Beyond PageRank: Machine Learning for Static Ranking

Matthew Richardson\textsuperscript{1}, Amit Prakash\textsuperscript{1} Eric Brill\textsuperscript{2}

\textsuperscript{1}Microsoft Research
\textsuperscript{2}MSN

World Wide Web Conference, 2006
Types of Ranking

- **Dynamic Ranking**
  - Query-dependent ranking
  - Focussed on improving the ordering of results returned to user

- **Static Ranking**
  - Query-independent ranking
Introduction

Ranking Algorithms

- **PageRank**
  - A Single measure - Easier to manipulate by spammers
  - Computationally very expensive
Introduction

Ranking Algorithms

- **PageRank**
  - A Single measure - Easier to manipulate by spammers
  - Computationally very expensive

- **fRank**
  - Machine Learning Approach
  - Ranking based on many features besides the PageRank
  - 100 times faster in computation compared to PageRank
PageRank Algorithm

\[ P(j) = \frac{(1 - \alpha)}{N} + \alpha \sum_{i \in B_j} \frac{P(i)}{|F_i|} \]  

\( P(j) \) is the page rank score of page \( j \)  
\( F_i \) is the set of pages that page \( i \) links to  
\( B_j \) is the set of pages that link to \( j \)
fRank Algorithm

Feature Set

- PageRank
- Popularity
- Anchor text and inlinks
- Page-level and Domain-level features
fRank Algorithm

\[ Z = \{ < i, j > \} \] be a collection of pairs of items, where item \( i \) should be assigned a higher value than item \( j \).

Goal: Learn a function \( f \) such that,

\[ \forall < i, j > \in Z, f(x_i) > f(x_j) \]

Ranking function is modeled by fully connected 2 layer neural network with 10 hidden nodes
Experimental setup

Training set

- Human judgements for 28000 queries resulting in an average of 18 ratings per query.
- The network is trained with 5 million pairings of web pages, where one page had a higher rating than the other.

Performance Measure

- Pairwise accuracy defined as the fraction of time that the ranking algorithm and human judges agree on ordering of a pair of web pages.
## Results

Table: Basic Results

<table>
<thead>
<tr>
<th>Technique</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>56.70</td>
</tr>
<tr>
<td>fRank</td>
<td>67.43</td>
</tr>
</tbody>
</table>
Results

Table: Results for individual feature sets

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>56.70</td>
</tr>
<tr>
<td>Popularity</td>
<td>60.82</td>
</tr>
<tr>
<td>Anchor</td>
<td>59.09</td>
</tr>
<tr>
<td>Page</td>
<td>63.93</td>
</tr>
<tr>
<td>Domain</td>
<td>59.03</td>
</tr>
<tr>
<td>All Features</td>
<td>67.43</td>
</tr>
</tbody>
</table>
Table: fRank performance as feature sets are added. At each row, the feature set that gave the greatest increase in accuracy was added to the list of features (i.e., we conducted a greedy search over feature sets)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>50.00</td>
</tr>
<tr>
<td>+Page</td>
<td>63.93</td>
</tr>
<tr>
<td>+Popularity</td>
<td>66.83</td>
</tr>
<tr>
<td>+Anchor</td>
<td>67.25</td>
</tr>
<tr>
<td>+PageRank</td>
<td>67.31</td>
</tr>
<tr>
<td>+Domain</td>
<td>67.43</td>
</tr>
</tbody>
</table>
## Results

**Table: Top ten URLs for PageRank vs. fRank**

<table>
<thead>
<tr>
<th>PageRank</th>
<th>fRank(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>google.com</td>
<td>google.com</td>
</tr>
<tr>
<td>apple.com/quicktime/download</td>
<td>yahoo.com</td>
</tr>
<tr>
<td>amazon.com</td>
<td>americanexpress.com</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>hp.com</td>
</tr>
<tr>
<td>microsoft.com/windows/ie</td>
<td>target.com</td>
</tr>
<tr>
<td>apple.com/quicktime</td>
<td>bestbuy.com</td>
</tr>
<tr>
<td>mapquest.com</td>
<td>dell.com</td>
</tr>
<tr>
<td>ebay.com</td>
<td>autotrader.com</td>
</tr>
<tr>
<td>mozilla.org/products/firefox</td>
<td>dogpile.com</td>
</tr>
<tr>
<td>ftc.gov</td>
<td>bankofamerica.com</td>
</tr>
</tbody>
</table>

Presented by Srinivas Pasupuleti

Beyond PageRank: Machine Learning for Static Ranking
Conclusions and Future Work

Conclusions

- **PageRank** doesn’t provide good static ranking
- **fRank** achieves higher pairwise accuracy
  - Page-level and popularity features are the significant contributors
  - Ordering using **fRank** correspond to what Web users, rather than Web page authors, prefer
Conclusions and Future Work

Future Work

- Many more features can be included
  - Existence, or lack there of certain words ("under construction")
  - Number of images in a page, size of those images
  - Searching habits of users, by time of the day

- Incorporating fRank and page-level features directly into the PageRank computation
My Review

- Need of more intuitive way of collecting training data
- “popularity” and PageRank generally go together
- Behavior of Web User and Web Author is not independent
References

Matthew Richardson, Amit Prakash and Eric Brill. Beyond PageRank: Machine Learning for Static Ranking. *WWW 2006*


Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *TR, Stanford 1998*