Dr. Searcher and Mr. Browser: A unified hyperlink-click graph

Barbara Poblete, Carlos Castillo, Aristides Gionis
University Pompeu Fabra, Yahoo! Research
Barcelona, Spain
CIKM 2008
In this talk

- New unified web graph: **hyperlink-click** graph
- Union of the hyperlink graph and the click graph
- Random walk generates robust results that match or are better than the results obtained on either individual graph
Motivation

- Two types of web graphs: **click** and **hyperlink** graph
- Represent two most common tasks of users: **searching** and **browsing**
- Correspond to two prototypical ways of looking for information: **asking** and **exploring**
- Edges capture semantic relations among nodes: e.g., similarity and authority endorsement
Motivation...

Drawbacks:

- Hyperlink graph: adversarial increase in PageRank scores, i.e., spam or link farms
- Click graph: sparsity, inherent bias in web-search engine ranking, dependency on textual match and click spam

Our claim: Hyperlink and click graphs are complementary and can be used to alleviate the shortcomings of each other
Motivation...

Drawbacks:

- Hyperlink graph: adversarial increase in PageRank scores, i.e., spam or link farms
- Click graph: sparsity, inherent bias in web-search engine ranking, dependency on textual match and click spam

Our claim: Hyperlink and click graphs are complementary and can be used to alleviate the shortcomings of each other
Contribution

Introduce the hyperlink-click graph, union of hyperlink and click graph

Consider random walk on the hyperlink-click graph

1. Model user behavior more accurately
2. Show that ranking with the hyperlink-click graph is similar to the best performance by either of the two graphs
3. The unified graph compensates where either graph performs poorly
4. More robust and fail-safe
Uses of the hyperlink-click graph:

- Ranking of documents
- Query ranking and query recommendation
- Similarity search
- Spam detection
Web graphs

hyperlink-click graph

clicked queries

frequency(q)

click graph

clicked documents

neighbor documents

hyperlink graph
Random walks on web graphs

Random walk on the hyperlink graph

\[ P_H = \alpha N_H + (1 - \alpha)1_H \]

Random walk on the click graph

\[ P_C = \alpha N_C + (1 - \alpha)1_C \]

Random walk on the hyperlink-click graph

\[ P_{HC} = \alpha \beta N_C + \alpha (1 - \beta)N_H + (1 - \alpha)1 \]
Random walks on web graphs

Random walk on the hyperlink graph

\[ P_H = \alpha N_H + (1 - \alpha)1_H \]

Random walk on the click graph

\[ P_C = \alpha N_C + (1 - \alpha)1_C \]

Random walk on the hyperlink-click graph

\[ P_{HC} = \alpha \beta N_C + \alpha (1 - \beta) N_H + (1 - \alpha)1 \]
Random walks on web graphs

Random walk on the hyperlink graph

\[ P_H = \alpha N_H + (1 - \alpha)1_H \]

Random walk on the click graph

\[ P_C = \alpha N_C + (1 - \alpha)1_C \]

Random walk on the hyperlink-click graph

\[ P_{HC} = \alpha \beta N_C + \alpha (1 - \beta) N_H + (1 - \alpha)1 \]
Web graphs

hyperlink-click graph

clicked queries

frequency(q)

clicked documents

neighbor documents

hyperlink graph

click graph
Evaluation and results

- Validate utility of the random walk scores on the hyperlink-click graph
- Compare scores with those produced from the hyperlink and the click graph
We focus on two tasks in which a good ranking method should perform well:

- ranking high-quality documents and
- ranking pairs of documents

The evaluation is centered on analyzing the dissimilarities among the different models
Evaluation and results...

Dataset:

- Yahoo! query log
- 9 K seed documents extracted from the query log
  - documents with at least 10 clicks
- 61 K queries with at least one click to the seed documents
- Using a web crawl expanded to 144 million documents
- The expansion considers all in- and out-neighbors of all documents in the seed set.
Evaluation and results...

Two types of data:

- Without sponsored results: query log only with *algorithmic* results
- With sponsored results:
Random walk evaluation

- we compute scores on the complete datasets, but
- we evaluate only on documents in the intersection of the three graphs

Two tasks:
1. ranking high-quality documents (DMOZ),
2. ranking pairs of documents (user evaluation).
DMOZ:

\( \Pi_Z \): Our first measure is the normalized sum of the \( \pi \) scores of \( D_Z \) documents.

\[
\Pi_Z = \frac{\sum_{d \in D_Z} \pi(d)}{\sum_{d \in D_C} \pi(d)}
\]

\( \Gamma_Z \): *Goodman-Kruskal Gamma* \( \Gamma \) measure between the rankings

\[
\Gamma = \frac{D - A}{D + A}
\]
User study:

- 13 users
- 1,710 assessments
- Users expressed not neutral opinion in 32% of document pairs
- Use $\Gamma$ to measure pairs of documents on which rankings agree with users
DMOZ:

- **Macro-evaluation**: captures the overall scores of high-quality documents

  $\Pi_Z$ and $\Gamma_Z$ are computed considering all the documents in the evaluation set.

- **Micro-evaluation**: at query level

  $\Pi_Z$ and $\Gamma_Z$ in the evaluation set are reduced to only those documents clicked from a particular query.

(User study: micro-evaluation)
DMOZ:

- **Macro-evaluation**: captures the overall scores of high-quality documents
  
  $\Pi_Z$ and $\Gamma_Z$ are computed considering all the documents in the evaluation set

- **Micro-evaluation**: at query level
  
  $\Pi_Z$ and $\Gamma_Z$ in the evaluation set are reduced to only those documents clicked from a particular query.

(User study: micro-evaluation)
### DMOZ

<table>
<thead>
<tr>
<th>Metric</th>
<th>Without Sponsored Results</th>
<th>With Sponsored Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Gamma_Z$</td>
<td>$G_C \approx G_{HC} &gt; G_H$</td>
<td>$G_C &gt; G_{HC} &gt; G_H$</td>
</tr>
<tr>
<td>$\Pi_Z$</td>
<td>$G_H \approx G_{HC} &gt; G_C$</td>
<td>$G_H \approx G_{HC} &gt; G_C$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Without Sponsored Results</th>
<th>With Sponsored Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Gamma_Z$</td>
<td>$G_C \approx G_{HC} &gt; G_H$</td>
<td>$G_{HC} \approx G_C &gt; G_H$</td>
</tr>
<tr>
<td>$\Pi_Z$</td>
<td>$G_{HC} \approx G_C &gt; G_H$</td>
<td>$G_{HC} \approx G_H \approx G_C$</td>
</tr>
</tbody>
</table>
User study

<table>
<thead>
<tr>
<th>metric</th>
<th>without sponsored results</th>
<th>with sponsored results</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Gamma$</td>
<td>$G_C &gt; G_{HC} &gt; G_H$</td>
<td>$G_H &gt; G_{HC} &gt; G_C$</td>
</tr>
</tbody>
</table>
Evaluation and results...

Exclude documents with very similar scores
Summary of $\Gamma_Z$ values for DMOZ:

<table>
<thead>
<tr>
<th></th>
<th>Macro (with variations)</th>
<th>Macro (no variations)</th>
<th>Micro (with variations)</th>
<th>Micro (no variations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.200</td>
<td>0.000</td>
<td>0.200</td>
<td>0.400</td>
</tr>
<tr>
<td></td>
<td>0.600</td>
<td>0.800</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

- click graph
- hyperlink-click graph
- hyperlink graph
Summary of $\Pi_Z$ values for DMOZ:

<table>
<thead>
<tr>
<th></th>
<th>Macro (with variations)</th>
<th>Macro (no variations)</th>
<th>Micro (with variations)</th>
<th>Micro (no variations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pi_Z$ value</td>
<td>0.200</td>
<td>0.300</td>
<td>0.400</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>0.600</td>
<td>0.700</td>
<td>0.800</td>
<td>0.900</td>
</tr>
</tbody>
</table>

Legend:
- click graph
- hyperlink-click graph
- hyperlink graph
Conclusions

- We study random walk on a unified web graph
- Intended to model user searching and browsing behavior
- Several combinations of metrics are used for evaluation

Experimental evaluation shows:
- The unified graph is always close to the best performance of either the click or the hyperlink graph
Future work

• Analyze how to deal with the inherent bias that exists in any ranking technique based on usage mining

• Study other web mining applications
  1. Link and click spam detection
  2. Similarity search
  3. Query recommendation
Thank you!