Topical TrustRank: Using Topicality to Combat Web Spam

Baoning Wu, Vinay Goel
and Brian D. Davison
Lehigh University, USA
Outline

- Motivation
- Topical TrustRank
- Experiments
- Conclusion
Background

- **Web spam**
  - Behavior having the effect of manipulating search engines’ ranking results

- **TrustRank** introduced notion of trust to demote spam pages
  - Link between two pages signifies trust between them
  - Initially, human experts select a list of seed sites that are well known and trustworthy
  - A biased PageRank algorithm is used
  - Spam sites will have poor trust scores
Formal TrustRank Definition

\[ t = \alpha \times T \times t + (1 - \alpha) \times d \]

- \( t \): TrustRank score vector
- \( T \): transition matrix
- \( \alpha \): decay factor
- \( d \): trust score vector of seed set
Issues with TrustRank

- Coverage of the seed set may not be broad enough
  - Many different topics exist, each with good pages
- TrustRank has a bias towards communities that are heavily represented in the seed set
  - inadvertently helps spammers that fool these communities
Bias towards larger partitions

\[ t = \frac{m_1}{\sum_{i=1}^{n} m_i} t_1 + \frac{m_2}{\sum_{i=1}^{n} m_i} t_2 + \ldots + \frac{m_n}{\sum_{i=1}^{n} m_i} t_n \]

- Divide the seed set into \( n \) partitions, each has \( m_i \) nodes
- \( t_i \): TrustRank score calculated by using partition \( i \) as the seed set
- \( t \): TrustRank score calculated by using all the partitions as one combined seed set
Outline

- Motivation
- Topical TrustRank
- Experiments
- Conclusion
Basic ideas

- Use pages labeled with topics as seed pages
  - Pages listed in highly regarded topic directories

- Trust should be propagated by topics
  - link between two pages is usually created in a topic specific context
Topical TrustRank

- **Topical TrustRank**
  - Partition the seed set into topically coherent groups
  - TrustRank is calculated for each topic
  - Final ranking is generated by a combination of these topic specific trust scores

- **Note**
  - TrustRank is essentially biased PageRank
  - Topical TrustRank is fundamentally the same as Topic-Sensitive PageRank, but for demoting spam
Comparison of topical contribution

\[ t = \frac{m_1}{\sum_{i=1}^{n} m_i} t_1 + \frac{m_2}{\sum_{i=1}^{n} m_i} t_2 + \ldots + \frac{m_n}{\sum_{i=1}^{n} m_i} t_n \]

\[ t = t_1 + t_2 + \ldots + t_n \]
Combination of trust scores

- Simple summation
  - default mechanism just seen
    \[ t = t_1 + t_2 + \ldots + t_n \]

- Quality bias
  - Each topic weighted by a bias factor
  - Summation of these weighted topic scores
    \[ t = w_1 t_1 + w_2 t_2 + \ldots + w_n t_n \]
  - One possible bias: Average PageRank value of the seed pages of the topic
Further Improvements

- **Seed Weighting**
  - Instead of assigning an equal weight to each seed page, assign a weight proportional to its quality / importance

- **Seed Filtering**
  - Filtering out low quality pages that may exist in topic directories

- **Finer topics**
  - Lower layers of the topic directory
Outline

- Motivation
- Topical TrustRank
- Experiments
- Conclusion
Data sets

- 20M pages from the Swiss search engine (search.ch)
  - 350K sites
  - 3,589 labeled spam sites
  - dir.search.ch for topics

- Stanford WebBase crawl for Jan, 2001
  - 65M pages
  - Dmoz.org Open Directory Project RDF of Jan, 2001
Ranking

- Each site/page has three rankings:
  - PageRank, TrustRank and Topical TrustRank (with different combination methods and improvement ideas)
- Sites/pages are distributed in decreasing order across 20 buckets, such that the sum of PageRank values in each bucket are equal.
Metrics

- Number of spam pages within top buckets
  - Top 10 buckets

- Overall movement
  - The sum of the movement in terms of buckets observed for each spam page
## Basic results on search.ch data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No. of spam sites within top 10 buckets</th>
<th>Overall movement</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>90</td>
<td>-</td>
</tr>
<tr>
<td>TrustRank</td>
<td>58</td>
<td>4,537</td>
</tr>
<tr>
<td>Topical TrustRank (simple summation)</td>
<td>42</td>
<td>4,620</td>
</tr>
</tbody>
</table>
## Improvements to Topical TrustRank

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of spam sites within top 10 buckets</th>
<th>Overall movement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple summation</td>
<td>42</td>
<td>4,620</td>
</tr>
<tr>
<td>Quality bias</td>
<td>40</td>
<td>4,620</td>
</tr>
<tr>
<td>Seed weighting</td>
<td>37</td>
<td>4,548</td>
</tr>
<tr>
<td>Seed filtering</td>
<td>42</td>
<td>4,671</td>
</tr>
<tr>
<td>Two-layer topics</td>
<td>37</td>
<td>4,604</td>
</tr>
<tr>
<td>Aggregation of above</td>
<td>33</td>
<td>4,617</td>
</tr>
</tbody>
</table>
Topical composition of spam sites

- TrustRank
- Topical TrustRank
Results for WebBase data

- For pages demoted by TrustRank, the spam ratio is 20.2%.

- For pages demoted by Topical TrustRank, the spam ratio is 30.4%.

- With improvements (seed filtering + seed weighting + quality bias), the spam ratio is 32.9%.
Spam pages in WebBase data set

- PageRank
- TrustRank
- Topical TrustRank

Number of spam pages

Buckets

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Outline

- Motivation
- Topical TrustRank
- Experiments
- Conclusion
Conclusion

- Topical TrustRank combines topical information with the notion of trust.
- Topical TrustRank (simple summation) demotes 27.6% additional highly ranked spam sites over TrustRank.
- Improvements to the Topical TrustRank algorithm achieved an additional 15.5%.
Future work

- Explore other partitioning strategies.
- Lessons learned may be applied to personalized search.
- Better techniques to combine trust scores.
- Better models for trust propagation.
Thank You!

- Baoning Wu
- baw4@cse.lehigh.edu
- http://wume.cse.lehigh.edu/