Moving Toward Next Generation Social Network Service: a Location Based Direction

Xing Xie
Microsoft Research Asia
Nov. 7, 2008
Social Network Services

**Social Network Services (Wiki):** building online communities of people who **share** interests and activities, or who are interested in **exploring** the interests and activities of others

- MySpace, Facebook, Hi5, LinkedIn, Xiaonei (校内网), Kaixin001 (开心网)
- Among top 25 sites, 5 are social networking sites (source: Alexa)

**Traditional social network services remain rooted in the virtual world**

- Not designed for mobile
  - Can not easily access the service without a desktop PC
- Not designed for local
  - Can not share real-time and location based information
Location Based Social Network (LBSN)

By adding a location dimension, we can bring social networking back from the virtual world into real life

- The development of wireless networks and location sensing technologies have made it easier to track and share personal location information on the fly
- Allow real-life experiences to be shared in a more convenient way

**Location Based Social Network (LBSN):** social network services where people can track and share location related information with each other, via either mobile or desktop computers
Location Based Social Network (LBSN)

- Turn a mobile device into a real-time digital life recorder that enables users to share their life with others when and where it happens
- Collect recommendations and rank interesting locations
- Creating a new and more trustworthy way to search for relevant information based on the opinion of friends
- Discovery of places, people and activities via user-generated content produced by the effort of the whole community
LBSN Users

Social Networking Subscriptions by Type, World Market, Forecast: 2006 to 2013

(Source: ABI Research)
Example LBSN Services

Loopt

Whrrl

Bedo(贝多)
Research Topics

- Spatial data mining and knowledge discovery
  - Personal location/trajectory data mining
  - Spatial query log mining
  - Geo-tagged multimedia mining
  - Opinion mining for location related information
- Spatial data indexing
  - Hybrid index for spatial, textual and multimedia data
  - Index for moving objects
- Location based social network analysis
- Geographical information extraction and retrieval
- Spatial data acquisition and pre-processing
- Spatial data visualization and computer human interface
- Location privacy, data sharing and security
- Navigation and traffic prediction
- Systems, architectures and middleware
- Other related topics
Understanding User Behaviors in LBSN

Understanding user behavior is essential for providing personal Web experiences and targeted advertisements.

Current systems try to understand users only from their online behaviors:
- How people search, read and write on the Web

However, we miss a large part of their everyday life, or called ‘physical’ behaviors:
- How people dine, shop and travel
- Location can reveal a lot of information about ‘physical’ behaviors
Understanding User Trajectories
Data and Devices
Data and Devices

- 40+ GPS-enabled devices
- 70+ people participated in our data collection program


<table>
<thead>
<tr>
<th>Mode</th>
<th>Segment Number</th>
<th>Distance(km)</th>
<th>Duration(h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>2137</td>
<td>1,777.5</td>
<td>475.5</td>
</tr>
<tr>
<td>Bike</td>
<td>705</td>
<td>2,387.5</td>
<td>262.6</td>
</tr>
<tr>
<td>Bus</td>
<td>706</td>
<td>4,138.7</td>
<td>226.2</td>
</tr>
<tr>
<td>Car</td>
<td>923</td>
<td>10,169.8</td>
<td>330.0</td>
</tr>
<tr>
<td>Train</td>
<td>183</td>
<td>4,434.6</td>
<td>69.1</td>
</tr>
<tr>
<td>Airplane</td>
<td>8</td>
<td>2,1586.8</td>
<td>52.4</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>425</td>
<td>1435.0</td>
<td>90.1</td>
</tr>
<tr>
<td>Total</td>
<td>5000+</td>
<td>45,000+</td>
<td>1,500+</td>
</tr>
</tbody>
</table>
Understanding User Mobility - 1

Inferring transportation modes from GPS data

- Differentiate driving, riding a bike, taking a bus and walking

Difficulties

- Velocity-based method cannot handle this problem well (<0.5 accuracy)
- People usually transfer their transportation modes in a trip
- The observation of a mode is vulnerable to traffic condition and weather
Understanding User Mobility - 2

The 1st finding: walking is a transition between other modes
- Partition a trajectory into segments of different modes
- Handle congestion to some extent

Table I. Transition matrix among transportation modes

<table>
<thead>
<tr>
<th>Transportation modes</th>
<th>Walk</th>
<th>Driving</th>
<th>Bus</th>
<th>Bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>/</td>
<td>41.1%</td>
<td>49.0%</td>
<td>9.0%</td>
</tr>
<tr>
<td>Driving</td>
<td>99.7%</td>
<td>/</td>
<td>0%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Bus</td>
<td>98.7%</td>
<td>0.6%</td>
<td>/</td>
<td>0.6%</td>
</tr>
<tr>
<td>Bike</td>
<td>99.8%</td>
<td>0%</td>
<td>0.2%</td>
<td>/</td>
</tr>
</tbody>
</table>

- Denotes a non-Walk Point: \( P.V>V_t \) or \( P.a>a_t \)
- Denotes a possible Walk point: \( P.V<V_t \) and \( P.a<a_t \)
Understanding User Mobility - 3

The 2nd finding: many features are more discriminative than velocity

- Heading Change Rate (HCR)
- Stop Rate (SR)
- Velocity change rate (VCR)
- $>0.65$ accuracy
Understanding User Mobility - 4

Post-processing
- Transition probability between different transportation modes $P(\text{Bike}|\text{Walk})$ and $P(\text{Bike}|\text{Driving})$
- Typical user behaviors based on location
- Constrains of the real world

<table>
<thead>
<tr>
<th>Segment[i-1]: Driving</th>
<th>Segment[i]: Walk</th>
<th>Segment[i+1]: Bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\text{Driving})$: 75%</td>
<td>$P(\text{Bike})$: 40%</td>
<td>$P(\text{Bike})$: 62%</td>
</tr>
<tr>
<td>$P(\text{Bus})$: 10%</td>
<td>$P(\text{Walk})$: 30%</td>
<td>$P(\text{Walk})$: 24%</td>
</tr>
<tr>
<td>$P(\text{Bike})$: 8%</td>
<td>$P(\text{Bus})$: 20%</td>
<td>$P(\text{Bus})$: 8%</td>
</tr>
<tr>
<td>$P(\text{Walk})$: 7%</td>
<td>$P(\text{Driving})$: 10%</td>
<td>$P(\text{Driving})$: 6%</td>
</tr>
</tbody>
</table>

Ground Truth

Inference result

Transition $P(\text{Walk}|\text{Driving})$
Transition $P(\text{Bike}|\text{Walk})$

Bus stop

Diagram showing transition probabilities and constraints in the real world.
Understanding User Mobility - 5

The 3rd finding: users’ GPS logs imply road network

- Use the location constrains and typical user behaviors as probabilistic cues
- Being independent of the map information

\[ M = \{ \text{Driving, Walk, Bike, Bus} \} \]

\[ \text{E.g., } P(M_0) = P(\text{Driving}); \quad P(M_3|M_1) = P(\text{Bus} \mid \text{Walk}); \]

(1) Change points and start/end points
(2) Building Graph
(3) Spatial indexing
(4) Probability calculation

- A start or end point
- A change point

\[ P_{185}(M_i|M_j), \quad P_{854}(M_i|M_j) \]
## Understanding User Mobility - 6

<table>
<thead>
<tr>
<th>Method</th>
<th>$A_D$</th>
<th>CP/P</th>
<th>CP/R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity-based method</td>
<td>0.49</td>
<td>0.15</td>
<td>0.58</td>
</tr>
<tr>
<td>Advanced features (SR+HCR+VCR)</td>
<td>0.65</td>
<td>0.25</td>
<td>0.72</td>
</tr>
<tr>
<td>Velocity-based features + advanced features</td>
<td>0.728</td>
<td>0.27</td>
<td>0.78</td>
</tr>
<tr>
<td>$EF$ + normal post-processing</td>
<td>0.741</td>
<td>0.31</td>
<td>0.77</td>
</tr>
<tr>
<td>$EF$ + graph-based post-processing</td>
<td>0.762</td>
<td>0.34</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Mining User Similarity Based on Location History - 1

- Friend recommendation → personalized location recommendation

Motivation

- First law of the geography
- Significance of user similarity in communities
- Increasing availability of user-generated trajectories

Difficulties

- How to uniformly model users’ location histories
- How to measure user similarity
Mining User Similarity Based on Location History - 2

Location history representation
- Stay point detection
- Hierarchical clustering
- Personal graph building

Latitude, Longitude, Time
P1: Lat1, Lngt1, T1
P2: Lat2, Lngt2, T2
...  
Pn: Latn, Lngtn, Tn
Mining User Similarity Based on Location History

**Similar sequences**
- Same visiting order: \( ai == bi \)
- Similar transition time: \( |\Delta t_i - \Delta t_i'| \leq t_{th} \)

**Similarity estimation**
- The length of the matched similar sequence
- The layer of the matched similar sequence

\[
\text{seq}_1 = \langle a_1(k_1) \rightarrow a_2(k_2) \rightarrow ... a_m(k_m) \rangle,
\]
\[
\text{seq}_2 = \langle b_1(k_1') \rightarrow b_2(k_2') \rightarrow ... b_m(k_m') \rangle,
\]
Mining User Similarity Based on Location History - 4

![Graph showing MAP and nDCG@5, nDCG@10 for different methods: Hie+Seq, Seq, Hie-Count, Count, Hie-Cosine, Cosine, Hie+Pearson, Pearson.](image)
Mining Co-located User Queries
Understanding Spatial Search Queries

A typical geographic search request has two fields:

- *Query* consisting of keywords
- *Search-location* that the user search for

*Co-located queries*: queries frequently associated with close search-locations in geographic space
Example Co-located Queries

• When classifying geographic queries into
  – Point of Interest (POI) queries
    • Specific location on the map, landmark ("Summer Palace") or particular business ("MGM Grand")
  – Category queries
    • A type of location, business category ("hotel") or brand ("Starbucks")

• Co-located queries fall in
  – POI vs. POI
    • {"World Trade Center", "Wall Street"}
    • {"Children’s Museum", "Experience Music Project"}
  – POI vs. Category
    • {"Microsoft campus", "Starbucks"}
    • {"Chinese restaurant", "Jeem"}
    • {"car repair", "Toyota"}
  – Category vs. Category
    • {"shopping mall", "parking"}
    • {"night club", "restaurant"}
Co-location Interest Measures

- **Participation ratio** ($pr$) of a query $q_i$ in a query set $C=\{q_1,\ldots,q_k\}$

$$pr(C, q_i) = \frac{|L_i'|}{|L_i|}$$

- $L_i$: search-location set of $q_i$
- $L_i'$: subset of $L_i$ that are in a neighbor-set of $C$
  - **neighbor-set** $N=\{l_1,\ldots,l_k\}$ of $C$: $l_j$ is a search-location of $q_j$, and all pairs of search-locations are neighboring, i.e., $N$ forms a clique

- **Participation index** ($pi$) of a query set $C$

$$pi(C) = \min_{q_i \in C} \{ pr(C, q_i) \}$$
Example Pr and Pi Calculation

Neighbor-set

Participation ratio

Participation index
Co-located Query Pattern

- **Co-located query pattern**: a query set $C$ with participation index no less than a given participation index threshold.

<table>
<thead>
<tr>
<th>A, D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_2, D_1</td>
</tr>
<tr>
<td>A_3, D_3</td>
</tr>
<tr>
<td>A_4, D_4</td>
</tr>
<tr>
<td>A_4, D_5</td>
</tr>
<tr>
<td>A_5, D_6</td>
</tr>
<tr>
<td>A_5, D_7</td>
</tr>
<tr>
<td>A_6, D_6</td>
</tr>
<tr>
<td>A_7, D_9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A, C, D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_2, C_2, D_1</td>
</tr>
<tr>
<td>A_4, C_4, D_4</td>
</tr>
<tr>
<td>A_4, C_4, D_5</td>
</tr>
</tbody>
</table>

- $2/7, 2/7, 3/9$

- $< 0.5$, not a pattern

| 6/7, 7/9 |

- $> 0.5$, co-located pattern
Problem of the Basic Approach

• Only find **global patterns** that are globally co-located in the entire search space
  – Many global patterns are common knowledge

• Queries tend to be co-located in specific regions, i.e., **local patterns**, which are more interesting, are lost
  – Example local patterns
    • {“Children’s Museum”, “Experience Music Project”}
    • {“hotel”, “Ilikai hotel”}
    • {“hotel”, “MGM Grand”}
Locality Degree of a Pattern

- **Locality degree** $H(C)$ of a pattern $C$:
  \[
  H(C) = - \sum_{\forall R_j \in S} \frac{|N_j|}{|N|} \log \left( \frac{|N_j|}{|N|} \right)
  \]
  - $N=\{N_1,...,N_t\}$: neighbor-set collection of $C$
  - $N_j$: the neighbor-set collection of $C$ in region $R_j$
  - $\frac{|N_j|}{|N|}$ indicates probability of $C$ occurs in region $R_j$

- High $H(C)$: $C$ appears evenly in all regions
- Low $H(C)$: $C$ sticks to a few specific regions
- **Local pattern**: a co-located pattern with $H(C) < a$ threshold
- **Global pattern**, otherwise
Lattice based Co-located Query Pattern Discovery

Geographic search log database: \( S, Q, L \)

Co-located query patterns in regions:
\[ P' = \{ P_1', \ldots, P_t' \} \]

Search-location sub-sets in regions:
\[ L = \{ G_1, \ldots, G_s \} \]

Pattern locality identification

Co-located query patterns in regions:
\[ P = \{ P_1, \ldots, P_t \} \]

Locality scores of patterns:
\[ H = \{ C_1, h_1 \}, \{ C_2, h_2 \}, \ldots \]
Data Set

- Two weeks’ search log from Live Local Search in August 2007

Data cleaning
- Identify query sessions
- Remove queries only containing place names
- Remove noisy search-locations
  - “City center” noise
  - “Implicit search” noise

After data cleaning, about 1 million search requests remain
Number of Patterns

<table>
<thead>
<tr>
<th>Basic approach</th>
<th>132</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell width (km)</td>
<td>Lattice based approach</td>
</tr>
<tr>
<td>5</td>
<td>4299</td>
</tr>
<tr>
<td>10</td>
<td>4182</td>
</tr>
<tr>
<td>20</td>
<td>4081</td>
</tr>
<tr>
<td>40</td>
<td>4031</td>
</tr>
<tr>
<td>80</td>
<td>3662</td>
</tr>
</tbody>
</table>
User Study

- Randomly pick 100 patterns discovered by the two approaches, respectively
- Asked 5 assessors to label the quality and the locality of each pattern
  - Three levels of quality
    - Geographically related (1.0)
    - Somewhat geographically related (0.5)
    - Not geographically related (0)
  - Two levels of locality
    - local (1)
    - global (0)
Quality & Locality Evaluation

- Patterns discovered by the lattice based method achieves higher quality scores in average

<table>
<thead>
<tr>
<th>Approach</th>
<th>Number of patterns with 3/5 agreement</th>
<th>Overall scores</th>
<th>Average score per pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>87</td>
<td>66</td>
<td>0.76</td>
</tr>
<tr>
<td>Lattice</td>
<td>93</td>
<td>83.5</td>
<td>0.90</td>
</tr>
</tbody>
</table>

- Lattice based method discovers more local patterns

<table>
<thead>
<tr>
<th>Approach</th>
<th>Number of patterns with 3/5 agreement</th>
<th>Number of local patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>100</td>
<td>64</td>
</tr>
<tr>
<td>Lattice</td>
<td>100</td>
<td>94</td>
</tr>
</tbody>
</table>
Conclusions

- Location based social networking is becoming popular
- Understanding user behavior from a geospatial perspective
  - Transportation mode classification
  - User recommendation based on GPS trajectories
  - Mining co-location patterns in spatial query logs

Future work

- More types of data: taxi data, cell-tower data, game logs, etc.
- More types of user interests: life pattern, social pattern, etc.
- Privacy issues
Thanks!