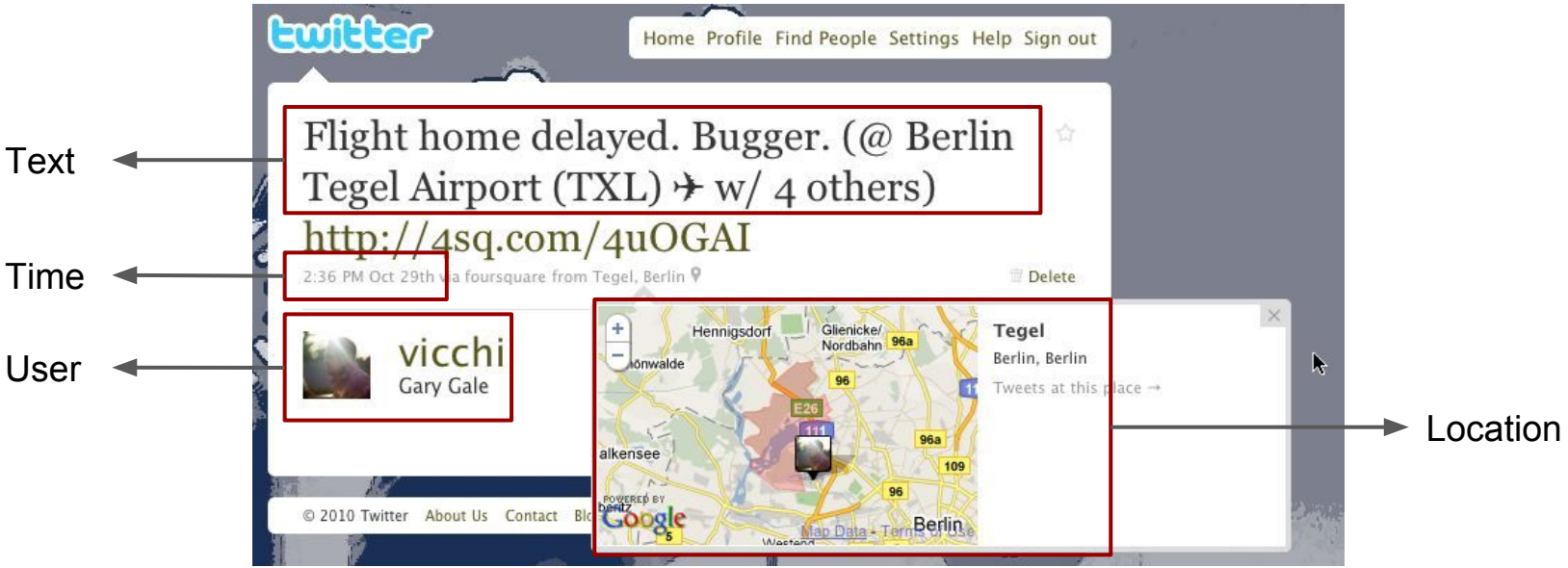
Bringing Semantics to Spatiotemporal Data Mining

Chao Zhang, Quan Yuan, Jiawei Han
University of Illinois at Urbana-Champaign
czhang82@illinois.edu

Slides: http://chaozhang.org/files/slides/slides-icde17.pdf

What is Semantics-Rich Spatiotemporal Data?

Spatiotemporal records attached with semantic information (e.g., category, text message, image, video, etc.)

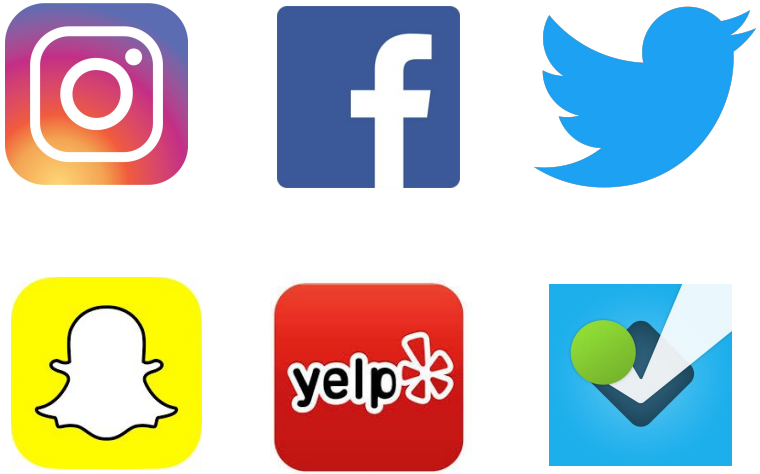


Example: User-Generated Social Media

Human sensing: millions of human sensors probe the physical world and create **geo-tagged social media** posts.

Daily stats:

- 10M+ geo-tagged tweets
- 20M+ geo-tagged Instagram posts
- 8M+ Foursquare check-ins



Example: Smartphone Usage Data

"Data collected from your smartphone is shaping the future of marketing -- and cell phone companies are cashing in."

Most smartphone carriers have access to your location, Web searches, App usage and other data.

Massive (anonymized) smartphone usage data can be acquired from carriers.

http://money.cnn.com/2013/12/16/technology/mobile/wireless-carrier-sell-data/

Example: Geo-Tagged Web Documents

Geo-tagging techniques have enabled inferring the locations of enormous Web resources.

- News, Blogs, Forum posts, Videos, etc.

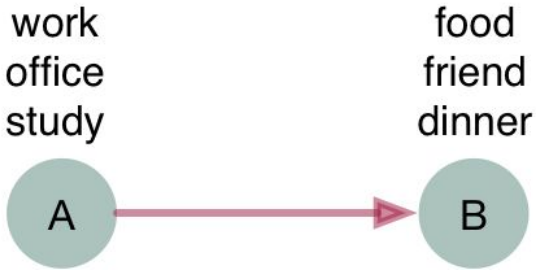
Bridging Semantics and Spatiotemporal Data

Reason 1: **Extract human-interpretable knowledge for decision making**.

Example:

- For mobility modeling, we can understand not only how people move, but also the semantics behind their movements.

Traditional mobility model



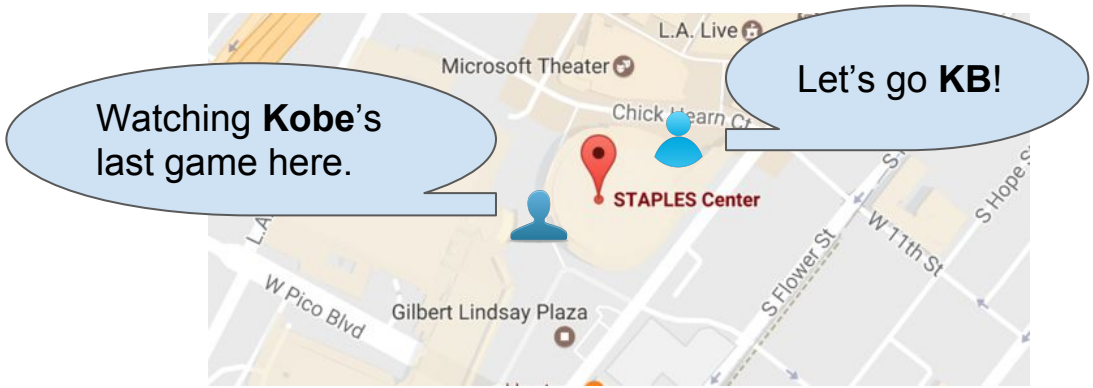
Semantics-rich mobility model

Bridging Semantics and Spatiotemporal Data

Reason 2: **Allow multiple modalities to complement and enhance each other.**

Example:

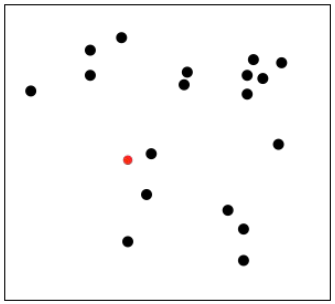
- The spatiotemporal distributions of different keywords (e.g., “Kobe” and “KB”) allows us to capture their correlations more accurately.



Key Challenges

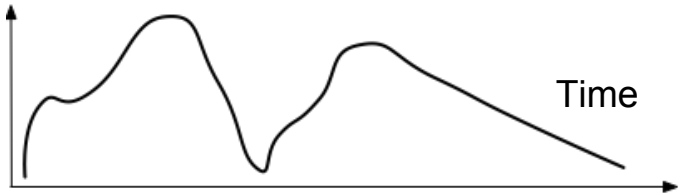
1. Joint modeling of multiple modalities

- Different representations, scales, and distributions.



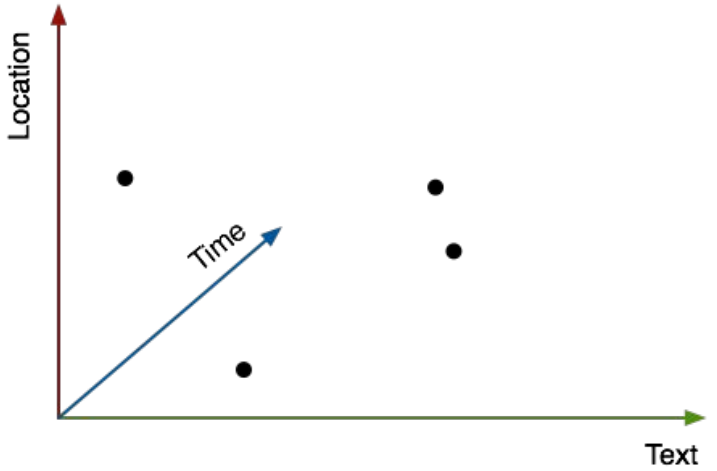
Location

Text



2. Sparsity in the multidimensional space.

- Training data scarcity
- High-dimensional text data



What Will Be Covered in This Tutorial?

1. Spatiotemporal activity modeling
   - *How to find the typical activities in different regions and periods?*
2. Spatiotemporal event detection and forecasting
   - *How to detect and forecast unusual spatiotemporal events?*
3. Spatiotemporal mobility modeling
   - *How to model human movement regularities from semantic trajectories?*

Outline

Introduction

Part I: Spatiotemporal Activity Modeling

Part II: Spatiotemporal Event Discovery

Part III: Spatiotemporal Mobility Modeling

Research Frontiers

Summary

Part I: Spatiotemporal Activity Modeling

Problem Description

Input: a collection of text-rich spatiotemporal records

- Each record has: a location, a timestamp, a text message.

Task: find people's typical activities in different regions and periods

- Multiple schemas, e.g.
  - region + time -> keywords
  - region + keyword -> time
  - time + keyword -> region

"What are the fun things to do around the Hilton hotel?"

"Where should I go to hang out with my friends at 9pm?"

An Overview of Representative Approaches

Similarity-Based Method

- Li et. al. WWW 2015

Probabilistic Graphical Modeling Method

- Sizov et. al. WSDM 2010

Representation Learning Method

- Zhang et. al. WWW 2017

An Overview of Representative Approaches

Similarity-Based Method

- Li et. al. WWW 2015

Probabilistic Graphical Modeling Method

- Sizov et. al. WSDM 2010

Representation Learning Method

- Zhang et. al. WWW 2017

Semantic Annotation of Mobility Data Using Social Media [Wu et. al. WWW 2015]

Input: GPS location history, a collection of spatiotemporal social media

Goal: annotate GPS location history with social media keywords to reflect user activities.

Why using social media?

- "Static annotation" vs "dynamic annotation"
- Including up-to-date event information.

Input: Location history of a mobile user

Record ID	Time	Longitude	Latitude
r_1	2013-1-20	40.75051	-73.99349
r_2	2013-2-10	40.68312	-73.97597
r_3	2013-2-19	40.75051	-73.993499

Annotation

Record ID	Annotations
r_1	madison, garden, rangers, penguins
r_2	nets, barclays, center, nba
r_3	rangers, madison, square, montreal, canadiens

Input: Geo-tagged tweets from the crowd

Time	Longitude	Latitude	Tweets
2013-1-20	40.75051	-73.99349	LETS GO RANGERS
2013-1-20	40.61219	-74.15814	I'm at Buffalo Wild Wings
2013-1-20	40.75050	-73.99350	I'm @ Madison Square Garden for Pittsburgh Penguins vs New York Rangers

Three Similarity-Based Annotation Strategies

Frequency-Based Method

- Use TF-IDF weighting to select representative keywords

Gaussian Model

- Model the distribution of a keyword as a mixture of Gaussians

Kernel Density Estimation

- Estimating the kernel densities of keywords based on observed samples

Kernel Density Estimation: More Details

Assumption: the semantics of a location record *ri* can be inferred from the documents posted at nearby locations within a short time

Time sensitivity: define a time window and collect documents *Di* that fall into it.

Relevance: Rank words based on the scores estimated by Kernel Density Estimation (KDE)

$$s_w^{KDE}(r_i; L_i(w), H) = \frac{1}{|L_i(w)|} \sum_{loc_j^D \in L_i(w)} K_H(loc_i^U - loc_j^D)$$

Locations of all the occurrences of word w

Kernel function

Geo-coordinates of record *ri*

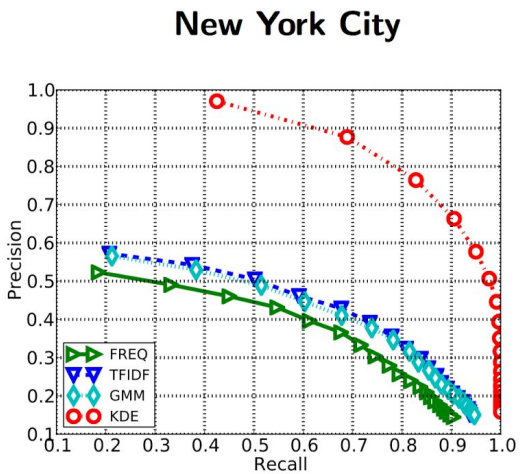
$$K_H(x) = |H|^{-1/2} K(H^{-1/2}(x))$$

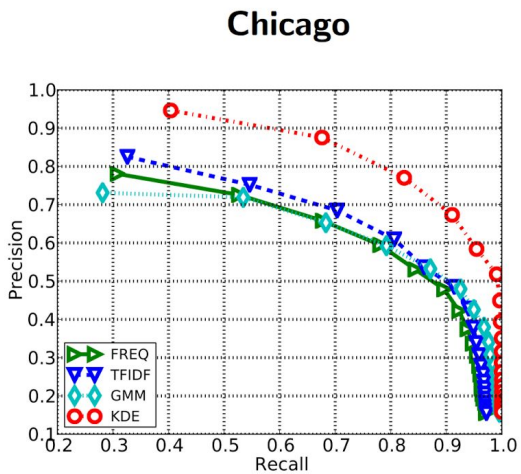
Bandwidth

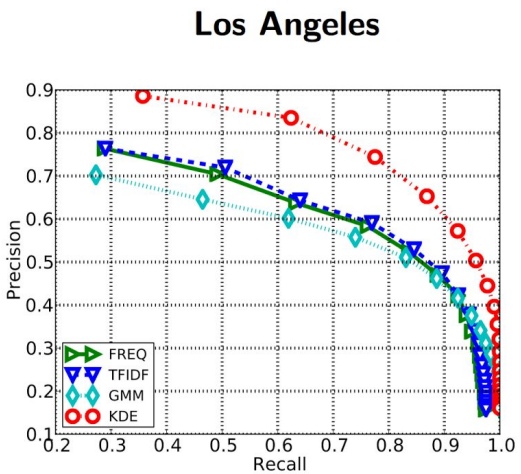
Experiments

Datasets

- Context documents: geo-tagged tweets from NYC, Chicago and LA
- Users' location histories: GPS coordinates and timestamps of tweets
- Groundtruth: manually judge whether the extracted keywords reflect user intention







An Overview of Representative Approaches

Similarity-Based Method

- Li et. al. WWW 2015

Probabilistic Graphical Modeling Method

- Sizov et. al. WSDM 2010

Representation Learning Method

- Zhang et. al. WWW 2017

GeoFolk: Latent Spatial Semantics in Web 2.0 Social Media [Sizov et. al. WSDM 2010, TIST 2012]

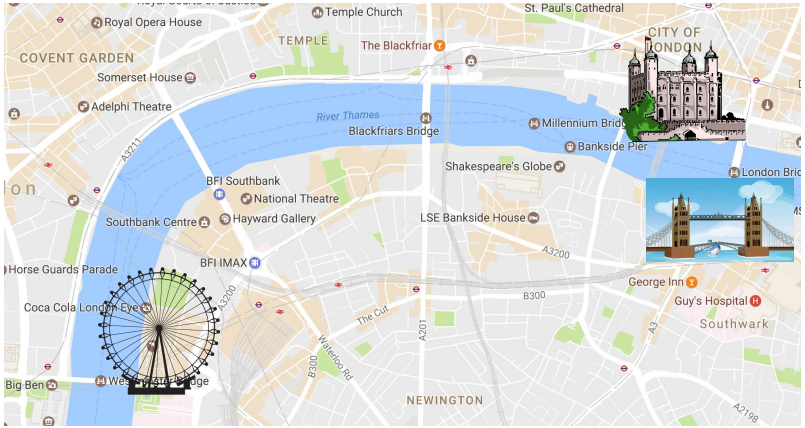
Goal: use topic modeling to uncover the latent activities from social media.

Each latent topic is **multidimensional** (location, time, text):

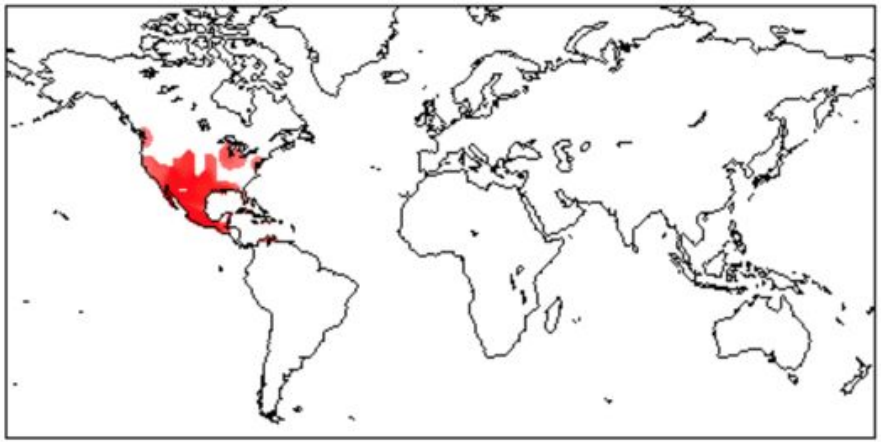
- Single dimension: insufficient for reliable content disambiguation
- Combination: better content characterization

Applications

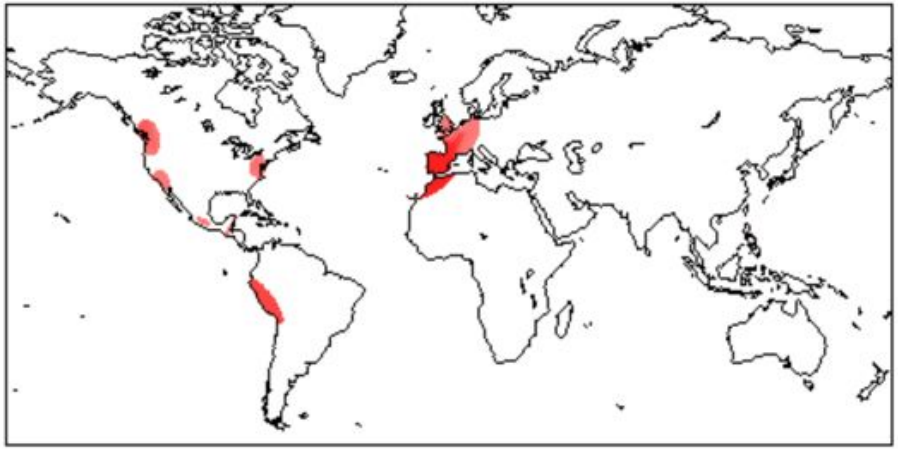
- Topic modeling for regions
- Recommender systems
- Content categorization



Examples



Mexican Food

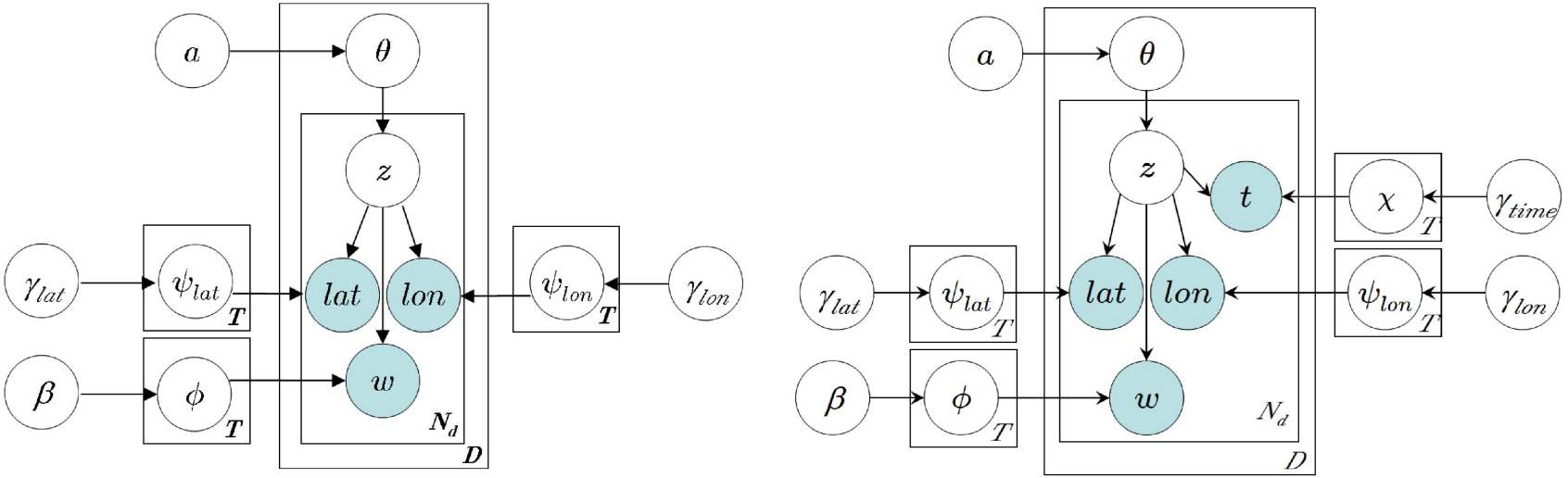


Spanish Food

Geographical Topic Discovery and Comparison, Yin et al, WWW 2011

GeoFolk: Geographical Topic Modeling

Input: each document is associated with tags, GPS coordinates (and time).



Experiments

Dataset: 28,770 flickr images with tags and coordinates

Tasks 1: Classification and clustering

Model	avg(accuracy)
GeoFolk	0.421
LDA	0.374
Tags	0.282
Coordinates	0.187

classification

Model	avg(accuracy)
GeoFolk	0.328
LDA	0.255
Tags	0.117
Coordinates	0.102

clustering

Task 2: Tag recommendation

Model	MRR
GeoFolk	0.212
LDA	0.119
Tags	0.073
Coordinates	0.027

An Overview of Representative Approaches

Similarity-Based Method

- Li et. al. WWW 2015

Probabilistic Graphical Modeling Method

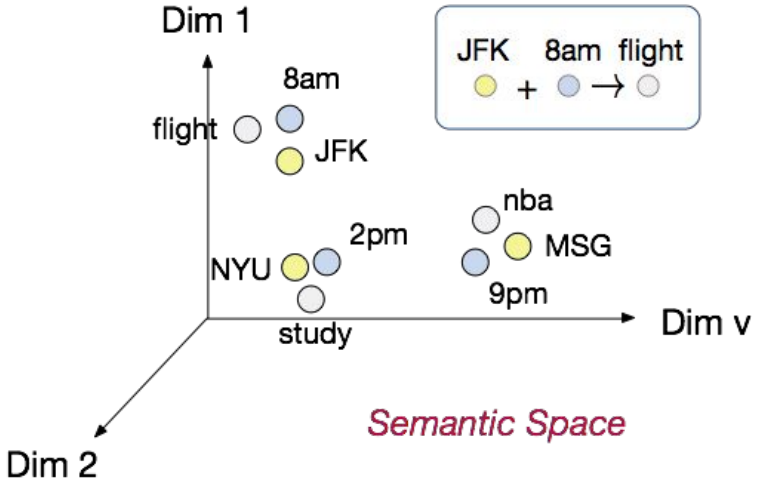
- Sizov et. al. WSDM 2010

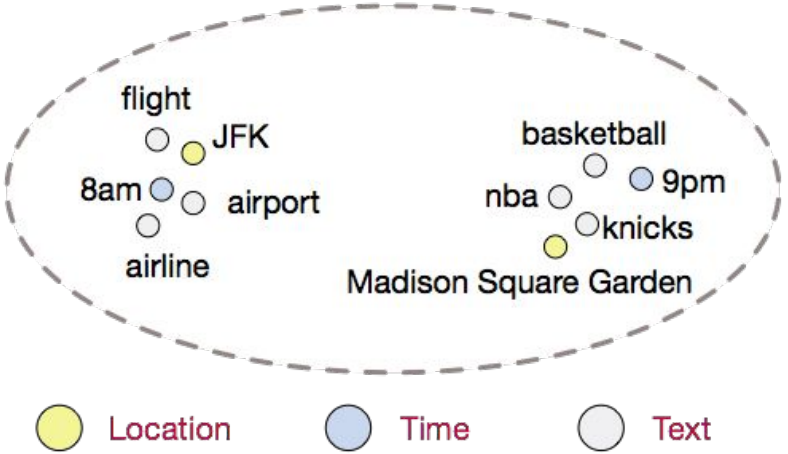
Representation Learning Method

- Zhang et. al. WWW 2017

Regions, Periods, Activities: Uncovering Urban Dynamics via Cross-Modal Representation Learning [Zhang et. al. WWW 2017]

Idea: map geographical regions, temporal periods, and textual keywords into a latent semantic space to preserve their correlations.

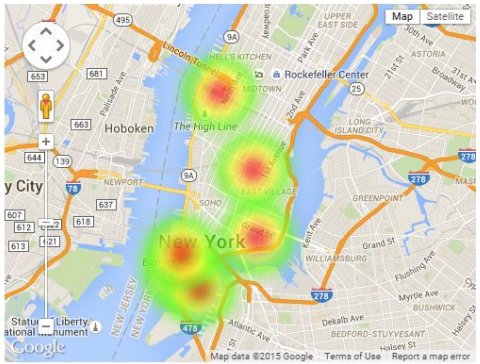




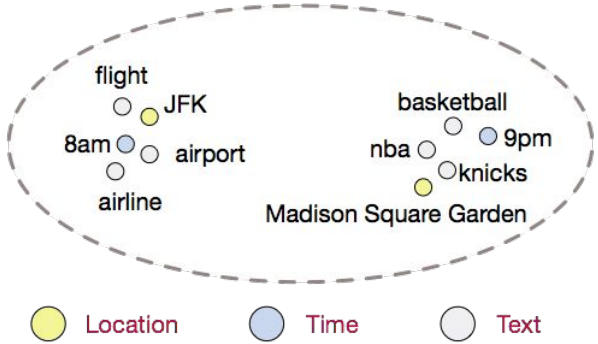
Regions, Periods, Activities: Uncovering Urban Dynamics via Cross-Modal Representation Learning [Zhang et. al. WWW 2017]

Hotspot Detection → Joint Embedding

Detect regions and periods where people's activities burst



Map regions, periods, and keywords into the same space



Hotspot Detection: A Mode-Seeking Procedure

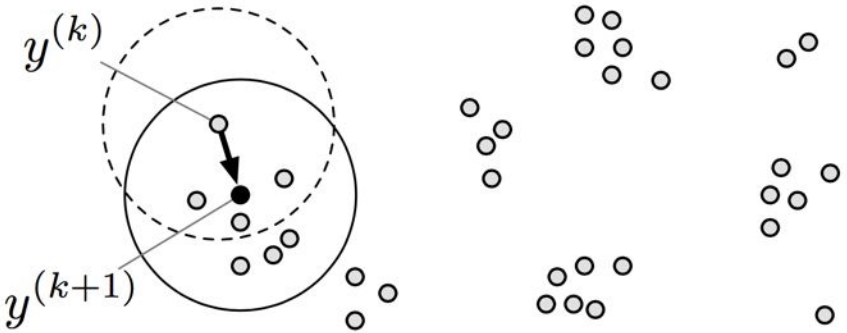
- A spatial (temporal) hotspot is a density maxima in the 2D (1D) space
- We design a fast mode seeking procedure to find the hotspots.

Benefits:

- Fast
- No distributional assumptions

Kernel density estimation:

$$f(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{x}_i}{h})$$



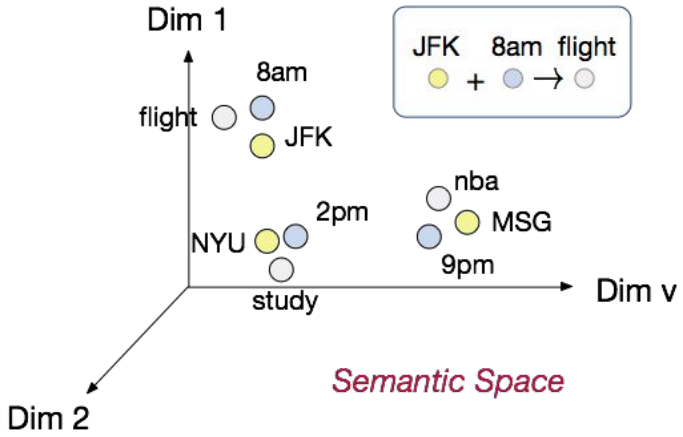
Cross-Modal Embedding: Designing Philosophy

Map regions, periods, and keywords into the same space:

- **Region**: a spatial hotspot
- **Period**: a temporal hotspot

Aim to preserve two types of correlations:

1. **Co-occurrence**: two units are correlated if there co-occur frequently
2. **Neighborhood**: two regions (periods) are correlated if they are adjacent



Cross-Modal Embedding: Two Strategies

Reconstruction-based embedding

1. Consider each record as a relation
2. Mark off one unit i and try to predict it from the observed units

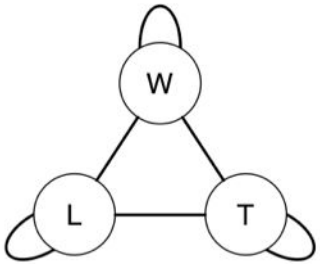
$$p(i|r_{-i}) = \exp(s(i, r_{-i}))/\sum_{j \in X} \exp(s(j, r_{-i}))$$

Overall objective function:

$$O = -\sum_{r \in \mathcal{C}} \sum_{i \in r} \log p(i|r_{-i})$$

Graph-based embedding

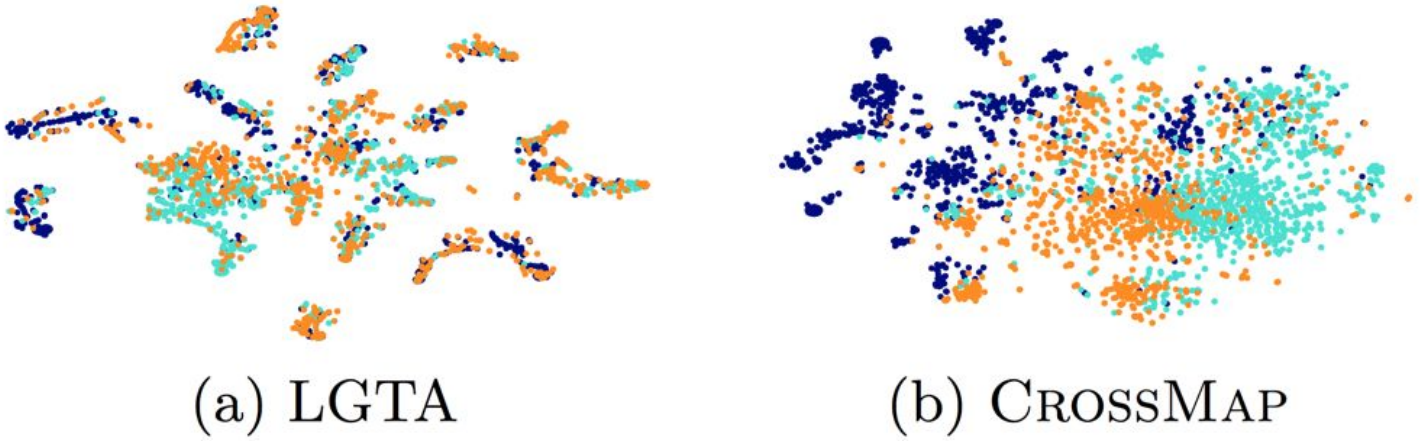
1. Use a graph to encode the correlations between regions, periods, and activities
2. Learn graph node embeddings to preserve the graph structure



$$O = O_{WW} + O_{LL} + O_{TT} + O_{WL} + O_{WT} + O_{LT}$$

$$O_{XY} = \sum_{i \in X} d_i \mathrm{KL}(p'(\cdot|i)||p(\cdot|i)) + \sum_{j \in Y} d_j \mathrm{KL}(p'(\cdot|j)||p(\cdot|j))$$

Embedding Visualization



(a) LGTA (b) CrossMap

Visualizing the feature vectors generated by LGTA and CrossMap for three activity categories: “Food” (cyan), “Travel & Transport” (blue), and “Residence” (orange).

Quantitative Evaluation: Attribute Recovery

Mark off one attribute (location, time, or text) and predict it based on the observed ones.

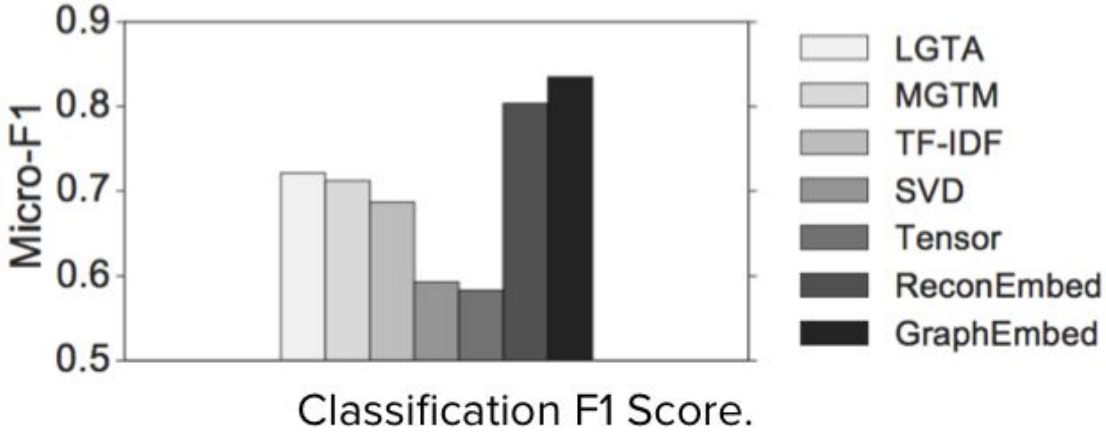
Mean reciprocal ranks:

	Text		Location		Time	
Method	Tweet	4SQ	Tweet	4SQ	Tweet	4SQ
LGTA	0.376	0.6107	0.3792	0.6083	-	-
MGTM	0.3874	0.5974	0.4474	0.5753	-	-
TF-IDF	0.62	0.8505	0.4298	0.7097	0.3197	0.3431
SVD	0.4475	0.7137	0.3953	0.646	0.3256	0.3187
Tensor	0.4382	0.6826	0.3871	0.6251	0.3179	0.2983
RECON	0.6877	0.9219	0.6526	0.9044	0.3582	0.3612
GRAPH	**0.7011**	**0.9449**	**0.6758**	**0.9168**	**0.3895**	**0.3716**

Application: Activity Classification

The embeddings can be used as feature vectors for downstream applications.

Example: Foursquare checkins belong to nine categories. We predict the category based on the embeddings.



Reference

Annotate POIs with categories

- On the semantic annotation of places in location-based social networks. Ye et al. KDD 2011
- Placer: semantic place labels from diary data. Krumm et al. UbiComp 2013

Annotate Regions with functions

- Inferring urban land use using large-scale social media check-in data. Zhang et al. Networks and Spatial Economics 2014
- Geographical topic discovery and comparison, Yin et al. WWW 2011
- Discovering regions of different functions in a city using human mobility and pois. Yuan et al. KDD 2012
- Latent geospatial semantics of social media, Sizov et al. TIST 2012

Annotate Visits with Semantics

- A conceptual view on trajectories. Spaccapietra et al. DKE 2008
- Semantic trajectories: Mobility data computation and annotation. Yan et al. TIST 2013
- Semantic annotation of mobility data using social media. Wu et al. WWW 2015

Multimodal Representation Learning

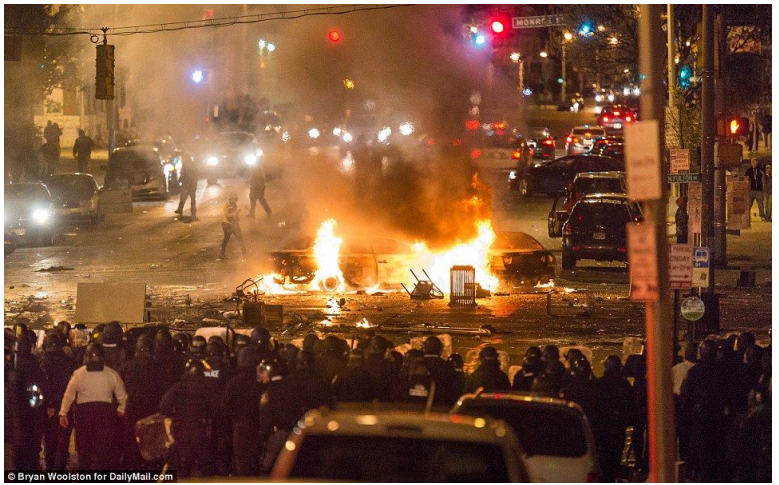
- Regions, Periods, Activities: Uncovering Urban Dynamics via Cross-Modal Representation Learning. Zhang et al. WWW 2017

Part II: Spatiotemporal Event Discovery

What is a Spatiotemporal Event?

An unusual activity bursted within a local area and a specific duration:

- E.g., protest, disaster, sport game, concert



Spatiotemporal Event Detection from Social Media

Batch Detection

Input: a static social media dataset

Output: all the spatiotemporal events occurred in history

Online Detection

Input: a continuous social media stream, a query window Q

Output:

- Detect all the spatiotemporal events occurred in Q
- Update the event list as Q shifts continuously

An Overview of Representative Approaches

Feature-Based Detection: First detect bursty keywords/phrases from the input, then group relevant features into events.

- E.g., Chen et. al. CIKM 2009, Abdelhaq et. al. PVLDB 2013

Document-Based Detection: Consider each document (e.g., tweet, check-in) as a basic unit and detect bursty document clusters as events.

- E.g., Zhang et. al. SIGIR 2016

Spatiotemporal Event Forecasting: Predict whether a specific type of spatiotemporal event will occur on a given day.

- E.g., Zhao et. al. KDD 2016

An Overview of Representative Approaches

Feature-Based Detection: First detect bursty keywords/phrases from the input, then group relevant features into events.

- E.g., Chen et. al. CIKM 2009, Abdelhaq et. al. PVLDB 2013

Document-Based Detection: Consider each document (e.g., tweet, check-in) as a basic unit and detect bursty document clusters as events.

- E.g., Zhang et. al. SIGIR 2016

Spatiotemporal Event Forecasting: Predict whether a specific type of spatiotemporal event will occur on a given day.

- E.g., Zhao et. al. KDD 2016

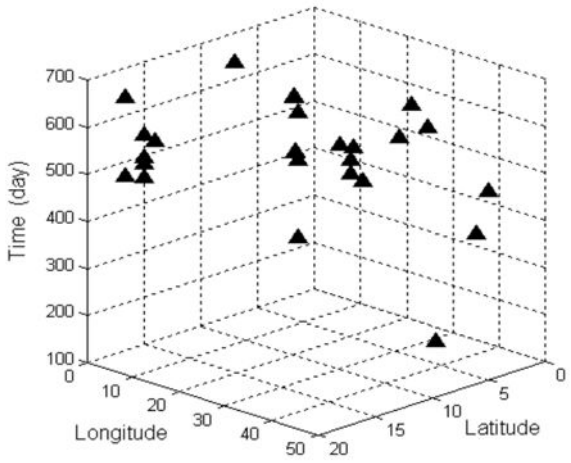
Event Detection from Flickr Data through Wavelet-based Spatial Analysis [Chen et. al. CIKM 2009]

A feature-based approach:

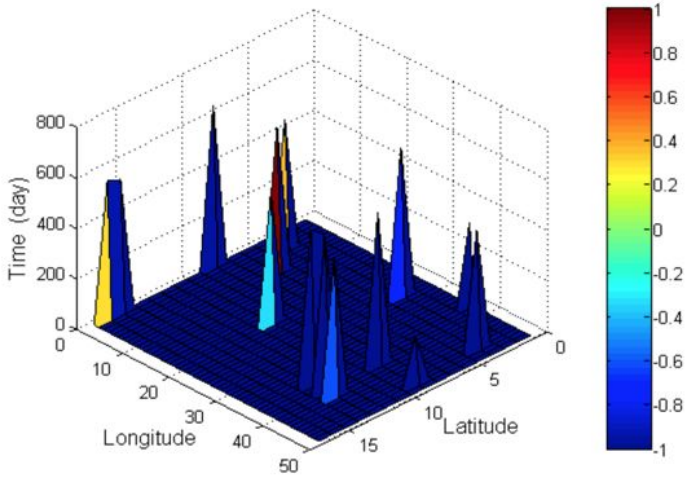
1. User Wavelet analysis to find spatiotemporally localized Flickr tags
2. Cluster event-related tags into events based on both semantic and spatiotemporal similarities

Detecting Event-related Tags

- For each tag, map every occurrence into a point in the 3D (lat, lng, time) space
- User Wavelet analysis handle 3D signals and find event-related (bursty) tags



(a) usage occurrences in the original 3D space



(b) surface plot in the original 3D space

Event Generation

Cluster bursty tags into spatiotemporal events:

Tag similarity

$$S(q_i, q_j) = \frac{SemSim(q_i, q_j)}{1 + SpaDist(q_i, q_j)}$$

Semantic similarity

Spatiotemporal similarity

$$SemSim(q_i, q_j) = \frac{|P(q_i) \cap P(q_j)|}{\min\{|P(q_i)|, |P(q_j)|\}}$$

$$KL^N(m_i, sd_i; m_j, sd_j) =$$

$$\frac{1}{2}(\log(\frac{sd_j^2}{sd_i^2}) + \frac{sd_i^2}{sd_j^2} + \frac{(m_i - m_j)^2}{sd_j^2} - 1)$$

Example Bursty Tags and Events

No.	Event Tags	Time	Location (la, lo)	Event Description
E_1	partnershipwalk akf agakhanfoundation	10/29/2006, 11/10/2007	(29.719322, -95.37212)	Partnership Walk is an initiative of Aga Khan Foundation USA to raise funds and awareness to help communities in Africa and Asia. It is held annually at Atlanta, Chicago, Dallas, Houston, Los Angeles.
E_2	southoaklandcountysoccer socs storm95	09/15/2007, 09/22/2007, 09/29/2007, 10/07/2007	(42.49387, -83.20573)	Weekly games of team SOCS Storm95 in south oakland country soccer club in 2007.
E_3	crosswalkamerica crosswalk scottgriessel creatista griessel	07/02/2006, 08/01/2006, 08/20/2006, 09/01/2006, 07/02/2007, 08/06/2007, 08/23/2007, 09/01/2007	(33.99294, -110.07808)	Crosswalk is a journey made by a couple of progressive Christians who trekked across the country from April to September. Griessel is the photographer of this walk.
E_4	f1 formulaone unitedstatesgrandprix	07/02/2006, 06/17/2007	(39.693844, -86.23974)	The United States Grand Prix was a Formula One race held on July 2, 2006, and June 15-17, 2007, at the Indianapolis Motor Speedway.
E_5	asl northpark deaf gpccd	04/22/2006, 04/14/2007	(34.239143, -116.894745)	The annual ASL fundraising picnic party at Pittsburgh North Park hosted by GPCCD in April.
E_6	beachjam amusementrides moreyspiers wildwoodbeachjam amusements beachcamping	05/20/2006, 05/20/2007	(38.987007, -74.81043)	The Beach Jam is an annual camping event on the Wildwood, NJ, beach at Morey's Piers that includ amusement rides. There is a 3-day Spring Beach Jam before Memorial Day.

EvenTweet: Online Localized Event Detection from Twitter [Abdelhaq et. al. PVLDB 2013]

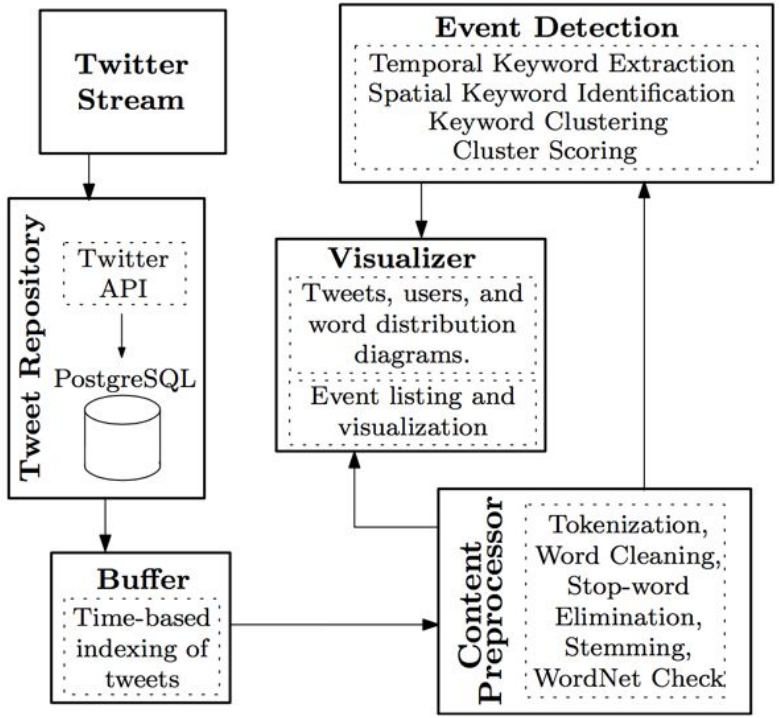


Figure 1: System overview of EvenTweet

Online local detection:

- Partition time into equal-size bins
- Trigger the detector when the current bin is saturated

Feature-based detection:

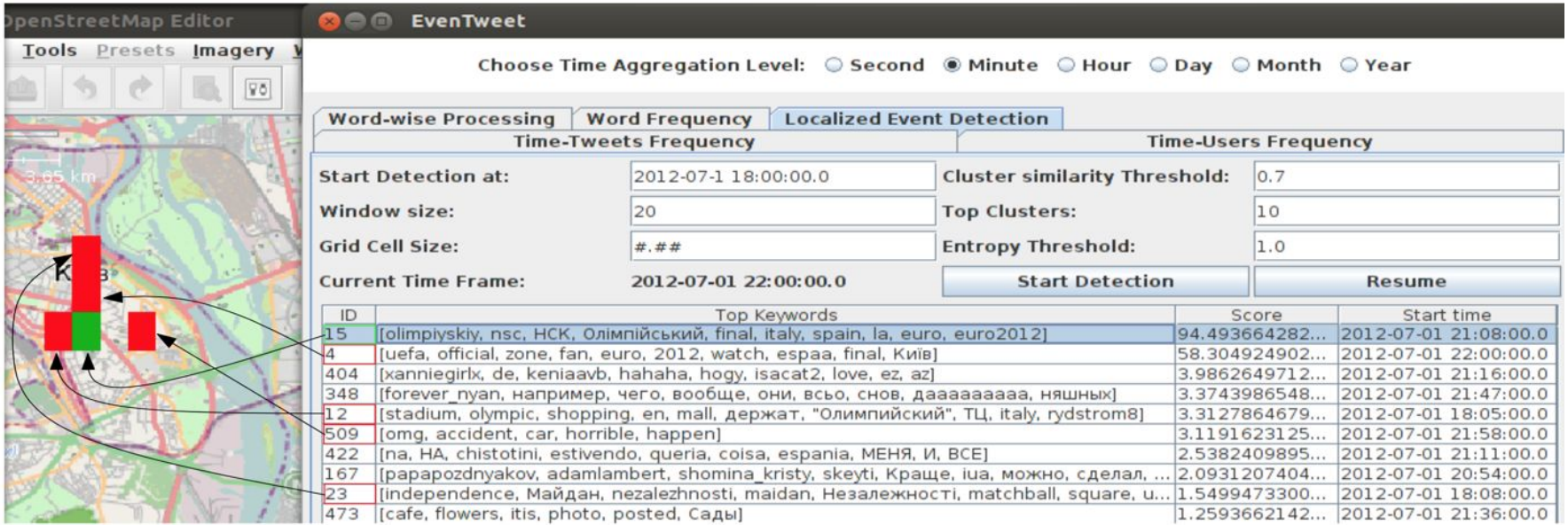
- Each keyword is considered as a feature
- Detect bursty features and cluster them into local events

The Online Detection Process

1. Select temporally bursty keywords from the current query bin by comparing against previous bins.
2. Select spatially localized keywords by computing keyword entropies.
3. Cluster the localized keywords into events based on spatial distributions.
4. Compute the burstiness score of the clusters and rank them.

Example [Abdelhaq et. al. PVLDB 2013]

A local event is represented as a collection of bursty and localized keywords:



An Overview of Representative Approaches

Feature-Based Detection: First detect bursty keywords/phrases from the input, then group relevant features into events.

- E.g., Chen et. al. CIKM 2009, Abdelhaq et. al. PVLDB 2013

Document-Based Detection: Consider each document (e.g., tweet, check-in) as a basic unit and detect bursty document clusters as events.

- E.g., Zhang et. al. SIGIR 2016

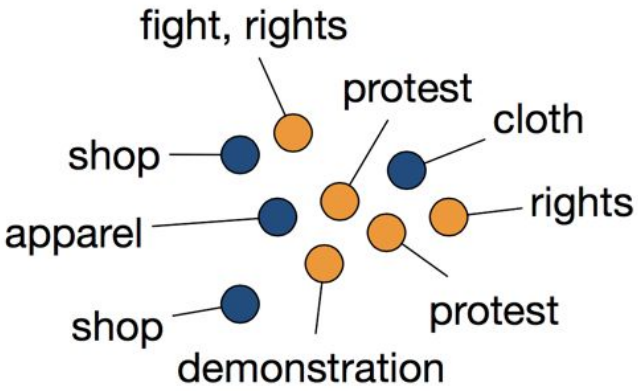
Spatiotemporal Event Forecasting: Predict whether a specific type of spatiotemporal event will occur on a given day.

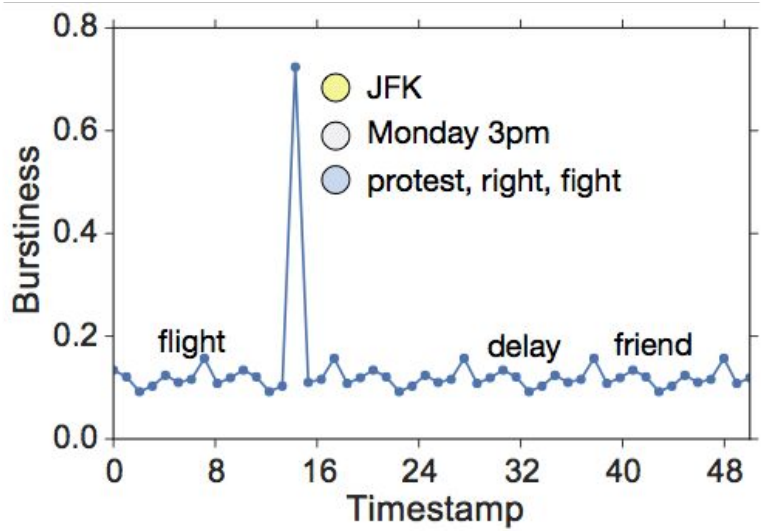
- E.g., Zhao et. al. KDD 2016

GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams [Zhang et. al. SIGIR 2016]

A document-based approach:

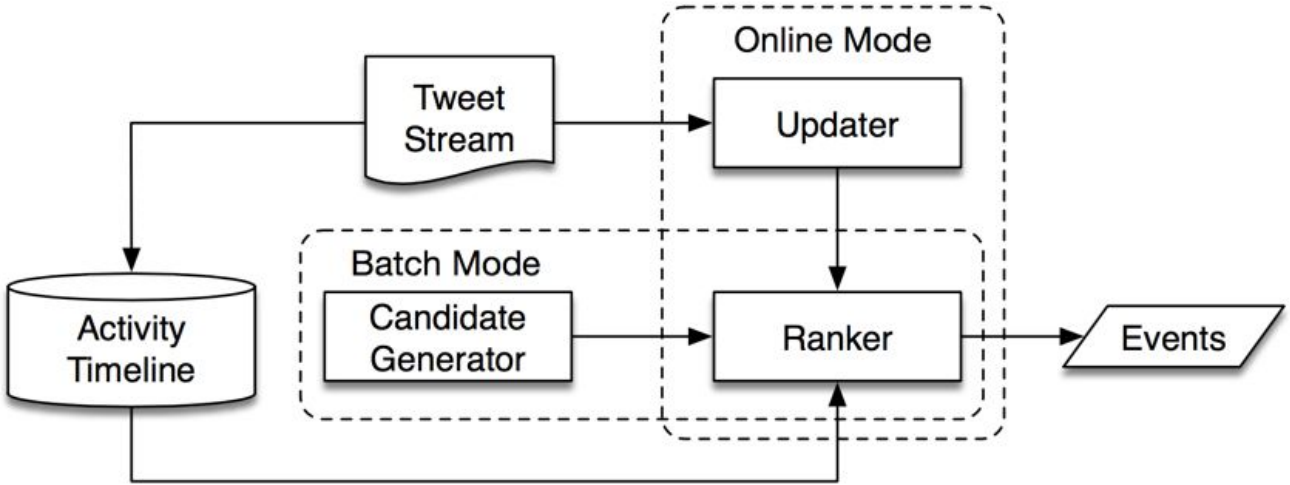
A local event is a geo-topic cluster that is spatiotemporally bursty.





GeoBurst: A Two-Step Detector

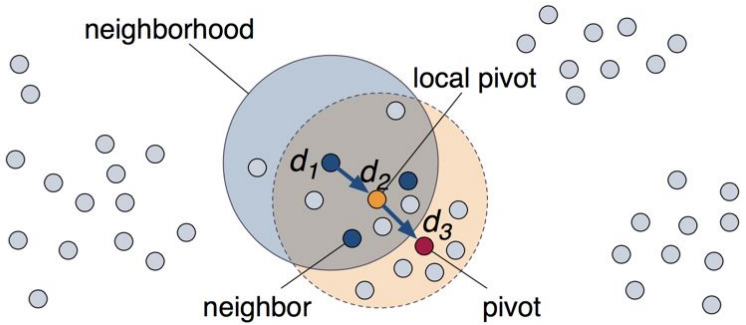
1. Candidate generation: detect all the geo-topic clusters in the query window
2. Candidate ranking: select top-K candidates by spatiotemporal burstiness



Candidate Generation & Ranking

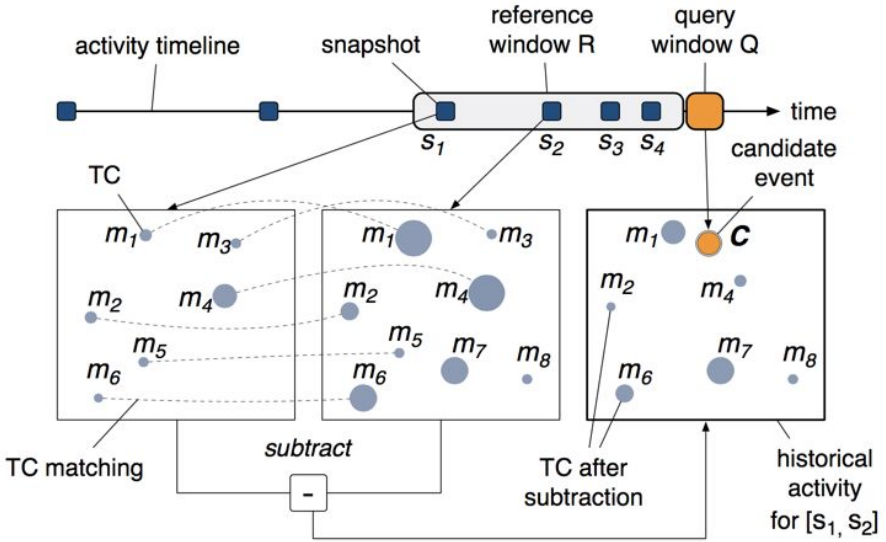
Candidate generation

- The event occurring spot acts as a ***pivot*** that produces relevant tweets around it.
- Define geo-topic authorities and perform authority ascent to find pivot tweets.



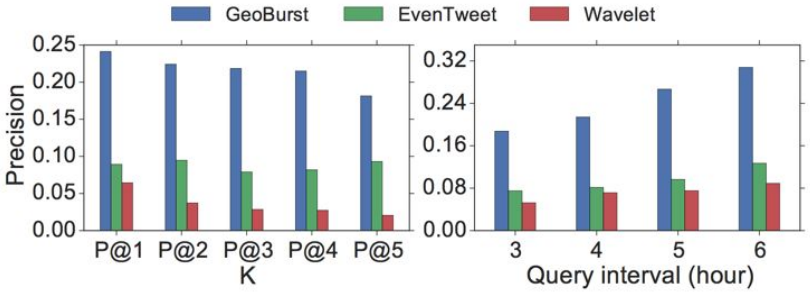
Candidate filtering

Summarize typical activities in different regions to rank the candidates.

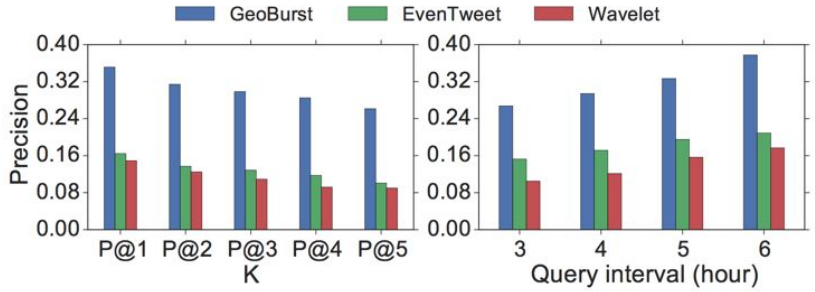


Empirical Performance

Effectiveness comparison:

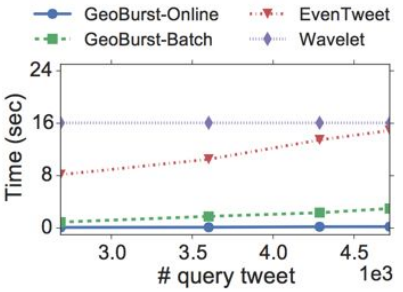


(a) Precision comparison (NY).

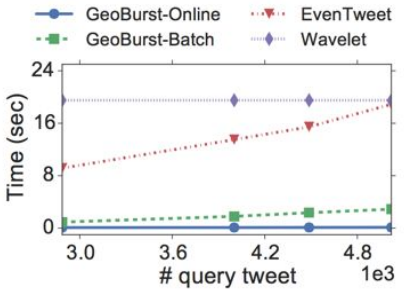


(b) Precision comparison (LA).

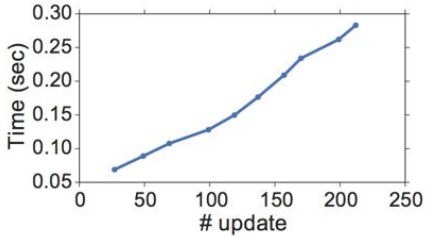
Efficiency:



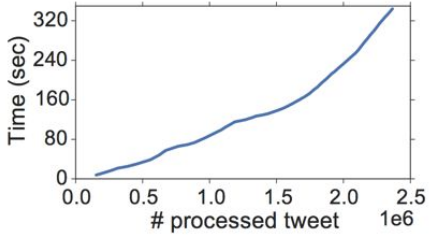
(a) Running time (NY).



(b) Running time (LA).



(a) Time v.s. # update (NY).



(a) Time v.s. # tweet (NY).

An Overview of Representative Approaches

Feature-Based Detection: First detect bursty keywords/phrases from the input, then group relevant features into events.

- E.g., Chen et. al. CIKM 2009, Abdelhaq et. al. PVLDB 2013

Document-Based Detection: Consider each document (e.g., tweet, check-in) as a basic unit and detect bursty document clusters as events.

- E.g., Zhang et. al. SIGIR 2016

Spatiotemporal Event Forecasting: Predict whether a specific type of spatiotemporal event will occur on a given day.

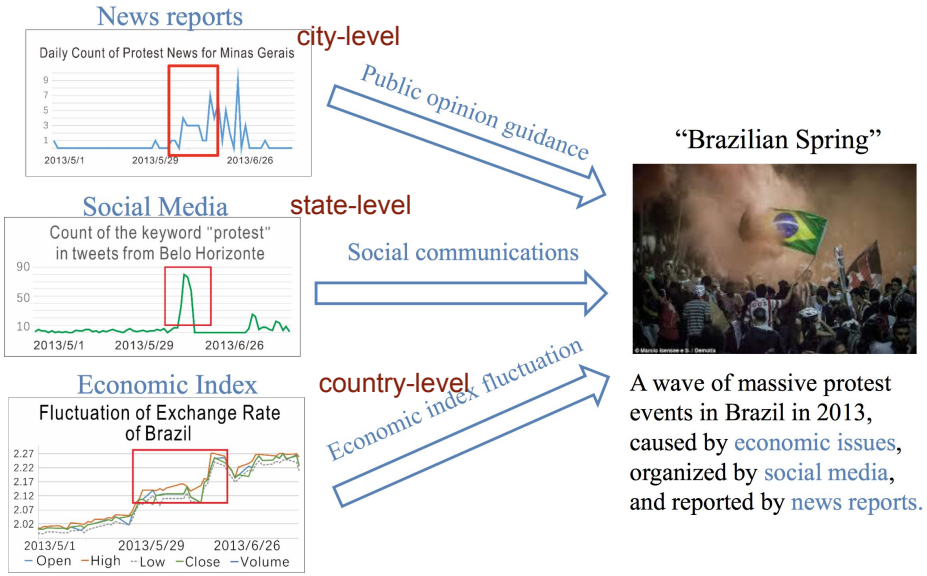
- E.g., Zhao et. al. KDD 2016

Hierarchical Incomplete Multi-Source Feature Learning for Spatiotemporal Event Forecasting [Zhao et. al. KDD 2016]

Task: use multiple data sources to predict whether certain event types will occur in the future.

Why multiple data sources?

- Spatiotemporal events are often influenced by different aspects of the society.
- Different data sources complement each other.
- One single source cannot cover all aspects of an event.



Slide from Liang Zhao:
http://people.cs.vt.edu/liangz8/materials/papers/HIML_slides.pdf

Challenges

Hierarchical topology

- E.g., country-level, state-level, city-level
- Higher-level features can influence lower-level ones

Interactive missing values

- Different data sources, different spans
- Need to consider the interactions among different sources.

Feature sparsity

- Only a small set of features are useful
- Need to use geo-hierarchy to select useful features

Hierarchical Incomplete Multi-source Feature Learning

Given the multi-source data for a location l at time t, predict whether the event will happen at time τ

$$f : \{X_{t,l_1}, \cdots, X_{t,l_N}\} \rightarrow Y_{\tau,l}$$

city, state,... ,country

- Each location has features at multiple levels *l=(l1, l2, ...lN)* E.g., (San Francisco, CA, USA)

Variables are dependent on the variables in their parent level

$$(level-1)\quad Y_{\tau,l} = \alpha_0 + \sum_{i=1}^{|\mathcal{F}_1|} \alpha_i^T \cdot [X_{t,l_1}]_i + \varepsilon$$ city-level

$$(level-2)\quad \alpha_i = \beta_{i,0} + \sum_{j=1}^{|\mathcal{F}_2|} \beta_{i,j}^T \cdot [X_{t,l_2}]_j + \varepsilon_i$$ state-level

$$(level-3)\quad \beta_{i,j} = W_{i,j,0} + \sum_{k=1}^{|\mathcal{F}_3|} W_{i,j,k}^T \cdot [X_{t,l_3}]_k + \varepsilon_{i,j}$$ country-level

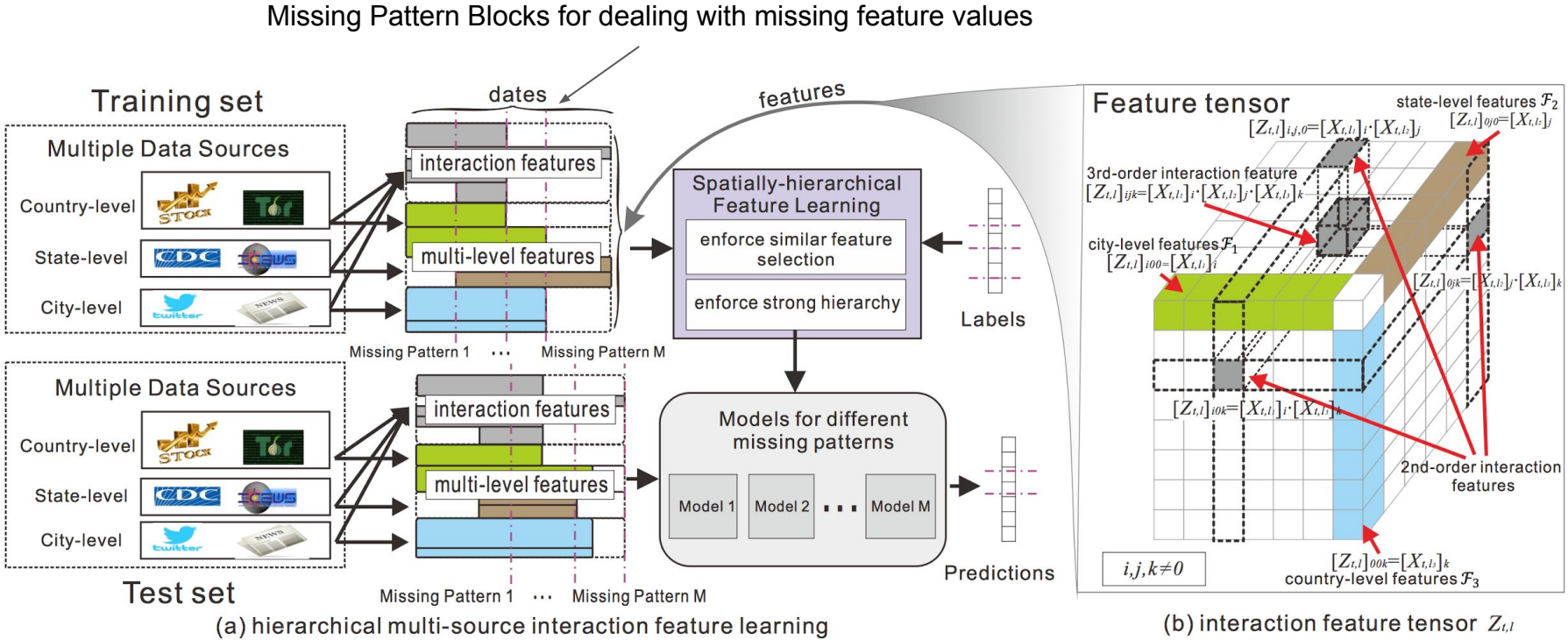
Encode hierarchical feature correlation by nth-order strong hierarchy

$$Y_{\tau,l} = \sum_{i=0}^{|\mathcal{F}_1|} \sum_{j=0}^{|\mathcal{F}_2|} \sum_{k=0}^{|\mathcal{F}_3|} W_{i,j,k} \cdot [X_{t,l_1}]_i \cdot [X_{t,l_2}]_j \cdot [X_{t,l_3}]_k + \varepsilon$$

Tensor form:

$$Y_{\tau,l} = W \odot Z_{t,l} + \varepsilon$$

Hierarchical Incomplete Multi-source Feature Learning



Experiments

Datasets: 10 datasets for civil unrest (CU) and influenza (FLU)

Missing data ratio (3%)

Method	Argentina	Brazil	Chile	Colombia	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
LASSO	0.5267	0.7476	0.5624	0.8032	0.3148	0.7823	0.5572	0.4693	0.8073
LASSO-INT	0.5268	0.7191	0.5935	0.7861	0.5269	0.777	0.4887	0.5069	0.7543
iMSF	0.4795	0.4611	0.5033	0.7213	0.5	0.5569	0.4486	0.4904	0.5
MTL	0.3885	0.5017	0.5011	0.4334	0.3452	0.4674	0.4313	0.3507	0.5501
Baseline	0.5065	0.7317	**0.6148**	0.8084	**0.777**	**0.8037**	0.7339	0.7264	**0.7846**
HIML	**0.5873**	**0.8353**	0.5705	**0.8169**	0.7191	0.7973	**0.7478**	**0.8537**	0.7488

CU forecasting performance AUC

	Missing data ratio				runtime
Method	21%	30%	50%	70%	(second)
LASSO	0.9180	0.9056	0.9036	0.8753	493.92
LASSO-INT	0.9142	0.9027	0.9073	0.8403	508.49
iMSF	0.8949	0.8899	0.8930	0.8628	88.90
MTL	0.6129	0.5303	0.6253	0.5568	223.78
Baseline	0.9044	0.9045	0.8562	0.4359	**31.97**
HIML	**0.9372**	**0.9368**	**0.9364**	**0.9357**	851.83

FLI forecasting performance AUC

Reference

Spatiotemporal Event Detection

- Event detection from flickr data through wavelet-based spatial analysis. Chen et al. CIKM 2009
- A probabilistic model for spatio-temporal signal extraction from social media. GIS 2013
- Identifying local events via space-time signals in twitter feeds. Krumm et al. GIS 2015
- Eventweet: Online localized event detection from twitter. Adbelhaq et al. PVLDB 2013
- GeoBurst: Real-time local event detection in geo-tagged tweet streams. Zhang et al. 2016
- Earthquake shakes twitter users: real-time event detection by social sensors. Sakaki et al. WWW 2010
- Crowd sensing of traffic anomalies based on human mobility and social media. Pan et al. GIS 2013
- Extracting city traffic events from social streams. Anantharam et al. TIST 2015
- Citywide traffic congestion estimation with social media. Wang et al. GIS 2015

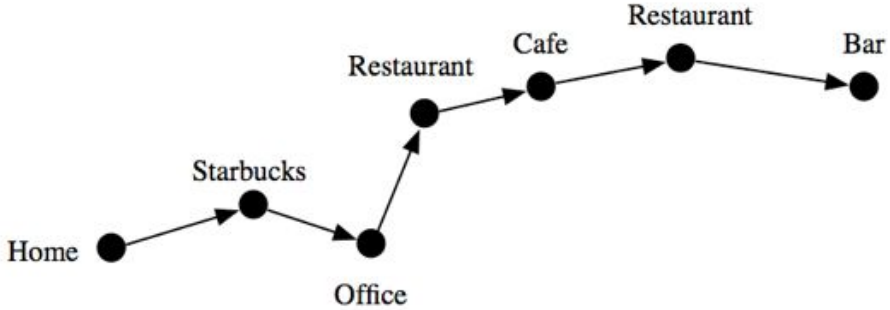
Spatiotemporal Event Forecasting

- “beating the news” with embers: forecasting civil unrest using open source indicators. Ramakrishnan et al. KDD 2014
- Spatiotemporal event forecasting in social media. Zhao et al. SDM 2015
- Combining heterogeneous data sources for civil unrest forecasting. Korkmaz et al. ASONAM 2015
- Hierarchical Incomplete Multisource Feature Learning for Spatiotemporal Event Forecasting. Zhao et al. KDD 2016
- Spatiotemporal model fusion: multiscale modelling of civil unrest. Hoegh et al. J. the Royal Statistical Society 2016

Part III: Spatiotemporal Mobility Modeling

Problem Description

Semantic trajectory: each GPS record also has semantic information (e.g., place category, text message)



Task: given a collection of semantic trajectories, how to model the movement regularities of the populace?

- Mobility modeling can occur at either **user level** or **crowd level**.

An Overview of Representative Approaches

Pattern-based approaches: mine movement patterns from semantic trajectories

- Sequential pattern mining (E.g., Zhang et. al. PVLDB 2014)

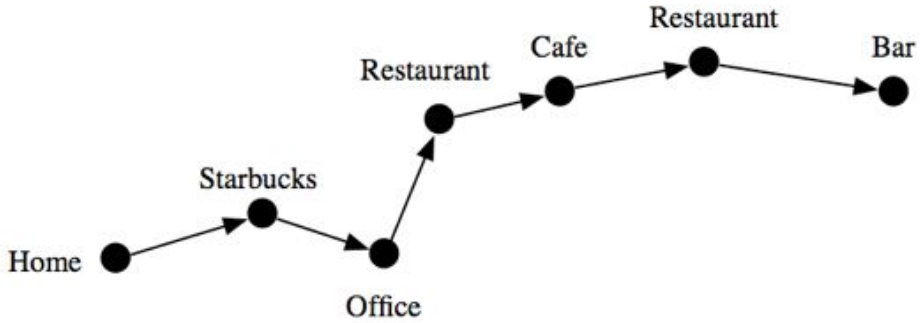
Model-based approaches: build statistical models to describe human movement

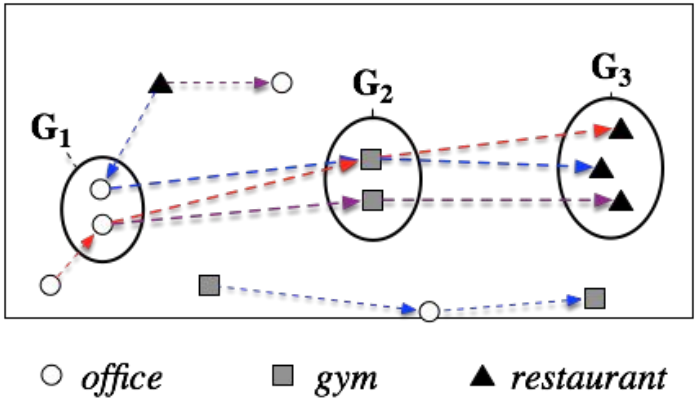
- Crowd-level sequential models (E.g., Zhang et. al. KDD 2016)
- User-level mobility models (E.g., Yuan et. al. KDD 2013)

Splitter: Mining Frequent Sequential Patterns from Semantic Trajectories [Zhang et. al. PVLDB 2014]

Each record in the trajectory has category information (e.g., office, hotel, gym)

Frequent sequential movement pattern: a movement sequence that frequently appear in the input trajectories





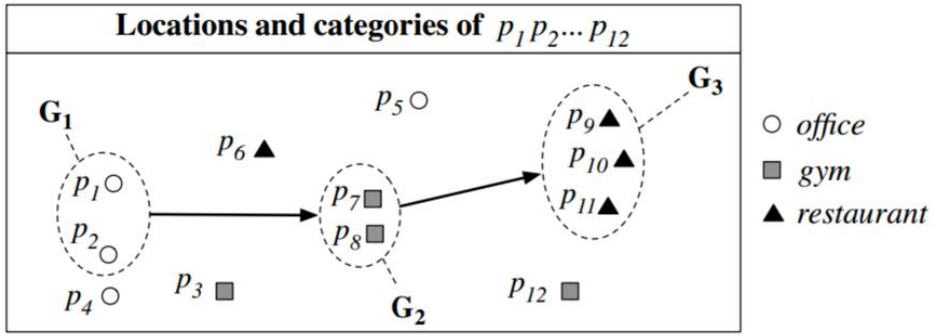
How to Define Sequential Movement Pattern?

We cannot consider each location as an independent item because of spatial continuity!

Similar places should be grouped while respecting:

- Semantic consistency
- Spatial compactness
- Temporal continuity

Example pattern: G1 -> G2 -> G3:



Problem Description

Input: a collection of N semantic trajectories, a support threshold K

Output: the sequential movement patterns that appear in no less than K trajectories

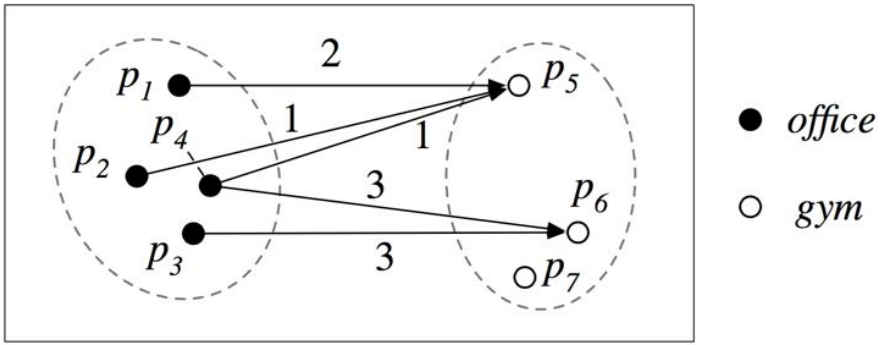
Splitter: A Top-down Mining Approach

How do we group similar places to form sequential movement patterns?

- There is an exponential number of possibilities, impossible to enumerate every option!

Key idea of Splitter:

- First mine category-level frequent transitions (coarse pattern)
- Then break each coase pattern into spatially compact ones (fine-grained pattern)



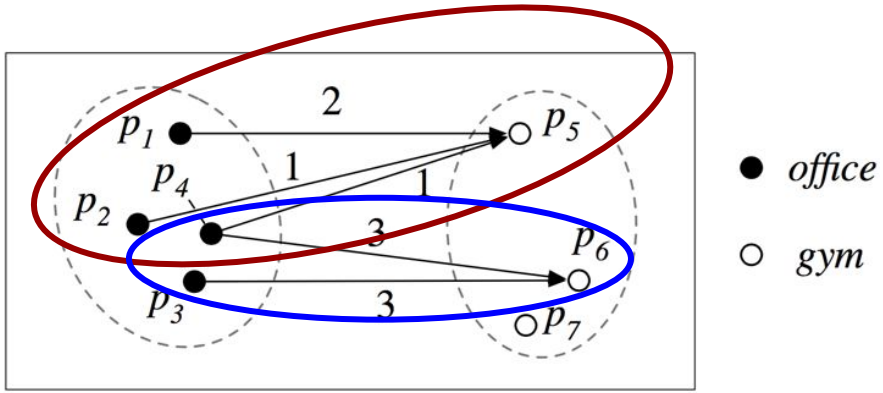
Splitter: A Top-down Mining Approach

How do we group similar places to form sequential movement patterns?

- There is an exponential number of possibilities, impossible to enumerate every option!

Key idea of Splitter:

- First mine semantics-level frequent transitions (coarse pattern)
- Then break each coase pattern into spatially compact ones (fine-grained pattern)



Splitter: A Top-down Mining Approach

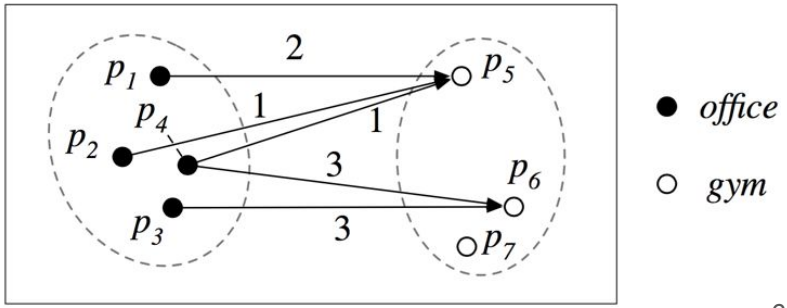
Coarse Pattern Mining: (1) Group the places with the same category; (2) Apply time-constrained sequential pattern mining

Fine-Grained Pattern Mining: (1) regard each transition as a high-D point; (2) perform iterative clustering in the high-D space to find patterns.

Object	Semantic Trajectory
o_1	$<(p_3, 0), (\boldsymbol{p_1}, 10), (\boldsymbol{p_7}, 30), (\boldsymbol{p_9}, 40)>$
o_2	$<(p_5, 0), (p_7, 30), (\boldsymbol{p_2}, 360), (\boldsymbol{p_7}, 400), (\boldsymbol{p_{10}}, 420)>$
o_3	$<(p_3, 0), (p_6, 30)>$
o_4	$<(p_2, 0), (\boldsymbol{p_1}, 120), (p_6, 140), (\boldsymbol{p_8}, 150), (\boldsymbol{p_{11}}, 180)>$
o_5	$<(p_{12}, 50), (p_8, 80), (p_{11}, 120), (p_4, 210)>$

Object	Timestamped item sequence
o_1	$\langle(G_2, 0), (G_1, 10), (G_2, 30), (G_3, 40)\rangle$
o_2	$\langle(G_1, 0), (G_2, 30), (G_1, 360), (G_2, 400), (G_3, 420)\rangle$
o_3	$\langle(G_2, 0), (G_3, 30)\rangle$
o_4	$\langle(G_1, 0), (G_1, 120), (G_3, 140), (G_2, 150), (G_3, 180)\rangle$
o_5	$\langle(G_2, 50), (G_2, 80), (G_3, 120), (G_1, 210)\rangle$

E.g., regard every length-2 transition as a 4D point: p1 -> p5, p2 -> p5, p4 -> p5, p4 -> p6, p3 -> p6.

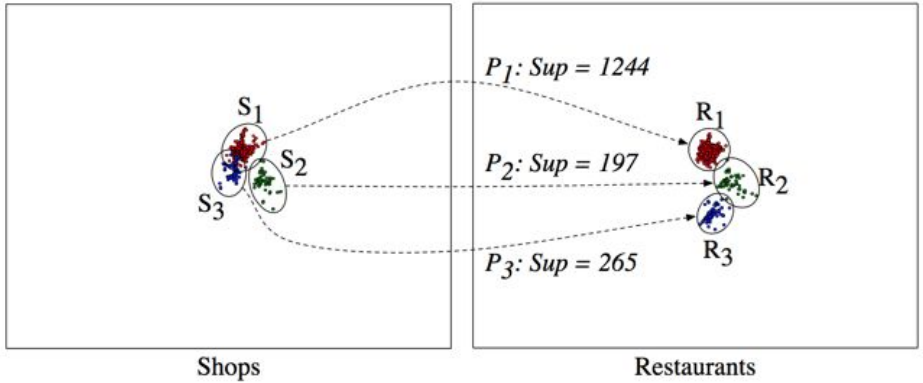


Example Patterns

Coarse Patterns

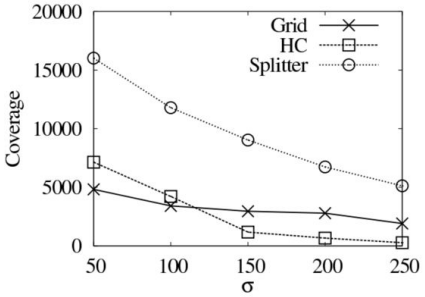
	Pattern	Sup
length=2	Shop → Food	1819
	Food → Shop	1464
	Professional → Nightlife Spot	1121
	Outdoor → Food	947
	Residence → College & University	647
length=3	Shop → Food → Shop	262
	Professional → Food → Nightlife Spot	240
	Entertainment → Food → Shop	178
	Transportation → Shop → Shop	174
	Residence → Outdoor → Food	163

Fine-grained Patterns

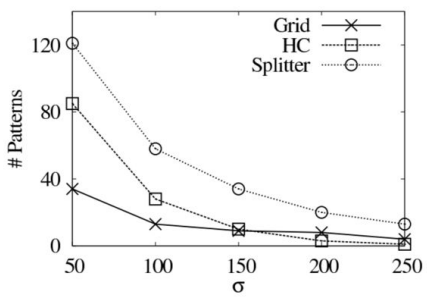


Experiments

Pattern Coverage

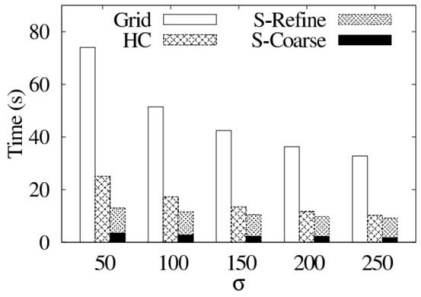


(a) Coverage *w.r.t.* σ.

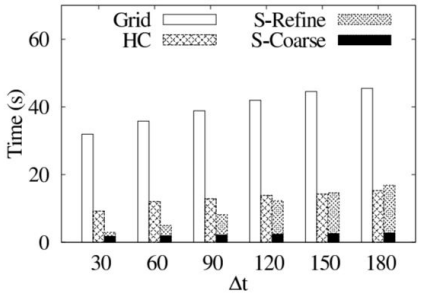


(b) Pattern number *w.r.t.* σ.

Efficiency



(a) Running time *w.r.t.* σ.

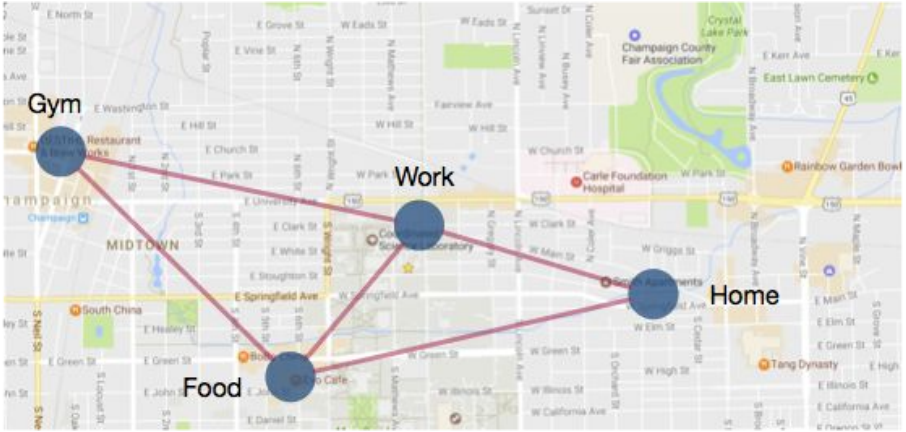


(b) Running time *w.r.t.* Δt.

GMove: Group-Level Mobility Modeling Using Geo-Tagged Social Media [Zhang et. al. KDD 2016]

Input: the semantic trajectories for a collection of users

Goal: (1) What are the intrinsic states underlying people's movements? (2) How do people move sequentially between those latent states?



Dilemma in Learning Mobility Models

Individual-level modeling: learn a model for each individual user

Each user has a limited number of records, we suffer from **data scarcity**!

Global-level modeling:

Different users have different moving behaviors, we suffer from **data inconsistency**!

GMove: Group-Level User Mobility Modeling

Idea: divide the users into coherent groups, and learn one model for each group.

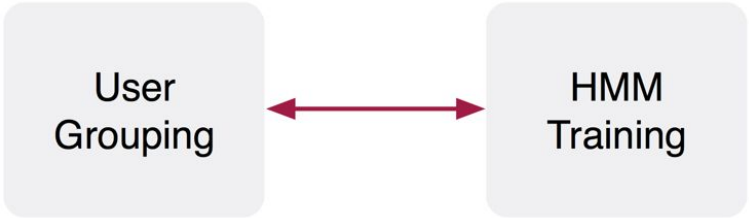
- Reduce data sparsity by aggregating the movements of multiple users.
- Ensure data consistency as the users in the same group have similar movement regularity.

	Data Sparsity	Data Consistency
Individual-level	X	O
Global-level	O	X
Group-level	O	O

HMM Ensemble Learner

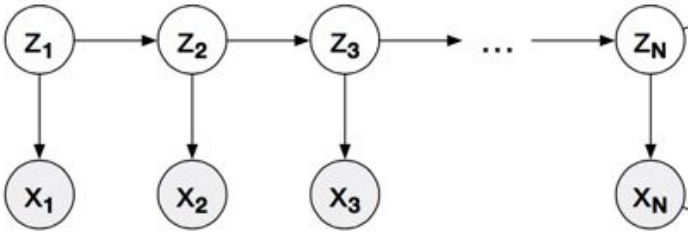
User grouping and mobility modeling mutually enhance each other:

- Better user grouping leads to more consistent training data
- Better mobility modeling helps infer the user membership more accurately

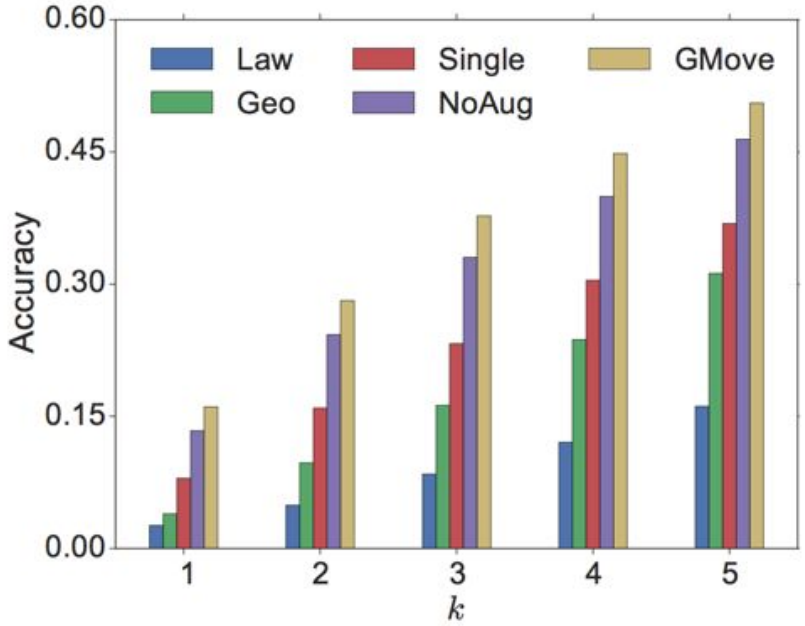


For each user u, compute the posterior probability that u belongs to group g:

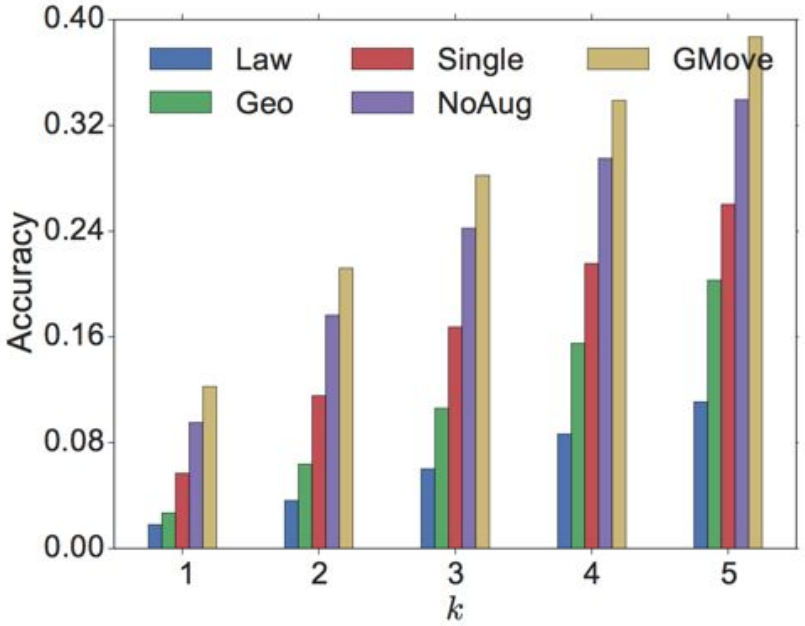
$$p(g|u;\mathcal{H}^{\text{new}}) \propto p(g)p(u|g;\mathcal{H}^{\text{new}})$$



Quantitative Evaluation: Next Location Prediction



(a) LA



(b) NY

Who, Where, When, What: User-Level Mobility Modeling [Yuan et. al. KDD 2013, TOIS 2015]

Input: a collection of checkins of different users

- Each post contains a user ID, a timestamp, a venue and a text message
- Each venue is associated with a venue ID and geo-coordinates

Output: multi-dimensional user-level mobility models

- who visits which place at what time for what activity

Previous studies: consider at most three factors out of the four

- Where What: geographical topic modeling
- Where When What: geographical event detection
- Who Where When: spatiotemporal mobility behavior modeling for users
- Who Where What: user-level geographical topic profiling

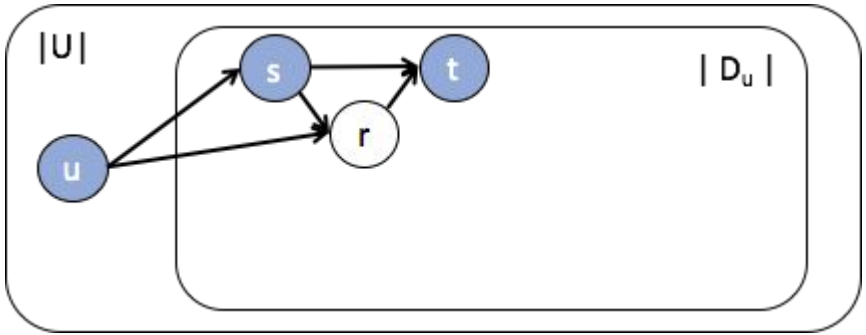
Overview

User u's mobility centers at several personal geographical regions r (home, work, …)
The region r where a user u stays is influenced by day s

- E.g., weekday: work region; weekend: shopping region

Visiting time is determined by region r and day s

- E.g., visiting shopping region at weekday evening & weekend afternoon



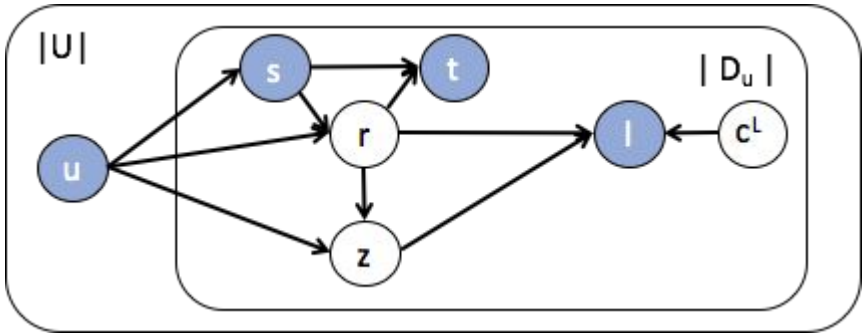
Overview

User u's topic interests is influenced by u's topic preference and region r

- E.g., u: "reading" and "shopping". u@Times Square: "shopping"

User u chooses a POI l based on either topic z or region r

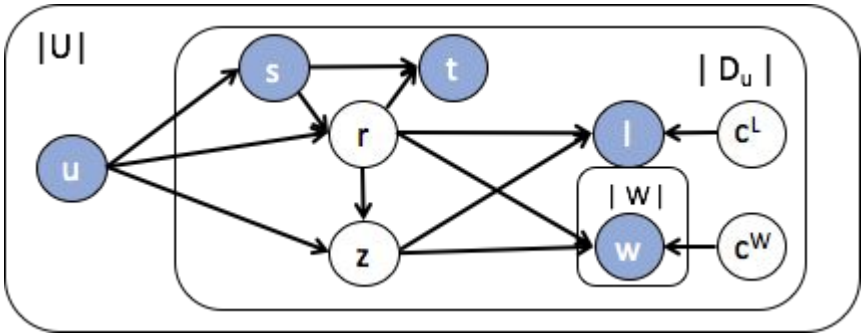
- Nearby POI within r that meets the topic requirement z (e.g., meal)
- Different users make different trade-offs between z and r



Overview

User u chooses a set of words w based on either topic z or region r

- Different user makes different trade-offs between z and r
- E.g., user u is shopping at home region: “grocery”, “family”



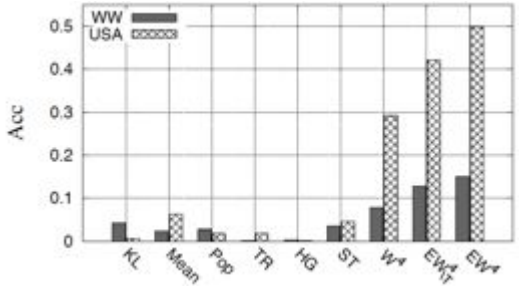
Experiments

Datasets
- 89,007 world-wide tweets (WW)
- 171,768 microblogs in USA (USA)

Venue prediction for tweet
- Rank venues by P(l|u,s,t,w)

Visitor prediction
- Rank users by P(u|s,t,l)

Venue prediction for user.
- Rank venues by P(l|u,s,t)



Acc	WW	USA
PMM	0.4163	0.4021
W^4	0.5063	0.5863
EW^4	0.5351	0.7679

Acc	WW	USA
PMM	0.0423	0.1102
W4	0.0776	0.2953
EW^4	0.1423	0.5054

Reference

Pattern-based Approaches

- Towards semantic trajectory knowledge discovery. Alvares el al. Data Mining and Knowledge Discovery 2008
- Diversified trajectory pattern ranking in geo-tagged social media. Yin et al. SDM 2011
- Mining travel patterns from geotagged photos. Zheng et al. TIST 2012
- Splitter: Mining fine-grained sequential patterns in semantic trajectories. Zhang et al. PVLDB 2014
- PRED: Periodic Region Detection for Mobility Modeling of Social Media Users. Yuan et al. WSDM 2017

Model-based Approaches

- Discovering routines from large-scale human locations using probabilistic topic models. Farrahi et al. TIST 2011
- Who, where, when and what: discover spatio-temporal topics for twitter users. Yuan et al. KDD 2013
- Constructing and comparing user mobility profiles. Chen et al. TWEB 2014
- You are where you go: Inferring demographic attributes from location check-ins. Zhong et al. WSDM 2015
- Who, where, when, and what: A nonparametric bayesian approach to context-aware recommendation and search for twitter users. Yuan et al. TOIS 2015
- Gmove: Group-level mobility modeling using geo-tagged social media. Zhang et al. KDD 2016

Reference

Application: Location Recommendation

- Personalized point-of-interest recommendation by mining users' preference transition. Liu et al. CIKM 2013
- Personalized recommendations of locally interesting venues to tourists via cross-region community matching. Zhao et al. TIST 2014
- On information coverage for location category based point-of-interest recommendation. Chen et al. AAAI 2015
- Joint modeling of user check-in behaviors for point-of-interest recommendation. Yin et al. CIKM 2015
- Experiments with a venue-centric model for personalized and time-aware venue suggestion. Deveaud et al. CIKM 2015
- Exploiting dining preference for restaurant recommendation. Zhang et al. WWW 2016
- Modeling check-in preferences with multidimensional knowledge: A minimax entropy approach. Wang et al. WSDM 2016

Application: Next Visit Prediction

- Mining user mobility features for next place prediction in location-based services. Noulas et al. ICDM 2012
- Mining geographic-temporal-semantic patterns in trajectories for location prediction. Ying et al. TIST 2013
- What's your next move: User activity prediction in location-based social networks. Ye et al. SDM 2013
- Learnext: learning to predict tourists movements. Baraglia et al. CIKM 2013
- On learning prediction models for tourists paths. Muntean et al.TIST 2015

Research Frontiers

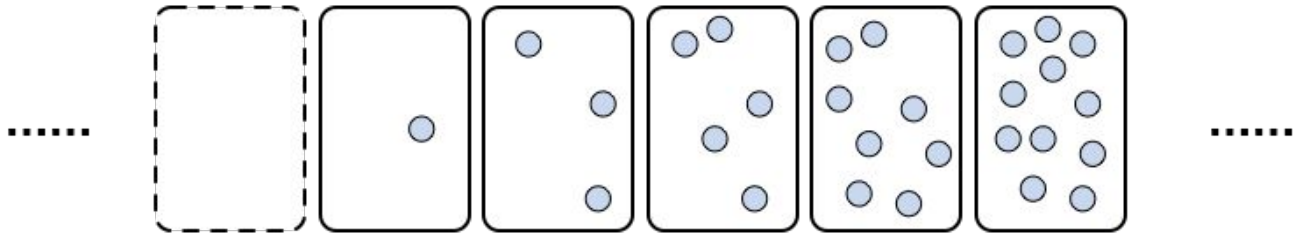
Online Learning from Streaming Data

In many scenarios, semantics-rich spatiotemporal data arrive in the stream form.

- E.g., social media, smartphone usage data

How do we design online and efficient algorithms for handling such stream data?

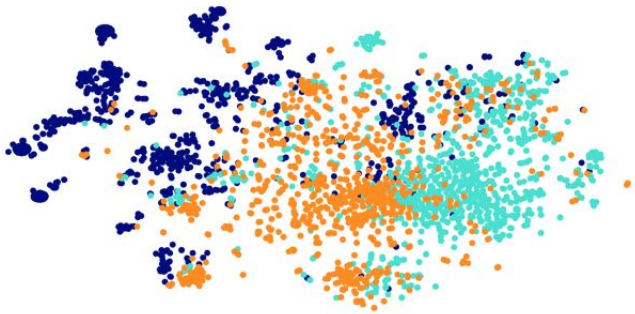
- High throughput
- Concept drift



Representation Learning for Semantics-Rich Spatiotemporal Data

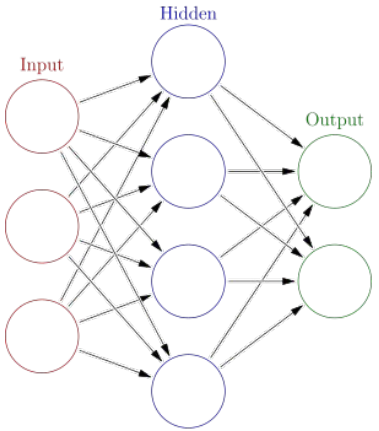
Unsupervised representation learning

- How to capture the correlations among different data types
- The learned representations are useful for downstream applications



Task-specific representation learning

- Optimize representations for the task at hand.
- E.g., location recommendation, activity classification



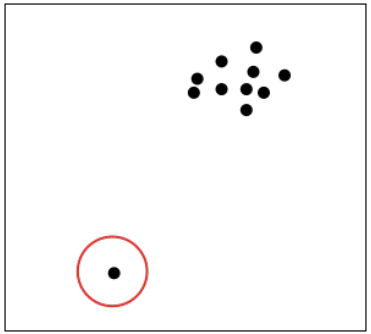
Handling Data Sparsity and Noise

Textual sparsity and noise

- How to accurately capture activity semantics from short and noisy text?

Instance sparsity and noise

- How to build reliable models if the training data is scarce?
- How to incorporate external knowledge effectively?
- How to detect noise and outliers in the input?



Freebase™

Summary

Spatiotemporal data mining is shifting from the semantics-free paradigm to the semantics-rich paradigm.

Bringing semantics into spatiotemporal data mining has large potential to improve a lot of applications.

- Location prediction, event detection, tour recommendation, etc.

Many research problems still remain open and to be explored.

- Data sparsity, representation learning, online learning

Thanks!

Slides available at: http://chaozhang.org/files/slides/slides-icde17.pdf