GeoEcho: Inferring User Interests from Geotag Reports in Network Traffic

Ning Xia (Northwestern University)
Stanislav Miskovic (Narus Inc.)
Mario Baldi (Narus Inc.)
Aleksandar Kuzmanovic (Northwestern University)
Antonio Nucci (Narus Inc.)
Background

Geotag: lat/long pair

<table>
<thead>
<tr>
<th>Host</th>
<th>HTTP requests</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.google.com">www.google.com</a></td>
<td>...S&amp;ll=44.xxxxxx, -69.xxxxxx&amp;…</td>
</tr>
<tr>
<td>api.twitter.com</td>
<td>...lat=39.xxxxxxx&amp; long=-91.xxxxxxx…</td>
</tr>
<tr>
<td>a.medialytics.com</td>
<td>...&amp;lat=33.xx&amp;lon=-78.xx&amp;d=HTC+…</td>
</tr>
</tbody>
</table>

Each application has its own geotags
Motivation

• Can we collect all geotags for a single user across applications?
• What do the geotags we see actually mean?
• What can we learn about each user from their reported geotags?

• CSP can see all geotags from different applications for the same user
• A large volume of geotags can be captured from user traffic, but not all of them are user locations
• From user locations, we can learn users’ real-world activities
Motivation (Cont.)

GeoEcho is designed to:

• Be fully passive and service-agnostic
• Learn users’ real-world interests from geotags
• Be utilized by traffic observers such as CSPs
• Enable better personalized services

GeoEcho analyzes user geotags to connect user online traffic to offline activities, which will enable CSPs to provide better services.
**Dataset**

- **Summary of datasets**

<table>
<thead>
<tr>
<th>Trace duration</th>
<th>2 weeks in summer 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>United States</td>
</tr>
<tr>
<td>Total user number</td>
<td>608,788</td>
</tr>
<tr>
<td>HTTP sessions with geotag</td>
<td>27,981,407</td>
</tr>
<tr>
<td>Base stations with known Coordinate</td>
<td>3,415</td>
</tr>
</tbody>
</table>

- **Point of Interest (PoI)**
  - Used to present user interests
  - Information from foursquare API
  - 8 categories and 400 subcategories

<table>
<thead>
<tr>
<th>Pol Categorises</th>
<th># of Pol subcategory</th>
<th>Subcategory examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art &amp; entertainment</td>
<td>41</td>
<td>Art gallery, casino...</td>
</tr>
<tr>
<td>College &amp; university</td>
<td>38</td>
<td>College gym, college stadium...</td>
</tr>
<tr>
<td>food</td>
<td>87</td>
<td>Coffee shop, Chinese restaurant...</td>
</tr>
<tr>
<td>Nightlife spots</td>
<td>18</td>
<td>Bar, night club...</td>
</tr>
<tr>
<td>Outdoors</td>
<td>46</td>
<td>Beach, ski area...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Methodology

Mobile traffic
(2 week from a CSP)

Geotag Extraction

Geotag Record

Geotag Position Identification

Trustable Host Identification

Trustable Seeds

User Locations

Geotag Preprocessing

Interest Vector Calculation

User Location Interest Vector

Analysis

Interest Analysis

PoI Inference

User Location Identification

Geotag Discovery & Extraction

Methodology
Geotag Extraction

- Raw geotag extraction from HTTP requests:
  - 2,500 keyword based geo-signature:
    - Hostname
    - Keywords
    - Regular expression
  - 2,246 individual hosts
  - 27,981,407 geotags from HTTP sessions

The extracted geotags may not be user locations.
User Location Identification

How to identify user locations from reported geotags?

- Geo-trustable hosts
  - HTTP hostnames that only collect user locations
  - Identified by the nearby base stations

Before location identification

After location identification
Geotag Characteristics

- Fine-grained or coarse-grained

<table>
<thead>
<tr>
<th>Geotag types</th>
<th>Digits after point</th>
<th>Coverage in meters</th>
<th>% of total geotags</th>
</tr>
</thead>
<tbody>
<tr>
<td>coarse-grained</td>
<td>1</td>
<td>10,000m*10,000m</td>
<td>0.25%</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,000m*1,000m</td>
<td>40.75%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>100m*100m</td>
<td>0.17%</td>
</tr>
<tr>
<td>fine-grained</td>
<td>4</td>
<td>10m*10m</td>
<td>0.15%</td>
</tr>
<tr>
<td></td>
<td>5+</td>
<td>1m*1m</td>
<td>58.68%</td>
</tr>
</tbody>
</table>

- Regular and bursty

  Bursty because of frequent reposts

  Regular geotag reports because of apps like weathers
Inferring User Interests

- User PoI Vector Calculation
  - Geotag Preprocessing:
    - Remove the geotag biases:
      - Temporal aspects
      - Locality aspects
  - Candidate PoI Selection
    - Select nearby PoIs for each geotag
      - Nearer PoIs have better chance

Pol vector calculation formalizes the PoI selection
Inferring User Interests

- Geotag Preprocessing

Geotag Biases
- Geotag are not regular in time
- More geotags around home or work place
- Coarse-grained geotags will cover too many PoIs

- **Group geotags into hours**: the same geotag will be considered once within each hour
- **Remove home and work places**: 30.7% geotags removed
- **Refine coarse-grained geotags**: coarse-grained geotags are replaced by inside fine-grained geotags
Inferring User Interests

- Candidate PoI Selection

**Fine-grained geotags:**
- Different PoI search radii
- $r_1$ (20m) < $r_2$ (50m)

**Coarse-grained geotags:**
- About 500m*500m coverage
- Consider all covered PoI

All selected PoIs from the same geotag are considered with equal user interest.
Inferring User Interests

User Interest Vector Calculation

- Calculate user interest vectors on different time scales (daily, month, etc.)
- Normalize the selected PoIs into vectors to enable comparison between different different users.

<table>
<thead>
<tr>
<th>PoI Category</th>
<th>PoI Subcategory</th>
<th>Interest Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>food</td>
<td>coffee_shop</td>
<td>0.05</td>
</tr>
<tr>
<td>food</td>
<td>chinese_restaurant</td>
<td>0.15</td>
</tr>
<tr>
<td>college</td>
<td>gym</td>
<td>0.25</td>
</tr>
<tr>
<td>college</td>
<td>stadium</td>
<td>0.2</td>
</tr>
<tr>
<td>college</td>
<td>library</td>
<td>0.3</td>
</tr>
<tr>
<td>nightlife</td>
<td>bar</td>
<td>0.05</td>
</tr>
</tbody>
</table>

User interest vector calculation formalizes the user interests from the user PoI vector for further analysis/comparison.
User Interests Analysis

With User Interest Vectors:

• Can we learn how many PoIs are interested in?
• Can we predict user movement by different time?
• Can we group different users with similar interests?

With user interest vectors, traffic observes such as CSPs can learn many details of end users and are possible to provide better services like recommendations and advertising.
User Interests Analysis

- **User Interest Vectors:**
  - PoIs can be used to present user real-world interests

The cardinality of user interest vectors is small (among 400 of them)
User Interests Analysis

- User Interest Patterns:

User interest vector can be calculated on different time duration (daily/monthly/yearly) to learn user interest patterns.
User Interests Analysis

- User Interest Uniqueness

The user interest vectors are largely unique

Similarity of PoI interests from 100 random users
Summary and Conclusions

Methodology:
- Extract user coordinates to get user locations
- Define and calculate user interest vectors
- Connect online traffic to \textit{offline physical activities}

Geotag characteristics
- Noisy, irregular and bursty

User interests:
- Cardinality is small
- User interests are largely unique

GeoEcho will generate formalized user interest vectors, which can be calculated on different time duration. CSPs can use such interest vectors to provide better personalized services, such as advertising, recommendation, etc.
GeoEcho: Inferring User Interests from Geotag Reports in Network Traffic

Thanks!