Mining Spatiotemporal and Social Media Data

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

JULY 20, 2015
Outline

- Introduction: Integrated Mining of Spatio, Temporal and Text Data
- Mining Spatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary
Introduction: Integrated Mining of Spatial, Temporal and Text Data

- Spatial Data
  - Spatial Data Mining
- Temporal Data
  - Trajectory Mining
  - Time-Series Analysis
- Text
  - Text-Rich Information Network Analysis
  - Latent Geographical Topic Analysis
  - Latent Periodic Topic Analysis

Link
Outline

- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary
Spatial Patterns and Associations

- Spatial frequent patterns and association rule: $A \Rightarrow B [s\%, c\%]$
- $A$ and $B$ are sets of spatial or non-spatial predicates, e.g.,
  - Topological relations: *intersects*, *overlaps*, *disjoint*, etc.
  - Spatial orientations: *left_of*, *west_of*, *under*, etc.
  - Distance information: *close_to*, *within_distance*, etc.
- Measures: $s\%$: support, and $c\%$: confidence of the rule
- Example: Rules likely to be found
  - $is\_a(x, large\_town) \land intersect(x, highway) \Rightarrow adjacent\_to(x, water) [7\%, 85\%]$
- Explore *spatial autocorrelation*: Spatial data tends to be highly self-correlated (nearby things are more related than distant ones)
- E.g., neighborhood, temperature
Mining Spatial Associations: Progressive Refinement

- Hierarchy of spatial relationship:
  - close_to is a generation of near_by, touch, intersect, contain, ...
  - Progressive refinement: First search for rough relationship and then refine it

- Two-step mining of spatial association:
  - Step 1: Rough spatial computation (as a filter)
    - Using MBR (Minimum Bounding Rectangle) or R-tree for rough estimation
  - Step 2: Detailed spatial algorithm (as refinement)
    - Apply only to those objects which have passed the rough spatial association test (no less than min_support)
Outline

- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary
Partition-Based Trajectory Pattern Mining

- Partition-Based Trajectory Pattern Mining (e.g., Mining T-Patterns) [1]:
  - First partition the space into equal-width grids and obtain Regions-of-Interests (RoIs)
  - Then transform each input trajectory into a time-annotated symbolic sequence
  - Use constraint-based sequential pattern mining to find trajectory patterns

Detecting Moving Object Clusters

- **Flock and convoy**: Both require $k$ consecutive time stamps
  - **Flock**: At least $m$ entities are within a *circular* region of radius $r$ and move in the same direction
  - **Convoy**: *Density-based clustering* at each timestamp; no need to be a rigid circle
- **Swarm**: Moving objects may not be close to each other for all the consecutive time stamps
- Efficient pattern mining algorithms for uncovering such swarm patterns
Trajectory Clustering: A Partition-and-Group Framework

- Grouping trajectories **as a whole** ⇒ cannot find **similar portions** of trajectories
- **Solution:** discovers common **sub**-trajectories, e.g., *forecast hurricane landfall*
- Two phases: **partitioning** and **grouping**
- Identify the points where the behavior of a trajectory changes rapidly ⇒ **characteristic points**
  - Based on the minimum description length (MDL) principle

\[
\begin{align*}
p_1 & \quad p_2 \quad p_3 \quad p_4 \quad p_5 \quad p_6 \quad p_7 \quad p_8 \\
p_{c_1} & \quad p_{c_2} & \quad p_{c_3} & \quad p_{c_4} \\
TR_1 & \quad TR_2 & \quad TR_3 & \quad TR_4 & \quad TR_5
\end{align*}
\]

⇒ characteristic point – – – : trajectory partition

Outline

- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary
Mining Frequent Movement Patterns

- **Frequent Movement Pattern:** A movement sequence that frequently appears in the input trajectory database

- **Frequent Movement Pattern vs. Frequent Sequential Pattern**
  - Both aim at finding frequent subsequences from the input sequence database
  - For mining frequent movement patterns, similar places may need to be grouped to collectively form frequent subsequences

An example trajectory

An example movement pattern
Mining Semantic-Rich Movement Patterns

- **Semantics-rich Movement Pattern**: In addition to knowing how people move from one region to another, we also want to understand the functions of the regions.

- **A two-step top-down mining approach**:
  - **Step 1**: Find a set of coarse patterns that reflect people’s semantics-level transitions (e.g., office → restaurant, home → gym)
  - **Step 2**: Split each coarse pattern into several fine-grained ones by grouping similar movement snippets

Outline

- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media
- From Mining Social Relationships
- Summary
Pattern Discovery in Sparse Movement Data: Finding Good Reference Points

- Pattern discovery in sparse data:
  - Periodicity shows up in some reference “spots” (or “cluster centers”)
  - Reference spots can be detected using **density-based method**
  - Periods are detected for each reference spot using **Fourier Transform** and **auto-correlation**

Finding being flying patterns? Bee hive is a good reference point

Period is more obvious in this binary sequence!
Pattern Discovery in Sparse Movement Data: Finding Good Reference Points

- Pattern discovery in sparse data:
  - Periodicity shows up in some reference “spots” (or “cluster centers”)
  - Reference spots can be detected using density-based method
  - Periods are detected for each reference spot using Fourier Transform and auto-correlation

Finding being flying patterns? Bee hive is a good reference point

Period is more obvious in this binary sequence!
Example: Mining Periodic Patterns with Sparse Data

- **Detecting periods**: Cluster data to find reference “points” and then detect multiple interleaved periods by Fourier Transform and auto-correlation.

- **Summarizing periodic patterns**: By clustering and pattern discovery.

3-yr Bird migration data: very sparse

- Jan.-Mar.
- Apr.-June
- July-Oct.
- Nov.
- Dec.
Periodicity Detection in Sparse Data

- Real movement data can be sparse, e.g., geo-location at phone calls
  - Segment the data using length 20
  - Overlay the segments
    - Observations are clustered in [5,10] interval.

- Projecting on the true period, it shows a highly skewed (clustered) distribution
  - Segment the data using length 16
  - Overlay the segments
    - Observations are scattered.

- Effective method can be developed based on this observation (Li, et al., 2015)
Outline

- Integrated Mining of Spatial, Temporal and Social Media Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media
- From Mining Social Relationships
- Summary
Social Media Are Popular in Today’s World

- Social media contains very rich spatial, temporal, text, photo, link, social network information
- Examples
  - Twitter: tweets from smart phones
  - Geo-tagged tweets
  - Flickr: geo-tagged photos
  - Advanced cameras with GPS receivers
  - Applications including Google Earth, Flickr, etc.
  - GPS functions in smart phones
  - Social media data mining: A rich frontier
LGTA: Mining Spatial Text Documents

- Applications
  - Analyze the cultural differences around the world
  - Explore the hot topics or events in different places
  - Compare the popularity of specific products in different regions
  - Discover different topics of interests those are coherent in geographical regions
  - Compare several topics across different geographical locations

- Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang, “Geographical Topic Discovery and Comparison”, Proc. of 2011 Int. World Wide Web Conf. (WWW'11), Hyderabad, India, Mar. 2011.
GeoTopic: From Geo-Tagged Text to Topic Clusters

- Input: Text with spatial information

<table>
<thead>
<tr>
<th>ID</th>
<th>Image</th>
<th>Text</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>dimsum breakfast dumplings...</td>
<td>22.377</td>
<td>114.185</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>sushi sashimi rawfish...</td>
<td>35.669</td>
<td>139.762</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>taco tacogrill crispybeef...</td>
<td>30.265</td>
<td>-97.680</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Output:
- Geographic topics: \{ p(w|z) \}
- \( p(w|z_1), p(w|z_2), p(w|z_3) \)
- Topic distribution \( p(z|l) \)

Location \( l = (40.70, 73.91) \)

| topic z                  | \( p(z|l) \) |
|--------------------------|--------------|
| Topic 1 (Chinese food)   | 22%          |
| Topic 2 (Japanese food)  | 14%          |
| Topic 3 (Mexican food)   | 18%          |
| ...                      | ...          |
Potential Solutions: Previous Work

- LDM: Location-driven model
  - Clustering based on document locations
  - One location cluster is a topic

- TDM: Text-driven model [Mei et al. WWW’08]
  - Topic modeling with network regularization
  - Documents that are close in space should have similar topic distributions

- GeoFolk [Sizov WSDM’10]
  - A topic modeling that uses both text and spatial information
  - The geographical distribution of each topic is Gaussian
LGTA: General Ideas (Location-Text Join Model)

- Geographical topic discovery
  - Topics are generated from regions instead of documents:
    - The geographic distribution of each region follows a Gaussian distribution
    - The words that are close in space likely belong to the same region and thus should be clustered into the same geographical topic
  - To generate a geographical document \( d \) in a collection \( D \):
    - Sample a region \( r \) from the discrete distribution of region importance \( \alpha \):
      - \( r \sim \text{Discrete} (\alpha) \)
    - Sample location \( l_d \) from Gaussian distribution of \( \mu_r \) and \( \Sigma_r \)
    - To generate each word in document \( d \):
      - (a) sample a topic \( z \) from multinomial \( \phi_r \)
      - (b) sample a word \( w \) from multinominal \( \theta_z \)
  - Each topic can be related to several regions
  - Parameters can be estimated using an EM algorithm
Performance Comparison: Geo-Tagged Photos Related to Landscape (coast vs. desert vs. mountain)
LGTA: Geographical Food Topic Comparison

The larger $p(\text{topic}|\text{location})$ is, the darker the location is

<table>
<thead>
<tr>
<th>Chinese Food</th>
<th>Japanese Food</th>
<th>Italian Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>chinese 0.552</td>
<td>japanese 0.519</td>
<td>italian 0.848</td>
</tr>
<tr>
<td>noodles 0.067</td>
<td>ramen 0.104</td>
<td>cappuccino 0.067</td>
</tr>
<tr>
<td>dimsum 0.064</td>
<td>soba 0.066</td>
<td>latte 0.048</td>
</tr>
<tr>
<td>hotpot 0.039</td>
<td>noodle 0.065</td>
<td>gelato 0.030</td>
</tr>
<tr>
<td>rice 0.038</td>
<td>sashimi 0.039</td>
<td>pizza 0.002</td>
</tr>
<tr>
<td>noodle 0.035</td>
<td>yakitori 0.030</td>
<td>pizzeria 0.002</td>
</tr>
<tr>
<td>tofu 0.020</td>
<td>okonomiyaki 0.026</td>
<td>mozzarella 0.001</td>
</tr>
<tr>
<td>dumpling 0.018</td>
<td>udon 0.026</td>
<td>pasta 0.001</td>
</tr>
<tr>
<td>duck 0.018</td>
<td>tempura 0.020</td>
<td>ravioli 0.000</td>
</tr>
<tr>
<td>prawn 0.017</td>
<td>curry 0.016</td>
<td>pesto 0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>French Food</th>
<th>Spanish Food</th>
<th>Mexican Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>french 0.564</td>
<td>spanish 0.488</td>
<td>mexican 0.484</td>
</tr>
<tr>
<td>bistro 0.070</td>
<td>tapas 0.269</td>
<td>tacos 0.069</td>
</tr>
<tr>
<td>patisserie 0.056</td>
<td>paella 0.076</td>
<td>taco 0.059</td>
</tr>
<tr>
<td>bakery 0.049</td>
<td>pescado 0.059</td>
<td>salsa 0.036</td>
</tr>
<tr>
<td>resto 0.044</td>
<td>olives 0.032</td>
<td>cajun 0.031</td>
</tr>
<tr>
<td>pastry 0.033</td>
<td>sticky rice 0.017</td>
<td>burrito 0.027</td>
</tr>
<tr>
<td>tarte 0.026</td>
<td>tortilla 0.013</td>
<td>crawfish 0.023</td>
</tr>
<tr>
<td>croissant 0.021</td>
<td>mediterranean 0.010</td>
<td>guacamole 0.022</td>
</tr>
<tr>
<td>baguette 0.019</td>
<td>mussels 0.008</td>
<td>margarita 0.020</td>
</tr>
<tr>
<td>mediterranean 0.018</td>
<td>octopus 0.008</td>
<td>cocktails 0.020</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media Data
- Latent Periodic Topic Discovery
- Real-Time Local Event Detection from Geo-Tagged Social Media
- Summary
Latent Periodic Topic Analysis [ICDM’11]

- Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "LPTA: A Probabilistic Model for Latent Periodic Topic Analysis", ICDM'11

- Periodic phenomena exist ubiquitously
  - Hurricanes
  - Music and film festivals
  - Product sales
  - TV program
  - Publicly traded company

- Most text articles have time associated with

- Ex. 1. News articles associated with pub. dates

- Ex. 2. Tagged photos annotated with dates in Flickr
Apply Periodicity Analysis on Text Data

- Periodicity detection for time series database [Elfeky et al. TKDE 2005]
- Some studies follow the similar strategies to analyze the time distribution of a single tag or query to detect periodic patterns [Vlachos et al. SIGMOD 2004]

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 is insufficient to discover periodic topics
  - E.g., “music”, “festival” and “chicago” may not have periodic patterns if considered separately, but there may be periodic topics if considered together
- Synonyms and polysemy words due to the language diversity
Latent Periodic Topic Analysis (LPTA)

Input:
Time-stamped documents

<table>
<thead>
<tr>
<th>ID</th>
<th>Text</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>coachella, music, arts, festival, ...</td>
<td>Apr 27 2008</td>
</tr>
<tr>
<td>2</td>
<td>sxsw, south by southwest, austin, ...</td>
<td>Mar 14 2008</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Output:
1. Periodic topics: \{ p(w|z) \}
2. Time distribution of topics

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

The distribution of the timestamps for the topic related to Coachella festival
Latent Periodic Topic Analysis: Problem Formulation

Input:
- A collection of time-stamped 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
General Idea of LPTA

- Related work
  - Periodicity Analysis in time-series DB [Elfeky et al., 2005]
  - Topic models: PLSA [Hofmann SIGIR 1999] and LDA [Blei et al. JMLR 2003]
  - Topic Over Time [Wang et al. KDD 2006], etc.

- LPTA (Latent Periodic Topic Analysis): General Ideas
  - Term co-occurrence
    - 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
  - 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**
  - A background topic is one covered uniformly over the entire period
  - The timestamps of the background topics are generated by a uniform distribution

- **Bursty topics**
  - A bursty topic is a transient topic that is intensively covered only in a certain time period
  - 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 of LPTA

- For each word in document \(d\) from collection \(D\):
  - Sample a topic \(z\) from multinomial \(\phi_d\) i.e., \(\{p(z \mid d)\}_{z \in Z}\)
  - (a) If \(z\) is a background topic, sample time \(t\) from a uniform distribution \([t_{\text{start}}, t_{\text{end}}]\), where \(t_{\text{start}}\) and \(t_{\text{end}}\) 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
  - Sample a word \(w\) from multinomial \(\theta_z\) i.e., \(\{p(w \mid Z)\}_{w \in V}\)
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

\[
L(\psi; D) = \log p(D | \psi) = \log \prod_{d \in D} p(w_d, t_d | \psi)
\]

\[
\log p(w_d, t_d | \psi) = \sum_d \sum_w n(d, w) \log \sum_z p(t_d | z) p(w | z) p(z | d)
\]

- If topic \(z\) is a background topic,
  \[
p(t | z) = \frac{1}{t_{end} - t_{start}}
\]
- If topic \(z\) is a bursty topic,
  \[
p(t | z) = \frac{1}{\sqrt{2\pi\sigma_z^2}} e^{-\frac{(t - \mu_z)^2}{2\sigma_z^2}}
\]
- If topic \(z\) is a periodic topic,
  \[
p(t | z) = p(k) \frac{1}{\sqrt{2\pi\sigma_z^2}} e^{-\frac{(t - \mu_z - kT)^2}{2\sigma_z^2}}
\]
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**
    \[
    p(w|z) = \frac{\sum_d n(d, w)p(z|d, w)}{\sum_d \sum_{w'} n(d, w')p(z|d, w')} \quad p(z|d) = \frac{\sum_w n(d, w)p(z|d, w)}{\sum_d \sum_{z'} n(d, w)p(z'|d, w)}
    \]
  - For bursty topic \(z\)
    \[
    \mu_z = \frac{\sum_d \sum_w n(d, w)p(z|d, w)t_d}{\sum_d \sum_w n(d, w)p(z|d, w)}
    \]
    \[
    \sigma_z = \left( \frac{\sum_d \sum_w n(d, w)p(z|d, w)(t_d - \mu_z)^2}{\sum_d \sum_w n(d, w)p(z|d, w)} \right)^{1/2}
    \]
  - For periodic topic \(z\)
    \[
    \mu_z = \frac{\sum_d \sum_w n(d, w)p(z|d, w)(t_d - I_d T)}{\sum_d \sum_w n(d, w)p(z|d, w)}
    \]
    \[
    \sigma_z = \left( \frac{\sum_d \sum_w n(d, w)p(z|d, w)(t_d - \mu_z - I_d T)^2}{\sum_d \sum_w n(d, w)p(z|d, w)} \right)^{1/2}
    \]
  - Complexity: \(O(\text{iter} \ K|W|)\) where \(\text{iter}\) 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
Experimental Datasets

- Seminar
  - Weekly seminar announcements for one semester from six research groups @CS, UIUC
  - 61 documents and 901 unique words
  - Set periodic interval $T$ as 1 week

- DBLP (Computer Science Digital Bibliography)
  - The paper titles of several confs (WWW, SIGMOD, SIGIR, KDD, VLDB and NIPS) from 2003 to 2007
  - The timestamps of the documents are determined w.r.t. 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

<table>
<thead>
<tr>
<th>Seminar</th>
<th>DBLP</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1 (DAIS) Tue 16:00 (0h0m0s)</td>
<td>Topic 1 (KDD) Aug 23 (10d3h11m)</td>
<td>Topic 1 (ACL) Sep 29 (10d13h20m)</td>
</tr>
<tr>
<td>Topic 2 (AIIS) Fri 14:00 (0h0m0s)</td>
<td>Topic 2(SIGIR) Aug 3 (9d6h56m)</td>
<td>Topic 2 (Bonnaroo) Jun 16 (2d14h21m)</td>
</tr>
<tr>
<td>model 0.0166</td>
<td>computer 0.0168</td>
<td>mining 0.0353</td>
</tr>
<tr>
<td>based 0.0158</td>
<td>learning 0.0158</td>
<td>data 0.0289</td>
</tr>
<tr>
<td>mining 0.0151</td>
<td>machine 0.0138</td>
<td>search 0.0233</td>
</tr>
<tr>
<td>text 0.0143</td>
<td>science 0.0128</td>
<td>clustering 0.0208</td>
</tr>
<tr>
<td>network 0.0135</td>
<td>algorithms 0.0128</td>
<td>based 0.0195</td>
</tr>
<tr>
<td>web 0.0119</td>
<td>language 0.0118</td>
<td>web 0.0168</td>
</tr>
<tr>
<td>problem 0.0111</td>
<td>work 0.0108</td>
<td>learning 0.0159</td>
</tr>
<tr>
<td>data 0.0111</td>
<td>problems 0.0108</td>
<td>networks 0.0114</td>
</tr>
<tr>
<td>query 0.0111</td>
<td>models 0.0108</td>
<td>analysis 0.0105</td>
</tr>
<tr>
<td>latent 0.0095</td>
<td>prediction 0.0108</td>
<td>large 0.0104</td>
</tr>
</tbody>
</table>
LPTA vs. Periodicity Detection

- AUTOPERIOD [Vlachos et al. SDM 2005], 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

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

<table>
<thead>
<tr>
<th>Seminar</th>
<th>LDA</th>
<th>PLSA</th>
<th>DBLP</th>
<th>LDA</th>
<th>Flickr</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>Topic 2</td>
<td>Topic 1</td>
<td>Topic 2</td>
<td>Topic 1</td>
<td>Topic 2</td>
<td>Topic 1</td>
</tr>
<tr>
<td>data</td>
<td>memory</td>
<td>problem</td>
<td>systems</td>
<td>web</td>
<td>search</td>
<td>web</td>
</tr>
<tr>
<td>latent</td>
<td>computer</td>
<td>algorithm</td>
<td>computer</td>
<td>data</td>
<td>text</td>
<td>mining</td>
</tr>
<tr>
<td>visualizati.</td>
<td>data</td>
<td>network</td>
<td>science</td>
<td>xml</td>
<td>databases</td>
<td>semantic</td>
</tr>
<tr>
<td>intel</td>
<td>mining</td>
<td>graph</td>
<td>algorithms</td>
<td>queries</td>
<td>relational</td>
<td>detection</td>
</tr>
<tr>
<td>talk</td>
<td>parallel</td>
<td>time</td>
<td>time</td>
<td>mining</td>
<td>user</td>
<td>automatic</td>
</tr>
<tr>
<td>analysis</td>
<td>science</td>
<td>networks</td>
<td>agent</td>
<td>semantic</td>
<td>analysis</td>
<td>services</td>
</tr>
<tr>
<td>computer</td>
<td>pattern</td>
<td>influence</td>
<td>visualizati.</td>
<td>search</td>
<td>ranking</td>
<td>applicatic.</td>
</tr>
<tr>
<td>systems</td>
<td>programm.</td>
<td>online</td>
<td>data</td>
<td>streams</td>
<td>structure</td>
<td>graph</td>
</tr>
<tr>
<td>machine</td>
<td>hardware</td>
<td>work</td>
<td>engineering</td>
<td>managem.</td>
<td>support</td>
<td>extraction</td>
</tr>
<tr>
<td>visual</td>
<td>algorithms</td>
<td>question</td>
<td>function</td>
<td>adaptive</td>
<td>evaluation</td>
<td>patterns</td>
</tr>
</tbody>
</table>
Integration of Text and Time

- 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 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.

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

Conclusion: The LPTA model discovers the latent periodic topics by combining the information from topical clusters and periodic patterns.
Outline

- Introduction
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media Data
- Latent Periodic Topic Discovery
- Real-Time Local Event Detection from Geo-Tagged Social Media
- Summary
GeoBurst: Real-time Local Event Detection in Geo-Tagged Tweet Streams [SIGIR’16]

- C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, J. Han, “GeoBurst: Real-time Local Event Detection in Geo-Tagged Tweet Streams”, SIGIR’16

- Local Event: A local events is an unusual activity bursted within a local area and specific duration while engaging a considerable number of participants

  - E.g., parade, riot, sport game, concert, accident, disaster

- The geo-tagged tweet stream brings new opportunities to this problem because of its (1) sheer size; (2) multi-dimensional information; and (3) real-time nature
Research Challenges

- Major challenges
  - Integrating multiple types of data: Location, time and text
  - Extracting interpretable events from tremendous noises (tweets are noisy and short)
  - On-line and real-time detection

- Previous work
  - Most existing event detection methods are designed for detecting *global events*
    - Bursty in the entire stream; but local events are “bursty” in a small region and involve a relatively small number of tweets
  - Some local event detection methods
    - Not model the correlations between keywords; or are incapable of detecting local events in real time
Insight and Problem Definition

- A local event usually leads to many related tweets around the location (a geo-topic cluster).
- But a geo-topic cluster is not necessarily a local event:
  - It may be a routine activity in that region (e.g., shopping).
  - It may be a global event rather than a local one (e.g., TV show).

We define a local event as a geo-topic cluster that shows clear spatiotemporal burstiness.

- Our task: Given the geo-tagged tweet stream, we aim to:
  - detect all local events in any query time window (batch mode).
  - update the result list in real time as the query window shifts continuously (online mode).
Overview of GeoBurst

- GeoBurst, a reference-based method for local event detection

- It consists of three key components:
  - A candidate generator that finds geo-topic clusters in the query time frame, and regard them as candidate events
  - A ranking module that summarizes the routine activities in different regions to filter non-event candidates
  - An updater that updates local events in real time as the query window shifts
Candidate Event Generation (I) Find Geo-Topic Clusters

- Find geo-topic clusters in the query time frame as candidate events
  - Geo-topic cluster (a group of tweets): geographically close & semantically relevant
- Intuition: the spot where the event occurs is acting as a pivot that produces relevant tweets around it
- Our clustering algorithm is based on:
  - a geo-topic authority score for each tweet
  - an authority ascent process to find authority maxima as pivots
- Computing geo-topic authority
  - Geographical impact: calculated by a kernel function ($d$ and $d'$ are tweets)
    \[ G(d' \rightarrow d) = K(||l_d - l_{d'}||/h) \]
  - Semantic impact: calculated by random walk on a keyword co-occurrence graph
    \[ S(d' \rightarrow d) = \frac{1}{mn} \sum_{e \in E_d} \sum_{e' \in E_{d'}} r_{e' \rightarrow e} \]
Candidate Event Generation (II) Pivot & Authority Ascent

- Geo-topic authority: A tweet gets an authority score from neighbor tweets where
  - The geographical impact is captured by kernel function
  - The semantic impact is captured by random walk on the keyword co-occurrence graph
- A pivot is an authority maximum: a prominent tweet that is surrounded by many relevant tweets
- A pivot attracts similar tweets to form geo-topic clusters
- Find all the pivots in the geo-topic space by Authority Ascent
The Ranking Module

- We design the **activity timeline structure** to summarize the activities in different spatial regions and time periods.
- The summaries in the activity timeline serve as background knowledge to quantify the spatiotemporal burstiness of candidates.

Each snapshot is a set of micro-clusters
Each cluster is an activity summary for a region

Retrieve the snapshots in a reference window as background knowledge
Compute z-score for each candidate as its ranking score
The Update Module

- In the entire process of GeoBurst, the most time-consuming step is pivot finding.
- How to avoid finding pivots from scratch as the query window shifts?
  - The key is to maintain the local pivot for each tweet.
- We design an updating strategy based on the additive property of authority score:
  - subtracting the contributions of outdated tweets
  - emphasizing the contributions of new tweets
Experiments (Algorithm Comparison)

- **Data:**
  - NYC: 9M geo-tagged tweets in New York during 3 months
  - LA: 8M geo-tagged tweets in Los Angeles during 3 months
- **Task:** 80 queries with different durations (3h, 4h, 5h, 6h), find top-5 local events in each query window
- **Compared Method:** EvenTweet (PVLDB’13), Wavelet (CIKM’09)
- **Evaluation:** The crowdsourcing platform CrowdFlower

- Ask the workers to judge whether the result is a local event

**Precision Comparison**

(a) Precision comparison (NY).

(b) Precision comparison (LA).
## Experiments (Illustrative Cases)

<table>
<thead>
<tr>
<th>GeoBurst</th>
<th>Is event?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Festival of Light! #nyfol (@ The Archway under the Manhattan Bridge in Brooklyn, NY)</td>
<td>Yes</td>
</tr>
<tr>
<td>2. #Lasers and beats under the Manhattan Bridge! #NewYorkFestivalofLight #NYFOL @ DUMBO</td>
<td></td>
</tr>
<tr>
<td>3. New York Festival of Lights #nyfol #dumbo @ DUMBO, Brooklyn</td>
<td></td>
</tr>
</tbody>
</table>

| # 2 | 1. Knicks vs. Nets at Barclays Center. @ Barclays Center http://t.co/PILk1xK3Tn          | Yes |
|     | 2. Brooklyn go hard @ Barclays Center http://t.co/iVUsJ5TNG                                      |     |
|     | 3. Let’s go Knicks! #NETS1107 (@ Barclays Center - @brooklynnets for @nyknicks vs @BrooklynNets) |     |

| # 3 | 1. #Thai Restaurant #spicythaifood (@ 104 2nd Avenue in New York, NY)                         | No  |
|     | 2. The ASIAN DISHES here are always my favorite. @ Ugly Kitchen                               |     |
|     | 3. Dinner time with my family. Suuuuper Nice Indian RESTAURANT! @ Malai Marke Indian Cuisine. |     |

<table>
<thead>
<tr>
<th>EvenTweet</th>
<th>Is event?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I practiced... Almost time for Amy Schumer. Jennifer (@ Carnegie Hall) <a href="https://t.co/HfqFTLmK2y">https://t.co/HfqFTLmK2y</a></td>
<td>No</td>
</tr>
<tr>
<td>2. 2014 Gold Glove Awards Ceremony with Hall of Famers, All-Stars Jay Leno @ The Plaza Hotel</td>
<td></td>
</tr>
<tr>
<td>3. My best attempt at a selfie with Hugh Jackman after The River at CITS @ The River on Broadway</td>
<td></td>
</tr>
</tbody>
</table>

| # 2 | 1. Knicks vs. Nets at Barclays Center. @ Barclays Center http://t.co/PILk1xK3Tn          | Yes |
|     | 2. Budweiser brings everyone together #family #novenfriends @ Alchemy Tavern, Brooklyn |     |
|     | 3. #Knicks vs #nets with my best gal. @ Barclays Center Brooklyn http://t.co/eXXMUKxplS |     |

| # 3 | 1. #katespade @ Kate Spade / Jack Spade HQ http://t.co/g6j1Fwyc4M | No  |
|     | 2. Inspiring keynote by Twitter CEO, Dick Costolo @GirlsWhoCode Gala. http://t.co/yEGh803CuT |     |
|     | 3. I wonder if Jake from Statefarm covers Jumanji? |       |
Outline

- Introduction
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media Data
- Latent Periodic Topic Discovery
- Real-Time Local Event Detection from Geo-Tagged Social Media
- Summary
Summary

- Emerging: Integrated mining spatiotemporal and social media data
  - Mining Geospatial Patterns
  - Mining and Aggregating Patterns over Multiple Trajectories
  - Mining Semantic-Rich Movement Patterns
  - Mining Periodic Movement Patterns
  - GeoTopic Discovery in Social Media Data
  - Latent Periodic Topic Discovery
  - Real-Time Local Event Detection from Geo-Tagged Social Media
- Integrated data mining with spatiotemporal, social and trajectory data
  - Integrated mining with four dimensions: Spatial + Temporal + Text + Network
References (I)

- F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory Pattern Mining, KDD’07
- Y. Huang, S. Shekhar, H. Xiong, Discovering colocation patterns from spatial data sets: A general approach, IEEE Trans. on Knowledge & Data Eng., 16(12), 2004
- K. Koperski, J. Han, “Discovery of Spatial Association Rules in Geographic Information Databases”, SSD’95
- J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory Clustering: A Partition-and-Group Framework", SIGMOD’07
- Z. Li, B. Ding, J. Han, R. Kays, “Swarm: Mining Relaxed Temporal Moving Object Clusters”, VLDB’10
- Z. Li, B. Ding, J. Han, R. Kays, P. Nye, “Mining Periodic Behaviors for Moving Objects”, KDD’10
- Z. Li, J. Wang and J. Han, "ePeriodicity: Mining Event Periodicity from Incomplete Observations", IEEE TKDE, 27(5): 1219-1232, 2015
- Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, “Geographical Topic Discovery and Comparison”, WWW'11
- Z. Yin, L. Cao, J. Han, J. Luo, and T. S. Huang, "Diversified Trajectory Pattern Ranking in Geo-tagged Social Media", SDM’11
- C. Zhang, J. Han, L. Shou, J. Lu, and T. La Porta, "Splitter: Mining Fine Grained Sequential Patterns in Semantic Trajectories", VLDB’14
- C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, J. Han, “GeoBurst: Real-time Local Event Detection in Geo-Tagged Tweet Streams”, SIGIR’16
Diversified Trajectory Pattern Ranking

Given a collection of geo-tagged photos along with users, locations and timestamps, how to rank the mined trajectory patterns with diversification into consideration?

- **Our Framework**
  - Extract trajectory patterns from the photo collection
  - Rank the trajectory patterns by estimating their importance according to user, location and trajectory pattern relations
  - Diversify the ranking result to identify the representative trajectory patterns from all the candidates

**Input:** A collection of geo-tagged photos (user, date time, GPS location)

1. Extract trajectory patterns
2. Rank trajectory patterns
3. Diversify ranked patterns

**Output:** Diversified trajectory pattern ranking result

Ex.: Top ranked trajectories in London, New York and Paris
Data Preprocessing and Pattern Discovery

- Cluster locations: mean-shift algorithm (27974 photos in London)

- Form sequences

- PrefixSpan [Pei et al. TKDE 2004]
  - Ex. (min-support = 2)

- Three frequent sequential patterns:
  - londoneye → bigben
  - londoneye → bigben → trafalgarsquare
  - londoneye → tatemodern

<table>
<thead>
<tr>
<th>ID</th>
<th>User</th>
<th>Date</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alice</td>
<td>04/26/11</td>
<td>londoneye -&gt; bigben -&gt; downingstreet -&gt; trafalgarsquare</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>04/27/11</td>
<td>londoneye -&gt; tatemodern -&gt; towerbridge</td>
</tr>
<tr>
<td>3</td>
<td>Bob</td>
<td>04/26/11</td>
<td>londoneye -&gt; bigben -&gt; tatemodern</td>
</tr>
</tbody>
</table>

ID User Date Sequence
1 Alice 04/26/11 londoneye -> bigben -> downingstreet -> trafalgarsquare
2 Alice 04/27/11 londoneye -> tatemodern -> towerbridge
3 Bob 04/26/11 londoneye -> bigben -> tatemodern

ID Travel sequence
1 londoneye → bigben → trafalgarsquare
2 londoneye → bigben → downingstreet → trafalgarsquare
3 londoneye → bigben → westminster
4 londoneye → tatemodern → towerbridge
5 londoneye → bigben → tatemodern
The top frequent trajectory patterns are short but not informative, e.g.,

How to rank trajectory patterns?

A trajectory pattern is important if many important users take it and it contains important locations

\[ P_T = M_{TU} \cdot P_U \quad P_T = M_{LT}^T \cdot P_L \]

A user is important if the user takes photos at important locations and visits the important trajectory patterns

\[ P_U = M_{UL} \cdot P_L \quad P_U = M_{TU}^T \cdot P_T \]

An location is important if it occurs in one or more important trajectory patterns and many important users take photos at the location

\[ P_L = M_{LT} \cdot P_T \quad P_L = M_{UL}^T \cdot P_U \]
Trajectory Pattern Ranking Algorithm

- $P_T$ is the eigen-vector for $M^TM$ for the largest eigen value, $M = M_{TU} M_{UL} M_{LT}$
- The algorithm is a normalized power iteration method to detect the eigen-vector of $M^TM$ for the largest eigen-value if the intial $P_T$ is not orthogonal to it
- Based on the algorithm, we can derive the top-trajectory in London

Algorithm: Trajectory pattern ranking

Input: $M_{TU}$, $M_{UL}$, $M_{LT}$

Output: A ranked list of trajectory patterns

1. Initialize $P_T^{(0)}$
2. Iterate
   \[
   P_L = M_{LT} \cdot P_T^{(t)} \quad P_U = M_{UL} \cdot P_L \\
   P_T = M_{TU} \cdot P_U \quad P_U = M_{TU}^T \cdot P_T \\
   P_L = M_{UL}^T \cdot P_U \quad P_T^{(t+1)} = M_{LT} \cdot P_L \\
   P_T^{(t+1)} = P_T^{(t+1)}/\|P_T^{(t+1)}\|_1
   \]
   until convergence.
3. Output the ranked list of trajectory patterns in the decreasing order of $P_T^*$, i.e., the converged $P_T$.

londoneye → bigben → downingstreet → horseguards → trafalgarsquare
Top-Ranked Trajectories Are often Highly Biased to only a few Locations

- Top-Ranked Locations in London
  - \( P_L \): the importance score for location L
  - \# user: \# users visited the location

<table>
<thead>
<tr>
<th>Location</th>
<th>( P_L )</th>
<th># User</th>
<th>Location</th>
<th>( P_L )</th>
<th># User</th>
</tr>
</thead>
<tbody>
<tr>
<td>londoneye</td>
<td>0.0157</td>
<td>528</td>
<td>southwark</td>
<td>0.0069</td>
<td>57</td>
</tr>
<tr>
<td>trafalgarstrue</td>
<td>0.0135</td>
<td>456</td>
<td>stpaulschurch</td>
<td>0.0058</td>
<td>77</td>
</tr>
<tr>
<td>bigben</td>
<td>0.0121</td>
<td>205</td>
<td>downingstreet</td>
<td>0.0053</td>
<td>52</td>
</tr>
<tr>
<td>tatemodern</td>
<td>0.0119</td>
<td>491</td>
<td>horseguards</td>
<td>0.0051</td>
<td>25</td>
</tr>
<tr>
<td>royaldealball</td>
<td>0.0093</td>
<td>175</td>
<td>londonbridge</td>
<td>0.0049</td>
<td>37</td>
</tr>
<tr>
<td>towerbridge</td>
<td>0.0089</td>
<td>185</td>
<td>embankment</td>
<td>0.0047</td>
<td>23</td>
</tr>
<tr>
<td>cityhall</td>
<td>0.0077</td>
<td>141</td>
<td>harrods</td>
<td>0.0047</td>
<td>39</td>
</tr>
<tr>
<td>waterloobridge</td>
<td>0.0076</td>
<td>198</td>
<td>toweroflondon</td>
<td>0.0046</td>
<td>91</td>
</tr>
<tr>
<td>parliamentsquare</td>
<td>0.0075</td>
<td>150</td>
<td>nationalhistorymuseum</td>
<td>0.0046</td>
<td>97</td>
</tr>
<tr>
<td>picadillycircus</td>
<td>0.0074</td>
<td>182</td>
<td>monument</td>
<td>0.0046</td>
<td>59</td>
</tr>
<tr>
<td>britishmuseum</td>
<td>0.0074</td>
<td>230</td>
<td>victoriaandalbertmuseum</td>
<td>0.0045</td>
<td>64</td>
</tr>
<tr>
<td>gherkin</td>
<td>0.0073</td>
<td>75</td>
<td>bank</td>
<td>0.0044</td>
<td>63</td>
</tr>
<tr>
<td>lloydsof</td>
<td>0.0070</td>
<td>121</td>
<td>royalacademy</td>
<td>0.0040</td>
<td>34</td>
</tr>
<tr>
<td>coventgarden</td>
<td>0.0070</td>
<td>169</td>
<td>oxfordstreet</td>
<td>0.0040</td>
<td>51</td>
</tr>
<tr>
<td>buckinghampalace</td>
<td>0.0064</td>
<td>107</td>
<td>bloomsbury</td>
<td>0.0038</td>
<td>27</td>
</tr>
</tbody>
</table>

- Top-Ranked Trajectories in London
  - highly biased to only a few locations
  - Trajectory 1 (londoneye → bigben → downingstreet → horseguards → trafalgarstrue)
  - Trajectory 5 (westminster → bigben → downingstreet → horseguards → trafalgarstrue)
From Top-Ranked to Diversified Ranked Trajectories

- Diversified Ranked Trajectories in London
  - Trajectories 2, 4, & 5 are popular routes to explore street arts in London

- Location Recommendation in London
  - Rank the locations by the scores of trajectories (append current trajectory with next destination)