Introduction to the Social Web
Content Search, Recommendation and Exploration

Sihem Amer-Yahia
CNRS/LIG
Nov 19th, 2014
Course Outline

• Nov 19th, 2014
  – Brief history of Recommendations: 1 hour
  – Blending Search and Recommendation: 2 hours
    • Hotlists Recommendation on Delicious
    • Top-k Algorithms
    • Network-aware Search on Collaborative Tagging Sites

• Nov 26th, 2014
  – User Studies on the Social Web: 1hour
    • Group Recommendation
  – Social Content Exploration: 2 hours
Top-

k Algorithms with Applications to Collaborative Tagging Sites

- Hotlist generation in *Delicious*
- Top-
k processing
- Network-aware search
The most popular bookmarks on Delicious right now

17 Best Free Online Fax Services
via savedelete.com

10 Interesting CSS3 Experiments and Demos
via sixrevisions.com

If the Earth Stood Still
via www.esri.com

Introduction to MySQL Triggers | Nettuts+
via net.tutsplus.com
Leveraging Tagging for Hotlist Generation

• Model current user interests based on social behavior: tagging and explicit social ties

• Apply to hotlist generation in del.icio.us
  – Hotlist -- simplest form of recommendation, a list of popular items presented to user
  – del.icio.us -- social tagging site, a kind of a social content site

• Specific contributions
  – Formalize the problem
  – Propose evaluation metrics
  – Propose and evaluate several hotlist generation methods
Data Model

- users $u \in U$, tags $t \in T$, items $i \in I$
- friends($u$) directional

- tags($u$)
- items($u$) & items($u$, $t$)

- taggers($i$) & taggers($i$, $t$)
del.icio.us Hotlists Experiment

- 116,177 del.icio.us users
  - who tagged 175,691 distinct URLs
  - using 903 tags
  - for a total of 2,322,458 tagging actions
  - for 1 month in 2006

- Evaluate how networks predict user’s interest

A/B testing: user behavior in first 3 weeks to predict 4th week
Tagging data has a long tail

- we have to clean it for efficiency (relational processing)
- we removed unpopular tags (< 4 uses) & URLs (< 10 uses), reduced to 27% of original size
Global

10 URLs that are tagged most often overall

Performance

coverage (global) = 3%
scope (global) = 100%

Global Top-10

<table>
<thead>
<tr>
<th>Rank</th>
<th>URL</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>google.com</td>
<td>980</td>
</tr>
<tr>
<td>2</td>
<td>facebook.com</td>
<td>820</td>
</tr>
<tr>
<td>3</td>
<td>iTunes.com</td>
<td>729</td>
</tr>
<tr>
<td>4</td>
<td>twitter.com</td>
<td>720</td>
</tr>
<tr>
<td>5</td>
<td>jonasbrothers.com</td>
<td>680</td>
</tr>
<tr>
<td>6</td>
<td>cnn.com</td>
<td>678</td>
</tr>
<tr>
<td>7</td>
<td>amazon.com</td>
<td>620</td>
</tr>
<tr>
<td>8</td>
<td>yahoo.com</td>
<td>525</td>
</tr>
<tr>
<td>9</td>
<td>youtube.com</td>
<td>524</td>
</tr>
<tr>
<td>10</td>
<td>techcrunch.com</td>
<td>492</td>
</tr>
</tbody>
</table>

Items (Chris)

<table>
<thead>
<tr>
<th>URL</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>jars.com</td>
<td>java</td>
</tr>
<tr>
<td>java.sun.com</td>
<td>java</td>
</tr>
<tr>
<td>techcrunch.com</td>
<td>news</td>
</tr>
<tr>
<td>devshed.com</td>
<td>tutorial</td>
</tr>
</tbody>
</table>

Items (Ben)

<table>
<thead>
<tr>
<th>URL</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>bbc.co.uk</td>
<td>news</td>
</tr>
<tr>
<td>pbs.org</td>
<td>news</td>
</tr>
<tr>
<td>tomwaits.com</td>
<td>music</td>
</tr>
<tr>
<td>nick-cave.com</td>
<td>music</td>
</tr>
<tr>
<td>loureed.com</td>
<td>music</td>
</tr>
</tbody>
</table>
If a user tags with *sports*, he is interested in sports-related content

\[
\text{interest}(u,t) = \frac{|\text{items}(u,t)|}{|\text{items}(u)|}
\]

Build one global hotlist per tag, use in one of two ways

- **best_tag**
  
  hotlist = top-10 for tag for which user has highest interest

- **dominant_tags**
  
  hotlist is a combination of up to 3 top-10 lists s.t. \( \text{interest}(u,t) \geq 0.3 \) (user has strong interest for these tags)
Performance of Tag-based

best_tag

coverage = 9%
scope = 100%

dominant_tags

1 tag   coverage = 10%   scope = 32%
2 tags  coverage = 14%   scope = 14%
3 tags  coverage = 18%   scope =  6%
Network-based

Choose 10 most popular URLs from those tagged by a user’s friends.

coverage (friends) = 43%
scope (friends) = 31%
Common Interest Networks: URL-interest

*Identify the seed -- a set of users who tag many of the same URLs as the user u ("agree with u").* Hotlist = 10 most popular URLs tagged by users in seed.

\[ agr(u,f) = \frac{|\text{items}(u) \cap \text{items}(f)|}{|\text{items}(u)|} \]

\[ U_{\text{scope}} = \{ u \in U \mid \exists f \in U, \text{agr}(u,f) > \text{threshold} \} \]

\[ U_{\text{seed}} = \{ f \in U \mid \text{agr}(u,f) > \text{threshold} \} \]

\[ \text{thresh} = 0.3 \quad \text{coverage} = 61\% \quad \text{scope} = 1.2\% \]

\[ \text{thresh} = 0.5 \quad \text{coverage} = 71\% \quad \text{scope} = 0.7\% \]
Agreement across the board is rare, let’s look at agreement per-tag: may agree with adviser on research, but with mom on cooking.

\[ \text{agr} (u,f,t) = \frac{|\text{items}(u,t) \cap \text{items}(f,t)|}{|\text{items}(u,t)|} \]

\( U_{\text{scope}}, U_{\text{scope}} \) defined as for url-interest, combined as in dominant-tags.

\( \text{scope} (\text{tag-url-interest}) = 7\% \)
### Tag/Interest-based Methods: a Comparison

Users in the intersection of **dominant-tags**, **url-interest** and **tag-url-interest**, with a strong interest in 2 tags, all thresholds = 0.3

|                | $|U_{\text{scope}}|$ | avg ($|U_{\text{seed}}|$) | coverage |
|----------------|----------------------|--------------------------|----------|
| dominant-tags  | 1235                 | 26,856                   | 17%      |
| tag-url-interest| 1235                 | 227                      | 82%      |
| url-interest   | 205                  | 203                      | 85%      |