Social Web Search

Filippo Menczer
Informatics and Computer Science

Leveraging online social behavior for collaborative Web search
NaN: Networks & agents Network

http://homer.informatics.indiana.edu/~nan/
Outline

• Part 1: GiveALink.org (last semester)

• Part 2: 6S
Outline

• Part 1: GiveALink.org

• Part 2: 6S
http://homer.informatics.indiana.edu/~nan/6S/
1. Provide seed URLs

2. Set topic parameters

3. Review setup

4. Launch crawler/indexer

<table>
<thead>
<tr>
<th>Crawling Progress</th>
<th>11/09 14:53:34</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexing Progress</td>
<td>11/09 14:56:28</td>
<td>39%</td>
</tr>
</tbody>
</table>

Information Summary

Please verify the information you entered.

- **Document Collection:**
  - /Users/fil/Documents/Homepages/Informatics

- **Bookmark File:**
  - /Users/fil/Library/Safari/Bookmarks.plist

- **Crawling Topic:**
  - web mining modeling search peer crawl networks agents

- **Seed URLs File:**

- **Number of Pages:**
  - 100

- **Your Email:**
  - fil@indiana.edu

- **Tag for URLs:**
  - web mining modeling search peer crawl networks agents
<table>
<thead>
<tr>
<th>Indexing Date</th>
<th>Document Title</th>
<th>Document URL</th>
<th>Document Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006/11/09</td>
<td>ClickZ Internet Marketing Solutions for Marketers</td>
<td><a href="http://www.clickz.com/">http://www.clickz.com/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Geeking with Greg</td>
<td><a href="http://glinden.blogspot.com/">http://glinden.blogspot.com/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Crawling the Web Speakers</td>
<td><a href="http://www.dl.fc.ul.pt/sobre/bibtex.php?entr">http://www.dl.fc.ul.pt/sobre/bibtex.php?entr</a>...</td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Things an Observatory should have (institutional...</td>
<td><a href="http://www.conf.ecs.soton.ac.uk/perl/oai2?ve">http://www.conf.ecs.soton.ac.uk/perl/oai2?ve</a>...</td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Luminous Landscape Forum (Powered by Invisio...</td>
<td><a href="http://www.luminous-landscape.com/forum/">http://www.luminous-landscape.com/forum/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>VDS Technologies - Map Data</td>
<td><a href="http://www.vdstech.com/map_data.htm">http://www.vdstech.com/map_data.htm</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Mark Pincus Blog</td>
<td><a href="http://markpincus.typepad.com/">http://markpincus.typepad.com/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>W3C I18N Tutorial: CSS3 and International Text</td>
<td><a href="http://www.w3.org/International/tutorials/css">http://www.w3.org/International/tutorials/css</a>...</td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>OntoTech</td>
<td><a href="http://ontotech.blogspot.com/">http://ontotech.blogspot.com/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Search</td>
<td><a href="http://www.roundtablegroup.com/about/">http://www.roundtablegroup.com/about/</a>...</td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>First Tuesdays at the Urban Institute</td>
<td>03/2002</td>
<td><a href="http://www.urban.org/Presentations/FT/FT3-">http://www.urban.org/Presentations/FT/FT3-</a>...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Clustly Cloud Creator</td>
<td><a href="http://cloud.clustly.com/">http://cloud.clustly.com/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>ピッツバーグ便り</td>
<td><a href="http://mawaha.exblog.jp/">http://mawaha.exblog.jp/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Powerset: See what people are saying right now...</td>
<td><a href="http://www.technorati.com/tag/Powerset/">http://www.technorati.com/tag/Powerset/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>AI Resources and Links</td>
<td><a href="http://labs.rightnow.com/resources.php">http://labs.rightnow.com/resources.php</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>The Next Net: The Search for the Next Big Leap...</td>
<td><a href="http://business2.blogs.com/business2blog/2">http://business2.blogs.com/business2blog/2</a>...</td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>slides.dvi</td>
<td><a href="http://www.comp.hkbu.edu.hk/WI05/speaker/">http://www.comp.hkbu.edu.hk/WI05/speaker/</a></td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Amazon.com: The Search: How Google and Its...</td>
<td><a href="http://www.amazon.com/Search-Rewrote-Bu">http://www.amazon.com/Search-Rewrote-Bu</a>...</td>
<td>web mining modeling...</td>
</tr>
<tr>
<td>2006/11/09</td>
<td>Industry Solutions from RightNow Technologies 2</td>
<td><a href="http://www.rightnow.com/products/industry-">http://www.rightnow.com/products/industry-</a>...</td>
<td>web mining modeling...</td>
</tr>
</tbody>
</table>
http://homer.informatics.indiana.edu/~nan/6S/
http://homer.informatics.indiana.edu/~nan/6S/
6S: Collaborative Peer Search

Data mining & referral opportunities

Emerging communities

WWW2004
WWW2005
WTAS2005
P2PIR2006
Distributed vs Centralized
Algorithm 1: Random Known
Algorithm 2: Greedy
Algorithm 3: Reinforcement
Learning Peers

- A focused profile $W^f$ of other peers for terms in queries
- An expanded profile $W^e$ of other peers for terms not in queries, but that co-occur with query terms and with higher frequency

$$w_{i,p}(t + 1) = (1 - \gamma) \cdot w_{i,p}(t) + \gamma \cdot \left( \frac{S_p + 1}{S_l + 1} - 1 \right)$$

- Query routing:

$$\sigma(p, Q) = \sum_{i \in Q} \left[ \alpha \cdot w_{i,p}^f + (1 - \alpha) \cdot w_{i,p}^e \right]$$
Simulating 500 Users

ODP (dmoz.org)

P@10

[Graph showing time vs. coverage and global coherence for different algorithms: adaptive, expanded, adaptive, best hit, and queries per peer.]
Small-world
Ongoing Work

- TREC based evaluation framework (Klampanos et al. 2005)
- User study with 6Search application
- Improve query routing algorithm
  - Coverage
- Spam protection
Questions?

http://homer.informatics.indiana.edu/~nan/6S/