Querying Geo-Textual Data: Spatial Keyword Queries and Beyond

Gao Cong
Nanyang Technological University

Christian S. Jensen
Aalborg University
The Web Is (Mostly) Mobile

- A quickly evolving mobile Internet infrastructure.
  - Mobile devices, e.g., smartphones, tablets, laptops, navigation devices
  - Communication networks and users with access
- Rapidly increasing device sales (millions)
- Mobile is a mega trend.
  - Google went “mobile first” in 2010.
  - Mobile data traffic 2020 = mobile data traffic 2010 x 1000
Mobile Is Spatial

- Increasingly sophisticated technologies enable the accurate geo-positioning of mobile users.
  - GPS-based technologies
  - Positioning based on Wi-Fi and other communication networks
  - New technologies are underway (e.g., GNSSs and indoor).
Spatial Web Querying

- Total web queries
  - Google: 2011 daily average: 4.7 billion (uncertain)

- Queries with local intent
  - "cheap pizza" vs. "pizza recipe"
  - Google: ~20% of desktop queries
  - Bing: 50+% of mobile queries

- Vision: Improve web querying by exploiting accurate user and content geo-location
  - Smartphone users issue keyword-based queries
  - The queries concern websites for places

- Balance spatial proximity and textual relevance
- Support different use cases
Geospatial and Textual (Geo-Textual) Data

- Spatial Web Objects: $p = \langle \lambda, \phi \rangle$ (location, text description)

- Example:

  $$\lambda = (56.158889, 10.191667)$$

  $\phi = \text{Den Gamle By Open-Air Museum}$

  Den Gamle By - "The Old Town" – was founded in 1909 as the world's first open-air museum of urban history and culture...
Geo-textual Data – Sources

• Static geo-textual data, POI data
  ■ Web pages with location
  ■ Online business directories
    ◆ E.g., Google My Business
  ■ Location-based social networks
    ◆ E.g., 65 million POIs at Foursquare

• Streaming geo-textual data
  ■ Geo-tagged micro-blog posts
    ◆ E.g., 10 million geotagged Tweets per day
  ■ Photos with tags and geo-location in social photo sharing websites
    ◆ E.g., Flickr
  ■ Check-in information at POIs in location-based social networks (e.g., Foursquare, Facebook places)
    ◆ E.g., Foursquare had 7 million check-in on 3rd Oct 2015
Example of Streaming Geo-Textual Data

- Components of streaming geo-textual data
  - Text
  - Location
  - Time

- Example: Geo-tagged tweets
Outline

• Background – brief motivation and the data

• Spatial keyword queries on static geo-textual data
   Standard queries
   Beyond single object result granularity
   Other queries

• Querying geo-textual streams

• Exploring geo-textual data

• Summary and challenges
Outline

• Background – brief motivation and the data

• Spatial keyword queries on static geo-textual data
  ■ **Standard queries**: spatial DB queries + IR queries
    ◆ Boolean range query
    ◆ Top-
      \( k \) \( kNN \) queries
  ■ Beyond single object result granularity
  ■ Other queries

• Querying geo-textual streams
• Exploring geo-textual data

• Summary and challenges
Boolean Range Query

- A query region
- A set of keywords

Keyword: pizza

- OChre Italian Restaurant: pizza, white wine, cherry tomatoes
- Student club, Gym, badminton, snooker
- Roadlink: bikes with various brands
- Far east restaurant: spring rolls, dumplings
- Pizza hut
- Adidas, Nike sports, New Balance Sports shoes
- Somerset mall: Adidas sports accessories retail...
- Adidas retails
Top-$k$ $k$NN Query (TkQ)

- A query location
- A set of keywords
- Ranking criteria: A combination of spatial proximity and text relevancy

$k = 2$
Keywords: Adidas, sports

OChre Italian Restaurant: pizza, white wine, cherry tomatoes
Pizza hut
Student club, Gym, badminton, snooker
Roadlink: bikes with various brands
Far east restaurant: spring rolls, dumplings
Adidas, Nike sports, New Balance Sports shoes
Somerset mall: ... Adidas sports accessories retail...
Adidas retails
Top-$k$ Spatial Keyword Query

- Objects: $p = \langle \lambda, \psi \rangle$ (location, text description)
- Query: $q = \langle \lambda, \psi, k \rangle$ (location, keywords, # of objects)

- Ranking function

$$
rank_q(p) = \alpha \frac{|| q.\lambda, p.\lambda ||}{\max D} + (1 - \alpha)(1 - \frac{tr_{q,\psi}(p.\psi)}{\max P}) \\
0 \leq \alpha \leq 1
$$

- Distance: $|| q.\lambda, p.\lambda ||$
- Text relevancy: $tr_{q,\psi}(p.\psi)$
  - Probability of generating the keywords in the query from the language models of the documents

- Generalizes the $k$NN query and text retrieval
Outline

• Background – brief motivation and the data

• Spatial keyword queries on static geo-textual data
   Standard queries
   Beyond single object result granularity
    ◆ Accounting for co-location for ranking
    ◆ The m-CK query
    ◆ Aggregate queries, including collective and group queries
   Other queries

• Querying geo-textual streams
• Exploring geo-textual data

• Summary and challenges
Accounting for Co-Location for ranking

• So far, we have considered data objects as independent, but they are not.

• It is common that similar places co-locate.
  - Markets with many similar stands
  - Shopping centers, districts
  - China town, little India, little Italy, …
  - Restaurant and bar districts
  - Car dealerships

• How can we capture and take into account the apparent benefits of co-location?
Local experts are asked to provide query keywords for locations and then to evaluate the results of the resulting queries.

The studies suggest that Prestige-Based Retrieval is able to produce better results.

Cao et al. \textit{Retrieving Top-k Prestige-Based Relevant Spatial Web Objects}. VLDB11.
The mCK Query

- The m-closest keywords (mCK) query
  - A query \( q \) consists of \( m \) keywords
  - Find a group of objects \( T \) covering all \( m \) query keywords
    \[ q \subseteq \bigcup_{o \in T} o.\psi \]
  - Objects should be close to each other
    - Minimize the diameter of a group
    - Group diameter: the maximum Euclidean distance between any pair of objects
    \[ Diam(T) = \max_{o_i, o_j \in T} Dist(o_i, o_j) \]

mCK Application

- Detecting geographic locations of web resources
  - Web resource can be *documents, photos, etc.*
    - These resources are associated with tags describing the content.
    - They may be posted without geographic location.
  - We can issue an mCK query using the tags as keywords.
    - The center of the mCK result can be used to geo-tag the resources approximately.
Computing the mCK Query

• The problem is NP-hard (Guo et al. SIGMOD’15)

• Exact solutions
  ■ Index structure + effective pruning + exhaustive search (Zhang et al., ICDE 2009, ICDE 2010)
  ■ Effective pruning + exhaustive search (Guo et al. SIGMOD’15)

• Approximate solutions
  ■ Three approximation algorithms with different approximation ratios and complexity (Guo et al. SIGMOD’15)

Guo et al. Efficient algorithms for answering the m-closest keywords query. SIGMOD’15
Inspiration for Collective Queries

10 Blue Links is Dead. Blended Search Lives...
The theme of the panel is that search results containing simply 10 blue links is dead. Search engines have determined that searchers would like to use a single search box for all...
www.technologyevangelist.com/2007/12/10_blue_links_is_dea.html · Cached page

10 Blue Links from Search Marketing Gurus | Online...
In regards – “10 Blue Links”. I am trying to bridge the delta between universal search and what marketing folks can to do capitalize on these inevitable changes.
www.searchmarketinggurus.com/search_marketing_gurus/2007/06/10-blue-links.html · Cached page

10 blue links News, 10 blue links Tips | WebProNews
SEO techniques typically linger long after their “good til” dates. 2008 should be no exception, but if you’re paying attention it’s time to move onto the stuff that works.This...
www.webpronews.com/tag/10-blue-links · Cached page

Live From Yahoo’s "End of the 10 Blue Links" Talk
We’re at OutCast Communication’s offices for a Yahoo Search event that they’ve dubbed “The End of the 10 Blue Links.” It looks to be a state of the union for Yahoo’s...
By MG Siegler · 67 posts · Published 5/19/2009
techcrunch.com/2009/05/19/live-from-yahoos-end-of-the-10-blue-links-talk · Cached page

Yahoo Vows Death to the ’10 Blue Links’ - PC World...
Yahoo previewed a new way of presenting search results that could be introduced within two to three months.
www.pcworld.com/businesscenter/article/165214/yahoo_vows_death_to_the_10_blue_links.html · Cached page
Maria Sharapova

Current tournament: Roland Garros (Women's Singles)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>M. Sharapova</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Finals</td>
</tr>
<tr>
<td>21</td>
<td>S. Errani</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jun 9, Completed</td>
</tr>
<tr>
<td>2</td>
<td>M. Sharapova</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semifinals</td>
</tr>
<tr>
<td>4</td>
<td>P. Kvitova</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Jun 7, Completed</td>
</tr>
<tr>
<td>2</td>
<td>M. Sharapova</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quarterfinals</td>
</tr>
<tr>
<td>23</td>
<td>K. Kanepi</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Jun 6, Completed</td>
</tr>
<tr>
<td>2</td>
<td>M. Sharapova</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4th Round</td>
</tr>
<tr>
<td></td>
<td>K. Zakopalova</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Jun 4, Completed</td>
</tr>
</tbody>
</table>

+ Show more matches

Maria Yuryevna Sharapova is a Russian professional tennis player. As of June 11, 2012 she is ranked world no. 1. A United States resident since 1994, Sharapova has won 27 WTA singles titles, including four Grand Slam singles titles. Wikipedia

Born: April 19, 1987 (age 25), Nyagan
Height: 6'2" (1.88 m)
Weight: 130.3 lbs (59.1 kg)
Grand slams: 4
Handed: Right-handed
Parents: Yelena Sharapov, Yuri Sharapov

People also search for

Victoria Azarenka
Serena Williams
Roger Federer
Rafael Nadal
Caroline Wozniacki

News for sharapova

At times, Maria Sharapova had doubts about coming back after shoulder surgery
SB Nation - 1 day ago
After winning the 2012 French Open, Maria Sharapova met with a small group of writers from various outlets, including Sports Illustrated. Here are...

Maria Sharapova, Novak Djokovic will carry flags at London Olympics
SI.com - 2 days ago
Sharapova savours her 'sweetest triumph' as reward for comeback
The Independent - 6 days ago
Collective Spatial Keyword Querying

- The spatial aspect offers natural ways of aggregating data objects and providing aggregate query results.

- We may want to return sets of objects that collectively satisfy a query.
The Collective Spatial Keyword Query

- Query location: ⭐
- Query keywords: theater, gym
The Collective Spatial Keyword Query

- **Objects**: \( o = \langle \lambda, \psi \rangle \) (location and text description)
- **Query**: \( Q = \langle \lambda, \psi \rangle \) (location and keywords)

The result is a group of objects \( \chi \) satisfying two conditions.
- \( Q.\psi \subseteq \bigcup_{o \in \chi} o.\psi \)
- \( \text{Cost}(Q, \chi) \) is minimized

\[
\text{Cost}(Q, \chi) = \alpha C_1(Q, \chi) + (1 - \alpha) C_2(\chi)
\]
- \( C_1(\ldots) \) depends on the distances of the objects in \( \chi \) to \( Q \).
- \( C_2(.) \) characterizes the inter-object distances among objects in \( \chi \).
- \( \alpha \) balances the weights of the two components.
Collective Query Variants

- **Type 1**: cost function:
  \[ Cost(Q, \chi) = \sum_{o \in \chi} Dist(o, Q) \]
  - Application scenario
    - The user wishes to visit the places one by one while returning to the query location in-between.
    - Go to the hotel between the museum visit and the jazz concert
    - NP-hard: proof by reduction from the Weighted Set Cover problem

- **Type 2**: Cost function:
  \[ Cost(Q, \chi) = \max_{o \in \chi} Dist(o, Q) + \max_{o_i, o_j \in \chi} Dist(o_i, o_j) \]
  - Application scenario
    - Visit places without returning to the query location in-between
    - E.g., go to a movie and then dinner
    - NP-hard: proof from reduction from the 3-SAT problem
Top-\textit{k} Groups Query Illustration

- Query location: ⭐ (Kenmore Hotel, SF)
- Query keyword: Restaurant
Top-$k$ Groups Query

- **Objects:** $p = \langle \lambda, \psi \rangle$ (location, text description)
- **Query:** $q = \langle \lambda, \psi, k \rangle$ (location, keywords, # of objects)

**Ranking function**

$$\text{rank}_q(G) = \alpha \beta \text{dist}(q.\lambda, G) + (1 - \beta) \text{diam}(G) \frac{1}{\max D} + (1 - \alpha) \text{TR}_G(q.\psi, G)$$

- $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$
- **Distance:** $\text{dist}(q.\lambda, G) = \min_{o \in G} \| q.\lambda, o.\lambda \|$
- **Diameter:** $\text{diam}(G) = \max_{o_1, o_2 \in G} \| o_1.\lambda, o_2.\lambda \|$
- The text relevance function favors large groups and groups where the query keywords are distributed evenly among group objects.
- Groups are disjoint

Skovsgaard and Jensen. *Finding top-k relevant groups of spatial web objects*. VLDB J. ’15
Outline

• Background – brief motivation and the data

• Spatial keyword queries on static geo-textual data
  ■ Standard queries
  ■ Beyond single object result granularity
  ■ Other queries
    ✷ Continuous top-k queries
    ✷ Reverse kNN queries
    ✷ Why not queries
    ✷ Similarity join queries

• Querying geo-textual streams
• Exploring geo-textual data
• Summary and challenges
Continuous Spatial Keyword Queries

- **Objects:** \( p = \langle \lambda, \psi \rangle \) (location and text description)
- **Query:** \( q = \langle \lambda, \psi, k \rangle \) (location, keywords, # of objects)
- A continuous query where argument \( \lambda \) changes continuously

**Ranking function**

\[
rank_q(p) = \frac{|| q.\lambda, p.\lambda ||}{tr_{q,\psi}(p.\psi)}
\]

- Euclidean distance (changes continuously)
- Text relevancy (query dependent)
Continuous Spatial Keyword Queries

• How can we process such queries efficiently?
  ■ Server-side computation cost
  ■ Client-server communication cost

• While the argument changes continuously, the result changes only discretely.
  ■ Do computation only when the result may have changed

• Use safe zones
  ■ When the user remains within the zone, the result does not change.
  ■ The user requests a new result when about to exit the safe zone.
Reverse Spatial-Keyword Query

- For service providers
- If adding a new shop at Q, which shops will be influenced?
- Influence facts
  - Spatial distance
    - Results: D, F
  - Textual similarity
    - *Services/Products*...
    - Results: F, C

Lu et al. *Reverse Spatial and Textual k Nearest Neighbor Search*. SIGMOD, 2011
Reverse Spatial-Textual $k$-NN Query

Motivating example: Social media advertising

Objective: Find the best location and the text content for a particular advertisement so that it is displayed to the maximum number of users.

Maximizing Bichromatic Reverse Spatial-textual $k$ Nearest Neighbor (MaxBRST$k$NN) query

Limited number ($k$) of most relevant (NN) advertisements

Choudhury et al. Maximizing bichromatic reverse spatial and textual $k$-NN queries. VLDB’16

Thanks Farhana for the slide
Why Not Spatial Keyword Top-k Queries

- An initial spatial keyword top-k query
  - $Q(q.loc, q.doc, k_0, \overrightarrow{w_0})$
- A set of missing objects: $\{o_1, o_2, ..., o_m\}$

- Return a refined query
  - $q'(q.loc, q.doc, k', \overrightarrow{w'})$
    - $penalty(k', \overrightarrow{w'}) = \lambda \cdot \frac{\Delta k}{R(\overrightarrow{w_0}, q.m) - k_0} + (1 - \lambda) \cdot \frac{\Delta \overrightarrow{w}}{\sqrt{1 + |\overrightarrow{w_0}|^2}}$

- Return a refined query
  - $q'(loc, doc', k', \overrightarrow{w})$
    - $penalty(q, q') = \lambda \cdot \frac{\Delta k}{R(M, q) - k_0} + (1 - \lambda) \cdot \frac{\Delta doc}{|doc_0 \cup M.doc|}$
Why Not Spatial Keyword Top-k Queries

\[ f(q, o) = w_1 \times (1 - SDist(q, o)) + w_2 \times TSim(q, o) \]

- Top-2 “clean/comfortable” hotels near COEX \((\overrightarrow{w}_0 = < 0.5, 0.5 >)\):
  - Rank 1: Intercontinental
  - Rank 2: Oakwood
  - Rank 3: Park Hyatt (not returned)

- Refined query:
  - Use larger \(k\)? Set \(k\) to 3 or larger

Why is Park Hyatt not returned?

<table>
<thead>
<tr>
<th></th>
<th>1-SDist()</th>
<th>TSim()</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercontinental</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Oakwood</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>Park Hyatt</td>
<td>0.3</td>
<td>0.9</td>
</tr>
</tbody>
</table>

- Set \(\overrightarrow{w}_0 = < 0.3, 0.7 >\)
- Modify both \(k\) & \(\overrightarrow{w}_0\)?
Why Not Spatial Keyword Top-k Queries

- Top-2 “Clean, Comfortable” hotels near Conference Venue:
  - Rank 1: Holiday Inn
  - Rank 2: Omena Hotel
  - Rank 3: Raddison Blu (not returned)

- Refined query:
  - Use a larger $k$?
  - User other query keywords?

Set $k$ to 3 or larger

Query with “Clean, Comfortable, Luxury”

- Modify both $k$ & $q_{.doc}$?

Chen et al. Answering why-not spatial keyword top-k queries via keyword adaption. ICDE’16
Spatio-Textual Similarity Join

- Text Similarity Threshold $T_{text}$
- Spatial Distance Threshold $T_{distance}$
- Objective: Retrieve all pairs of geo-textual objects $(o_i, o_j)$ s.t.
  - (1) $\text{TextSim}(o_i, o_j) \geq T_{text}$
  - (2) $\text{Distance}(o_i, o_j) \leq T_{distance}$

---

Map of various locations and text descriptions:
- OChre Italian Restaurant: pizza, white wine, cherry tomatoes
- Student club, Gym, badminton, snooker
- Roadlink: bikes with various brands
- Far east restaurant: spring rolls, dumplings
- Pizza hut
- Adidas, Nike sports, New Balance Sports shoes
- Adidas retails
- Somerset mall: Adidas sports accessories retail...

Bouros et al. *Spatio-Textual Similarity Join*, PVLDB'12
Spatio-Textual Similarity Query

- A query region (rectangle)
- A set of keywords
- Thresholds of text similarity and spatial similarity

Retrieve the set of geo-textual regions $R$ s.t. for all $r \in R$
1) $\text{TextSim} (r, q) \geq q.T_{\text{text}}$
2) $\text{SpatialSim} (r, q) \geq q.T_{\text{spatial}}$

SpatialSim ($r, q$) is measured based on the overlap of $r$ and $q$.

Fan et al.: SEAL: Spatio-Textual Similarity Search, PVLDB 12
More Types of Queries – Examples

- **Approximate** String Search in Spatial Databases. Yao et al, ICDE’10
- Top-k Spatial Keyword Queries **on Road Networks**. Rocha-Junior and Nørvåg. EDBT’12
- **Diversified** Spatial Keyword Search On **Road Networks**. Zhang et al. EDBT’14
- Desks: **Direction-Aware** Spatial Keyword Search. Li et al. ICDE’12.
- **Distributed** Spatial Keyword Querying on Road Networks. Luo et al. EDBT’14
- **Authentication** of Moving Top-k Spatial Keyword Queries. Wu et al. TKDE’15
- **Reverse Keyword Search** for Spatio-Textual Top-k Queries in Location-Based Services. Lin et al. TKDE’16
- Keyword-Aware **Continuous kNN** Query on **Road Networks**. Zheng et al. ICDE’16
- Finding the minimum **spatial keyword cover**. Choi et al. ICDE’16
- ...
Outline

• Background – brief motivation and the data

• Spatial keyword queries on static geo-textual data
  ■ Standard queries
  ■ Beyond single object result granularity
  ■ Other queries

• Querying geo-textual streams
• Exploring geo-textual data

• Summary and challenges
Outline

• Background and motivation
• Spatial keyword queries on static geo-textual data

• Querying geo-textual streams
  ▪ Boolean publish/subscribe query
  ▪ Top-k publish/subscribe query
  ▪ Moving publish/subscribe query
  ▪ Other query

• Exploring geo-textual data
• Summary and challenges
Motivation of publish/subscribe

- Streaming geo-textual data (e.g., geo-tagged tweets) often has the quickest first-hand reports of:
  - **Breaking news**
    - E.g., Osama Bin Laden’s death\(^1\)
  - **Disasters**
    - E.g., Bomb blast in Mumbai in Nov. 2008\(^3\), flooding of Red River Valley in Mar 2009\(^2\)
  - **Public Health – Disease Outbreaks**
    - E.g., Norovirus outbreak at universities\(^3\), influenza epidemic 2009\(^3\)

---

Applications of Publish/Subscribe

• Applications
  - Location-based services, e.g., Location-aware event, Local news subscription, Location-based E-coupon
    - location-based and keyword-based requirements
    - Real-time requirement (instant feeding)
  - Annotation of Points-of-Interest (POIs) with social media feeds: Bridge dynamic (streaming) world and offline world

• Challenges:

  ➢ High arrival rate of geo-textual objects.
    • Over 10 million new tweets with coordinates per day \(^1,^2\)
    • Over 100 million new tweets with semantic locations per day \(^1,^2\)
  ➢ A large number of subscription queries.

**Boolean Range Subscription Query**

- **Boolean Range Subscription (BRS) Query**

  \[ q = (\psi, r) \]
  
  - \( \psi \): a set of keywords connected by AND or OR semantics
    
    - (dengue AND fever, vomit OR poisoning)
  
  - \( r \): the query region (within 1 km from Hyatt Regency in SF)

- To trigger the action of “pushing”, the following conditions should be satisfied:

  - The Boolean expression, as indicated in \( \psi \), should be satisfied by the object terms.
  
  - The location of object should be within the query region \( r \).
Boolean Range Subscription Query

- Problem: answering a stream of BRS queries in real time on a stream of geo-textual objects continuously.

Query for tweets containing *vomit* AND *fever* with their *distance to Hyatt Regency in SF* smaller than *1km*. 
Location-Aware Publish/Subscribe Model

Publisher

Subscription Queries Index

Subscription Query

Subscription Query

Subscription Query

Subscription Query

Subscription Query

Subscription Query

Subscription Query

index

published result

register

geo-textual object

Thanks Lisi for the slide
Boolean Publish/Subscribe Solution

• High-level framework
  - Organize subscription queries s.t. each group can be processed together
  - Indexing Subscription queries, which is also a stream
• Quad-tree + Inverted file for indexing (Chen et al.)
  - Handle Boolean Expression with AND, OR.
• R-tree: Each node records the set of keywords from descend node (Li et al.)
  - Handle Boolean Expression with AND
• AP-tree: Divide the space either by words, or by space (Wang et al)
  - Handle Boolean Expression with AND

Chen et al. *An efficient query indexing mechanism for filtering geo-textual data*. SIGMOD’13
Li et al. *Location-aware publish/subscribe*. KDD’13
Wang et al. *AP-Tree: Efficiently Support Continuous Spatial-KeyWord Queries Over Stream*, ICDE’15
What is the definition of a Subscription Query?

- An interest region + a Boolean expression
  - Such a query may receive very few (or too many) matching geo-textual objects.
  - It may be difficult for a subscriber to specify the query keywords, and especially the size of a spatial region when they are used as Boolean filters.

- Users would prefer to be updated with a few most relevant geo-textual objects in terms of distance, text similarity, and recency.
Temporal Spatial-Keyword Subscription (TaSK) Query

- **TaSK Query**
  - A set of keywords: **food poisoning vomiting**
  - Location: **Hyatt Regency in SF**
  - \( k \) - the number of results: **10**
  - **Objective:** Maintain up-to-date top-\( k \) most **relevant** results for each TaSK query over a stream of geo-textual objects.

- **How to measure ‘relevance’?**
  - \( S_{tsk} \): **Temporal spatial-keyword score**, a combination of distance (**spatial**), text relevance (**keyword**), and object freshness (**temporal**).

\[
S_{tsk} = S_{sk} (q, o) \cdot D - (t_e - o_t_e)
\]

\[
S_{sk} = \alpha \cdot S_{dist} (q, \rho, o, \rho) + (1 - \alpha) \cdot S_{rel} (q, \psi, o, \psi)
\]

- **\( S_{sk} \): Spatial-Keyword Score**
  - \( S_{dist} \): Score of spatial proximity
  - \( S_{rel} \): Score of text relevance

- **\( D^{\Delta t} \): Exponential Decaying Factor**
Main Idea

 []) For each new geo-textual object we compute its ranking score w.r.t. each query, and update the current top-kth result.
- Scores of the current top-k results changes!

How to represent, group, and index subscription queries such that queries in one group can be evaluated simultaneously to reduce the computation.

Chen et al. Temporal Spatial-Keyword Top-k Publish/Subscribe. ICDE’ 15
We use **Conditional Influence Region (CIR)** to generate **filtering conditions** for a query by considering text relevance, spatial proximity, and recency.

**Filtering condition:** When a new object \( o \) arrives at time \( t_0 \), \( o \) cannot be a top-k result if
- text relevance to \( q \) is smaller than 0.3,
- \( o \) falls outside the region.

\[
\begin{align*}
\text{TextSim} &= 1.0 \\
\text{Timestamp} &= t_0 \\
\text{TextSim} &= 0, \\
\text{Timestamp} &= t_0 + \Delta t \\
\text{TextSim} &= 0.3, \\
\text{Timestamp} &= t_0 + \Delta t \\
\text{TextSim} &= 0, \\
\text{Timestamp} &= t_0 \\
\text{TextSim} &= 0.3 \\
\text{Timestamp} &= t_0
\end{align*}
\]
Solution: Representing a TaSK Query

- The radii of CIRs of a query $q$ depend on:
  1. The temporal spatial-keyword score between $q$ and its $k$-th result ($R_{q[k]}$)
  2. The text relevance between $q$ and a new geo-textual object $o$
  3. The arrival time of $o$

- **Problem:** Infeasible to generate a CIR for $q$ and then use the region as filtering condition because the radius of the region relies on the abovementioned three aspects.

- **Solution:** Generating CIRs and deriving their corresponding filtering conditions for a TaSK query with respect to a set of spatial cells (e.g., Quad-tree cells).
How to generate CIR

If a new object $o$ falls in cell $c$, and the text relevance between $o$ and $q$ is smaller than $\minT(q, c)$, then $o$ can be filtered out.

- Given a cell, we generate a circle of radius $\minD(q, c)$.
- Given current time, and $k$th result, for each generated circle, we compute its minimum conditional text score (i.e., $\minT(q, c)$).

Other Challenge: $\minT(q, c)$ will change over time
Moving Subscription Queries

• Moving Query: Region + Boolean Expression
• Solution: Reduce communication overhead by safe region

Guo et al. Location-Aware Pub/Sub System: When Continuous Moving Queries Meet Dynamic Event Streams. SIGMOD’15
Other Work

• **SKYPE: Top-k Spatial-keyword Publish/Subscribe Over Sliding Window**, Wang et al. VLDB’16
  - Sliding Windows setting: Each geo-textual object arrives and expires
  - New techniques to prune the search space for maintaining top-k results

• A location-aware publish/subscribe framework for **parameterized** spatio-textual subscriptions. Hu et al. ICDE’15
  - Subscription queries: Defined based on both text similarity and spatial proximity. Each subscription has a pre-given threshold

• **Tornado**: A **Distributed** Spatio-Textual Stream Processing System. Mahmood et al. PVLDB 2015
Outline

• Background and motivation
• Spatial keyword queries on static geo-textual data
• Querying geo-textual streams
• Exploring geo-textual data
  ▪ Region search
    ◆ Most dense region
    ◆ Best region search
    ◆ Best region search on road network
    ◆ Interactive exploration
  ▪ Region exploration

• Summary and challenges
MaxRS problem:
- Input: a set of spatial object, and a rectangle of a given size.
- Output: the position of the rectangle such that the sum of the objects covered by the rectangle is maximized.

**In-memory solution:**
Imai and Asano, Journal of Alg. 1983
Nandy et al. Computers & Mathematics with Applications 1995

**Secondary memory solution:**
Region Search on Road Networks

Dining, NYC
Region Search on Road Networks

Shopping, NYC
Problem Formalization

- Road network graph $G$
  - A node represents a road junction point or a location, associated with a set of keywords
  - An edge represents a road segment
- Nodes are weighted w.r.t. Query
  - Relevance to the query
  - Other query-independent weights (e.g., popularity or rating) are also possible
- Query Region $R$
  - A connected subgraph of $G$
  - Shape of the region is not easily described by predefined shape and size.

Cao et al. *Retrieving regions of interest for user exploration*. VLDB’14
“Hot Region” Query

• $q = \langle \lambda, \psi, \Delta \rangle$
  - $\lambda$: a rectangular query range
  - $\psi$: keywords
  - $\Delta$: a road segment length constraint

• Retrieves the region with largest weight given the length constraint and the query range

• Example: $\lambda = \text{the whole graph}, \Delta = 6$

• Result: $<v2, v4, v5, v6>$
Region Search: Most Diversified Region

Query for a region of a given size that has the most diverse collection of attractions

- There are more POIs in region $r_1$ (5 POIs)
- There are more types of POIs in region $r_2$ (3 types)
Region Search

• Users would like to search for “best” region of a given size to explore the data
  ■ Different users prefer regions of different sizes to explore.
  ■ Different users prefer “best” regions based on different criteria
    ◆ Most dense region: SUM as aggregation function (Choi et al. VLDB02, Imai and Asano, Journal of Alg. 1983)
    ◆ More general aggregation function---submodular monotone function
Submodular function

• In the example, user specifies the size of query rectangle, and an aggregation function.
• The aggregation score function:
  - It’s submodular, with a “diminishing return” property.

Example:
Given $L(o_1) = \{a, b, c\}, L(o_2) = \{a, d\}, L(o_3) = \{c\}$, consider function $f(X) = | \cup_{o \in X} L(o) |$, we have

$$1 = f(\{o_1, o_2, o_3\}) - f(\{o_2, o_3\}) < f(\{o_1, o_3\}) - f(\{o_3\}) = 2$$

Marginal gain from adding $o_1$ to $\{o_2, o_3\}$
Marginal gain from adding $o_1$ to $\{o_3\}$

The marginal gain from adding an element to an input set decreases as the size of the input set increases
Problem definition

• Given a set of spatial objects \( O \), a submodular monotone function \( f : 2^O \rightarrow \mathbb{R} \), and the size \( a \times b \) of query rectangle,

• The best region search (BRS) problem aims at finding a location \( p \) from space \( P \)

\[
p = \arg\max_{p \in P} f(O_{r_p}^{a,b})
\]

Here

\( f(O_{r_p}^{a,b}) \) is the aggregate score of the region of size \( a \times b \) centered at \( p \)

• Challenges of solving such a general problem
  - Efficiency: Infinite points in the space to consider.
Interactive Data Exploration Using Semantic Windows

Motivation: Users perform various exploration tasks interactively.

Setting:
- A set of n-dimensional objects $S$, a grid is defined on top of $S$.
- A window is a union of adjacent cells that constitutes an n-dimensional rectangle.

Problem statement
- Conditions:
  - Content-based: aggregation of the objects inside the window
  - Shape-based: the shape of the window
- Given a set of conditions, the SW query finds all possible windows satisfying all conditions, and return online results quickly

Solution
- Divide space into grid, and enumerate windows (from cells)
- Data-driven search, based on a sample to guide the search

Looking for bright clusters of stars

Kalinin et al. *Interactive data exploration using semantic windows*. SIGMOD ‘14
Outline

• Background and motivation
• Spatial keyword queries on static geo-textual data
• Querying geo-textual streams
• Exploring geo-textual data
  ◦ Region search
  ◦ Region exploration
    ◦ Top-k Spatio-temporal term search
    ◦ Selectivity estimation
    ◦ Geo-spatial event/topic exploration
• Summary and challenges
Top-$k$ Spatio-Temporal Term Querying

• Example

Marathon runners...
Cambridge Science Festival
Preparing for marathon...
Boston marathon...
Film festival...
International Film Festival...

Word cloud

Spatio-temporal streaming posts

• Problem

Input: a top-$k$ term query with region $R$ and timespan $[t_b, t_e]$

Output: top-$k$ terms

$\begin{array}{cccc}
    t_1 & t_2 & \ldots & t_{k-g} & \ldots & t_{k-1} & t_k \\
\end{array}$

Top $k_g$ Exact results

Approximate results

Selectivity Estimation on Geo-textual Streams

- **Input**
  - A spatio-textual object stream

- **Query**
  - Query region \( q.R \)
  - A set of query keywords \( q.T \)

- **Task**
  - Estimate the number of objects that fall in query region \( q.R \) and contain the query keywords \( q.T \).

- **Solution**
  - Three different types of summary for estimation (ASP-tree, KMV, BN)

Twitterstand

- Twitterstand: Continuously acquire breaking news
  - Online clustering algorithm to generate breaking news
  - Users can specify regions of interest and topic (Boolean filter)
  - Periodically update users with new clusters of tweets

Note: The figure is from the paper.
Exploring Spatio-temporal Events

• Motivation - Zoomable event cube

Spatio-temporal streaming posts with hashtags

• Problem

Input: $k$ and a selected spatio-temporal cell with some granularity

Output: Top-$k$ events in the cell:

1. \{#BostonMarathon, #BostonMarathon2016\}
2. \{#FilmFest\}
3. \{#ScienceFestival\}

Feng et al. *STREAMCUBE: hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream*. In ICDE’15
Exploring Spatio-temporal Events

- **Step 1. Event detection using Hashtag clustering**
  - Each hashtag is represented as a **bag-of-words** and a **bag-of-hashtags** that co-occurs with the target hashtag within a document.
  - Each event is a **set of hashtags** with similar semantics.
  - Threshold hold based clustering.

- **Step 2. Event aggregation to coarser granularities**
- **Step 3. Event ranking** (popularity, Localness, burstness, etc.)

---

New hashtag: 

- #marathon
- #BostonMarathon
- #BostonMarathon16
- #Film Fest
- #Science Fest

To the closest cluster when distance < Threshold

New hashtag: 

- #iphone
- #BostonMarathon
- #BostonMarathon16
- #Film Fest
- #Science Fest

To a new cluster when distance ≥ Threshold
Geo-spatial Event Detection

Marathon runners ...
Cambridge Science Festival
Preparing for marathon...
Boston marathon...
Film festival
International Film Festival...

Research Problem:
• How to cluster tweets
• Extracting features for each cluster to determine if it is an event

Spatio-temporal streaming posts

Walther et al. Geo-spatial event detection in the Twitter stream. ECIR’ 13
The topic exploration problem:
• Given a collection of spatio-temporal documents $D$;
• **Input:** query rectangle region $R$ and timespan $[t_b, t_e]$;
• **Output:** $K$ topics in the region and timespan.

Topics on July 4th, 2015

- **topic 1**
  - independence
  - parade
  - firework

- **topic 2**
  - exercise
  - Nike
  - tired
  - jogging
  - running
  - exhaustion

- **topic 3**
  - happy
  - night
  - espresso
  - pizza
  - bar

Zhao et al. *Topic Exploration in Spatio-Temporal Document Collections.* SIGMOD’16
**Topic models**, e.g., Latent Dirichlet Allocation (Blei, 2005):  

**Model Assumptions:**
- Document $\rightarrow$ a probabilistic distribution over topics
- Topic $\rightarrow$ a probabilistic distribution over words.

**How it works:**

- Alan Turing was a pioneering English computer scientist, ... he broke German ciphers, ... Turing's parents enrolled him at St Michael's school ... he was prosecuted ...

<table>
<thead>
<tr>
<th>topics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>computer: 0.003</td>
<td></td>
</tr>
<tr>
<td>ciphers: 0.002</td>
<td></td>
</tr>
<tr>
<td>school: 0.04</td>
<td></td>
</tr>
<tr>
<td>parents: 0.025</td>
<td></td>
</tr>
<tr>
<td>prosecuted: 0.025</td>
<td></td>
</tr>
</tbody>
</table>
Challenges

**Efficiency issue:**

- Training topic models online is time consuming
  - The complexity of training LDA is $O(|W|KI)$
  - 3-months tweets (250M) in NYC $\rightarrow$ 13.85 hrs training time
  - In real scale, the number of tweets in a user specified region and timespan will be large

- User could consider to modify the query in an *exploratory manner*, and we could not train topic models for each query offline in advance.
Main idea: Organize the documents into a hierarchy structure and pre-train some models on the cells to accelerate online training.

Problem 1: Combining Algorithm with Bounded Error

Machine learning techniques + Data management principles
Summary

- The web is going mobile and has a spatial dimension.
- Many queries have local intent
- Spatial keyword queries on static geo-textual data
- Querying geo-textual streams
- Exploring/Mining geo-textual data
Next Steps

- More sophisticated ranking!
  - Which signals to use?
    - Webpage: quality of a web page, click through, diversity, etc.
    - POI: Popularity and rating, etc
  - What is the relevant context for this?
    - Dependence on location
    - Dependence on keywords
    - Dependence on search history
    - Dependence on social network
    - Dependence on time
  - How to combine them into a function (e.g., as a sum)?
  - Which weight parameters to use (e.g., a weight for each term)?
Next Steps

• **Personalized** queries (Location recommendation + Queries)

• **Queryless Search**: Which functionality to serve when?
  - Ex: mineral water, dumplings
  - How can context be used for determining user intent?

• **Evaluation**?
  - Which functionality is best where and when and for who?
  - GeoCLEF
Next steps

• Querying and mining geo-textual data streams
  ■ Storing, indexing data streams: snapshot queries, standing queries, analytics queries
    ◆ relevant trending events
    ◆ Casual relationship between events
  ■ Distributed systems for supporting querying and mining geo-textual data streams
  ■ By bridging the static geo-textual data and streaming geo-textual data, exciting opportunities for data analytics emerge

• Region search and exploration
  ■ What are interesting exploratory search and mining tasks on geo-textual data?
  ■ How to perform them efficiently?