Agglomerative clustering of a search engine query log

Doug Beeferman  
Lycos

Adam Berger  
CMU

KDD 2000

Presented by: Cangming & Ronan

Motivation

- Query log analysis
  - Identifying late-breaking trends
- Clustering URLs
  - Generating ontology, organizing bookmarks, grouping search results
- Clustering queries
  - Query recommendations

Disclaimer

<table>
<thead>
<tr>
<th></th>
<th>2000</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workstation speed</td>
<td>266 MHz</td>
<td>&gt; 3 GHz</td>
</tr>
<tr>
<td>Lycos used</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>Queries / day</td>
<td>10 million</td>
<td>400 million</td>
</tr>
<tr>
<td>Size of the internet</td>
<td>?</td>
<td>5 million TB</td>
</tr>
</tbody>
</table>
Content ignorant

- Traditional approach
  - Extract feature vector from document

- Content ignorant approach
  - Less computationally expensive
  - Handle text-free pages, pages with restricted access or dynamic content

Graph-based Iterative Clustering

- Graph construction
- Similarity measure
- Iterative clustering
- Complexity / Optimization

Graph Construction

- Input: Query / URL pairs
- Each distinct query becomes a white node
- Each distinct URL becomes a black node
- For each pair, add an edge between the corresponding nodes

Example

(jean-Baptiste Jeannin, facebook.com/people/jean-Baptiste)
(jean-Baptiste Poquelin, en.wikipedia.org/wiki/Molière)
(Molière, en.wikipedia.org/wiki/Molière)
(Molière, imdb.com/title/tt0796335)
(Don Juan, imdb.com/title/tt0016804)
(Don Juan, en.wikipedia.org/wiki/Molière)
(Don Juan, imdb.com/title/tt0796335)
(Jane Winton, imdb.com/title/tt0016804)
(Fabrice Luchini, imdb.com/title/tt0796335)
(—, imdb.com/title/tt0796335)
(—, imdb.com/title/tt0796335)

Similarity measure

\[ \sigma(x, y) = \begin{cases} \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|} & \text{if } |N(x) \cup N(y)| > 0 \\ 0 & \text{otherwise} \end{cases} \]
Iterative Clustering

- Repeatedly merge the most similar pair, alternating between queries and URLs
- Until some stopping criterion is met
- Iterative approach helps to group queries (or URLs) that are otherwise uncorrelated

Complexity

- Naïve analysis: \( \Theta(n_0^2 + n_r^2) \times \text{iteration} \)
- Incremental computation of distances
  - Compute only non-zero distances
  - One-time computation
  - Re-compute only distances that are susceptible to have changed

Experiment: Query Recommendations

- Comparing 3 methods for building suggestion lists:
  - Baseline
  - Full-replacement
  - Hybrid
- Measure performance by clickthrough rate

Data

- Learning data: 500,000 query/URL pairs
- Test data: around 6 million impressions for each method

Results

<table>
<thead>
<tr>
<th>period</th>
<th>strategy</th>
<th>impressions</th>
<th>clicks</th>
<th>clickthrough rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 7-9</td>
<td>baseline</td>
<td>6,320,341</td>
<td>71,130</td>
<td>1.41%</td>
</tr>
<tr>
<td>April 16-15</td>
<td>hybrid</td>
<td>6,056,737</td>
<td>70,143</td>
<td>1.31%</td>
</tr>
<tr>
<td>April 21-22</td>
<td>full-replacement</td>
<td>5,846,907</td>
<td>65,272</td>
<td>1.44%</td>
</tr>
</tbody>
</table>

Discussion

- Similarity metric limitation
  - Two URLs shared should be better than one
  - More clicks should mean better correlation
  - Sensitivity to noisy clickthroughs

- Adding weights to edges
  - “Clustering Search Engine Query log containing noisy clickthroughs”, Wing Shun Chan et al., 2004.
Discussion

- Quality of the search engine
  - Clustering only as good as the search engine itself
  - Perhaps, it might possibly improve the search experience, and the IR.

- Limitation of experiments
  - Clickthrough does not reflect the quality of clustering
  - Users are likely to prefer search refinements over related searches