|z')p(w|z')p(z'|d)}$$

- M-step

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

For bursty topic z

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

For periodic topic z

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

Complexity: O(iter K|W|) where iter is the number of the iterations in EM, K is the number of topics, |W| is the total count of the words in all the documents

Datasets

- Seminar
  - The weekly seminar announcements for one semester from six research groups in computer science department at University of Illinois at Urbana-Champaign
  - 61 documents and 901 unique words
  - Set periodic interval T as 1 week
- DBLP (Digital Bibliography Project)
  - 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
  - 4070 documents and 2132 unique words
  - Set periodic interval T as 1 year
- Flickr
  - The photos for several music festivals from 2006 to 2010 including SXSW (South by Southwest), Coachella, Bonnaroo, Lollapalooza and ACL (Austin City Limits)
  - The tags of a photo are considered as document text, while the time when the photo was taken is considered as document timestamp
  - 84244 documents and 7524 unique words
  - Set periodic interval T as 1 year

Topics Discovered by LPTA

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

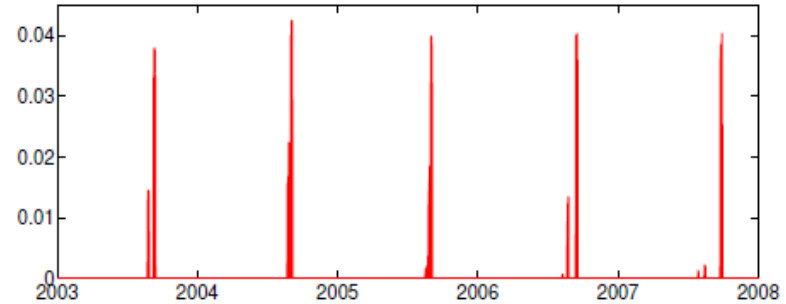
Seminar		DBLP		Flickr	
Topic 1 (DAIS) Tue 16:00 (0h0m0s)	Topic 2 (AIIS) Fri 14:00 (0h0m0s)	Topic 1 (KDD) Aug 23 (10d3h11m)	Topic 2(SIGIR) Aug 3 (9d6h56m)	Topic 1 (ACL) Sep 29 (10d13h20m)	Topic 2 (Bonnaroo) Jun 16 (2d14h21m)
model 0.0166	computer 0.0168	mining 0.0353	retrieval 0.0495	acl 0.0945	bonnaroo 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

LPTA vs. Periodicity Detection

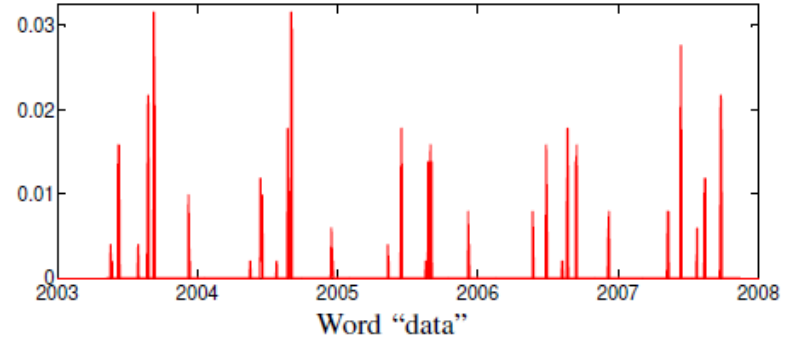
- AUTOPERIOD*, a two-tier approach by considering the information in both the autocorrelation and the periodogram, fails to detect meaningful periodic words because the time series are sparse and few words have apparent periodic patterns.
- Compared with single word representation, LPTA uses multiple words to describe a topic.
  - In DBLP, topic "VLDB": data 0.0530, xml 0.0208, query 0.0196, queries 0.0176, efficient 0.0151, mining 0.0142, database 0.0136, streams 0.0112, databases 0.0111

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

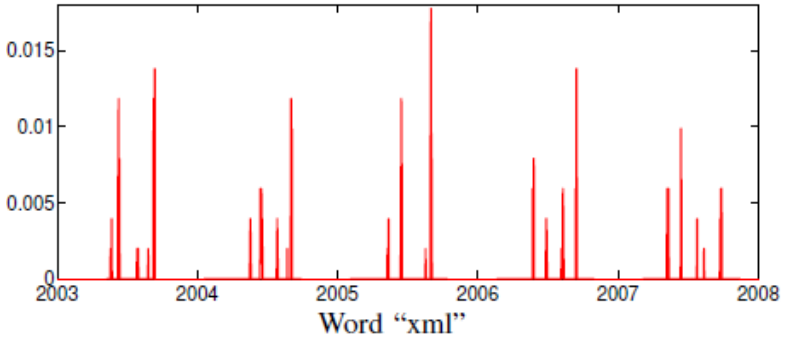
LPTA vs. Periodicity Detection (Cont.)



Topic *VLDB* discovered by LPTA



Word "data"



Word "xml"

Time distribution of topic VLDB discovered by LPTA and time distributions of the words in the topic

LPTA vs. Topic Models

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.

Integration of Text and Time Information

Periodic topics for SIGMOD vs. VLDB and SIGMOD vs. CVPR datasets by using LPTA. The date and the duration are the mean and standard deviation of the timestamps.

SIGMOD vs. VLDB		SIGMOD vs. CVPR	
Topic 1 (SIGMOD) Jun 17 (7d11h6m)	Topic 2 (VLDB) Sep 11 (9d5h29m)	Topic 1 (SIGMOD) Jun 20 (7d15h42m)	Topic 2 (CVPR) Jun 21 (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	system	segmentation
web	system	databases	detection
system	processing	queries	motion

SIGMOD and VLDB are two reputed conferences in database area, and it is difficult to differentiate these two conferences based on text only

SIGMOD and CVPR are held in June, so it is difficult to differentiate these two if we rely on time information only

Periodic vs. Bursty Topics

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.

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

The words will fit into the corresponding periodic or bursty topics if they have periodic or bursty patterns

Quantitative Evaluation

The latent topics discovered by the topic modeling approaches can be regarded as clusters. Accuracy and normalized mutual information (NMI) can be used to measure the clustering performance.

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**

Conclusion

- Introduce the problem of latent periodic topic analysis on timestamped documents
- Propose a model called LPTA (Latent Periodic Topic Analysis) that exploits both the periodicity of the terms and term co-occurrences.
- Demonstrate that our LPTA model works well for discovering the latent periodic topics by combining the information from topical clusters and periodic patterns

Future Work

- Effectively analyzing large scale data
- Automatically determining the optimal number of topics in real life
- Incorporating the social networks into periodicity detection

Thanks!