Understanding User Behavior From Online Traces

ELAD KRAVI

ADVISORS
YARON KANZA
BENNY KIMELFELD
The Data Revolution

People share large amount of data
  ◦ Explicitly and implicitly
  ◦ Attributes collected including
    ◦ locations, timestamps, textual content etc.

A great opportunity to *improve* online services, to *enhance* existing infrastructure and to *engage* users
Goals

Leverage analysis of online traces for
- Improving measurement of *users’ similarity*
- Enhancing *online services*
- Engaging *online activity*

What affects users online behavior?
- Do people have *different needs* in *different places*?
- How do *social relationships* affect online behavior?
Outline

Location and text effect
Social networking effect
## Location and Text Effect

<table>
<thead>
<tr>
<th>Location</th>
<th>Textual</th>
<th>Location + Textual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>The <em>City Nexus</em> tool [SIGSPATIAL 2014]</td>
<td>Multi-Clicked Queries [under review]</td>
</tr>
</tbody>
</table>
Does the familiarity of environment matter?

- Pizza dough?
- Pizza place?
A note on the dataset

Our dataset included more than a billion queries log traces of a popular commercial search engine. Using these traces one can calculate the rank of query auto-completion completion terms.

<table>
<thead>
<tr>
<th>Query</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>”pizza dough”</td>
<td>5</td>
</tr>
<tr>
<td>”pizza place”</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>auto-completion</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>pizza</td>
<td>dough</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>place</td>
<td>2</td>
</tr>
</tbody>
</table>
Hypothesis: Information need is affected by *familiarity* of the environment

<table>
<thead>
<tr>
<th>Category</th>
<th>Familiar</th>
<th>Unfamiliar</th>
</tr>
</thead>
<tbody>
<tr>
<td>pizza</td>
<td>dough:3, places:5</td>
<td>places:3, dough:5</td>
</tr>
<tr>
<td>gas</td>
<td>fireplace:3, station:6</td>
<td>station:3, fireplace:8</td>
</tr>
<tr>
<td>wild</td>
<td>rice:2, horse:5</td>
<td>horse:2, rice:4</td>
</tr>
</tbody>
</table>
What is a *location*?

- Using the IP address

What is a *familiar* location?

- Significance
- Travels

How to *verify* that the model works?
A place $p$ is familiar to a user $u$ if:

- $\%$ of days on which $u$ was active from $p \geq t$
- $p$ is significant to $u$

**OR**

- $u$ returned to $p$ at least $r$ times
- All $u$’s activities are in $p$

**AND**

re-occurrence
e.g. desktop
For 53.9% of the users, the distance from declared home was smaller than 20 KMs.

For 75.4% of the users it was smaller than 100 kilometers.
Queries Per Unique Users (QPU)

A place having **large QPU** (many queries few users) is expected to obtain **many familiar queries**

- and vice versa
Queries Per Unique Users (QPU)

A clear correlation is observed
Compare To A Baseline Model

Consider the following baseline:

- A *familiar* location is *every place around 20 KM from declared home*

---

**Our Model**

A clear correlation can be seen

**20 KM Model**

No correlation can be seen
# Difference in Language Models

<table>
<thead>
<tr>
<th>Uni-grams</th>
<th>Bi-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F-Search</strong></td>
<td><strong>U-Search</strong></td>
</tr>
<tr>
<td>facebook</td>
<td>google</td>
</tr>
<tr>
<td>sale</td>
<td>restaurant</td>
</tr>
<tr>
<td>free</td>
<td>schedule</td>
</tr>
<tr>
<td>games</td>
<td>football</td>
</tr>
<tr>
<td>ebay</td>
<td>ny</td>
</tr>
<tr>
<td>how</td>
<td>lyrics</td>
</tr>
<tr>
<td>login</td>
<td>ct</td>
</tr>
<tr>
<td>online</td>
<td>store</td>
</tr>
<tr>
<td>craigslist</td>
<td>movie</td>
</tr>
<tr>
<td>recipes</td>
<td>hours</td>
</tr>
<tr>
<td>porn</td>
<td>locations</td>
</tr>
<tr>
<td>tube</td>
<td>mall</td>
</tr>
<tr>
<td>Location</td>
<td>The City Nexus tool [SIGSPATIAL 2014]</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Textual</td>
<td>Multi-Clicked Queries [under review]</td>
</tr>
<tr>
<td>Location + Textual</td>
<td>Familiarity of environment [SIGIR 2015]</td>
</tr>
<tr>
<td></td>
<td>Correlation Between Textual Content and Geospatial Locations [GeoRich 2014]</td>
</tr>
</tbody>
</table>
Posts Origin in Microblogs

Correlation?
Application #1: Associating Posts from Different Networks

Text-based social network

Location-based social network

Who will have a greater success, Alice or Bob?
An Example – Measuring User’s Similarity

We compared between similarity based on the following measures:

◦ only the *locations* of the messages using *nearest neighbor distance*
◦ only the *content* of the messages using *TF-IDF*
◦ combination of *both*
Identification Test
Problem Definition – Identification test

A post $p$ is denoted by $p = (l, c)$
- $l$ – location, on sphere
- $c$ – textual content

Each user $u$ is associated with the set $p_u$ of her posts

- Split $u$ into $u_1, u_2$, such that $p_{u_1} \cup p_{u_2} = p_u$ and $p_{u_1} \cap p_{u_2} = \emptyset$
- Let $K$ be the $k$-most-similar users to $u_1$ among $U \cup u_2$
- Consider success as the case where $u_2 \in K$ and failure otherwise

Goal – maximize success rate
Content is better than locations and the combination of the two provides the best results.
Outline

Location and text effect

Social networking effect
Social Networks

Recommending content items to community owners [SIGIR 2014]
- Using recommender-system approach to recommend content items to owners of online communities in a corporate social network

Measuring the effect on activity level [TOCHI 2015]
- Further extending previous work to examine the effect of recommendation over the activity in the communities
Main Challenges

**Formal modeling** that allows automatic detection
- avoiding detection of erroneous patterns
- yet, portraying the diversity of human behavior

**Verifying** a proposed **model**
- lack of ground truth and tagged data
Summary

We examined the utilization of spatial, textual and social information toward understanding online behavior

- **Spatial and Textual:** jointly-visited locations, multi-clicks, familiarity of environment and similarity between users
- **Social:** recommending content items to community owners, engaging community’s activity

Leveraging online data one can improve measurement of users’ similarity, enhance online services and engage online activity
Future Work

Building tools for finding complex patterns that combine traces from different datasets

- For example - improving web search by using social activity

Developing infrastructure for allowing users to define their models of online behavior patterns, and later on detecting these patterns on real datasets
Thank You!