Recommending Interesting Activity-Related Local Entities

Jie Tang, University of California at Berkeley, jietang@eecs.berkeley.edu
Ryen W. White and Peter Bailey, Microsoft, {ryenw, pbailey}@microsoft.com

Local Entities

- Real world attractions or businesses
- Recommended for local queries
- Proximity insufficient to find entities related to location of interest
- **Activity-related** entities determined based on current activity
  - Relationship not commutative
  - Shopping $\rightarrow$ Bar, does not mean Bar $\rightarrow$ Shopping
- Nature of entity determines willingness to travel
  - Willing to travel far for major tourist attractions, but not substitutable entities like bars and cafés

Identifying **Activity-Related** Local Entities

Data Pre-processing
- Find URLs for location of interest e.g., pikeplacemarket.org for Seattle
- Extract SERP clicks from queries for Seattle-based addresses
- Filter infrequent sites via browsing logs, non-local sites via geocoding

Entity Resolution
- Find URLs about same entity
- Identify canonical query (with the most traffic)
- Select other URLs > 5% of clicks for canonical query
- URLs comprise entity cluster

Entity Recommendation
- Generate affinity matrix, $A$
- $A_{i,j}$ relatedness $i$ and $j$ using:
  - SERP click graph
  - Search session co-occurrence
  - Merge click graph + session, inc. Max-flow and Hitting Time

Judging Entity Relatedness

Ground truth from Amazon Mechanical Turk
- URL-URL pairs via graph walk on $A$
- Rated relationship: 1=URLs unrelated, 2=URLs related, 3=URLs obviously related, and 4=URLs same entity. Authors judged 20 entities and many URLs for a gold standard set ($\kappa = 0.74$)
- Each HIT had three questions, one from gold standard set and two novel questions
- Quality control via qualification test and continual assessment
- 3,426 URL-URL ratings, each pair rated by three turkers ($\kappa = 0.43$)
- Evaluated using 800 pairs where all agreed

Findings

**Entity Resolution**
Two URLs point to same entity iff human rating = 4
Performance: $F_1 = 0.63$ (accurately find duplicates)

**Entity Recommendation**
Two entities related if rating = 2/3, unrelated if 1

$F_1$ scores for entity recommendation algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Click graph</th>
<th>Search sessions</th>
<th>Merged</th>
<th>Merged Max-flow</th>
<th>Merged Hitting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$-score</td>
<td>0.39</td>
<td>0.28</td>
<td>0.44</td>
<td>0.48</td>
<td>0.49</td>
</tr>
</tbody>
</table>

- Merged click-graph + search-session algorithm better than either method alone
- Merged model using max-flow and hitting time algorithms led to further improvements
- Session had less data, still strong performance