Topical TrustRank:
Using Topicality to Combat Web Spam

Baoning Wu, Vinay Goel, Prof. Brian D. Davison
CSE 450 Term Project
Introduction

- Problem of spam
- No universal technique to combat all types of spam
- TrustRank introduced notion of trust to demote spam pages
- We improve on TrustRank using topical information - Topical TrustRank
TrustRank

- 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
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 help spammers that fool these communities
Explanation

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

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

- Propose the use of pages listed in well maintained topic directories as seed pages.
- Trustworthiness of a page should be differentiated by topics.
  - link between two pages is usually created in a topic specific context.
Topical TrustRank

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

- Partition the seed set
- Compute TrustRank for each partition
- Combine the trust scores of each partition
Partitioning

- Random
- By topic
Combination of trust scores

- Simple summation
- Quality bias
  - each topic weighted by a bias factor
  - summation of these weighted topic scores
  - one such bias: Average PageRank value of the seed pages of the topic
Improvements

- Seed Weighting
- Seed Filtering
- Finer topics hierarchy
Seed Weighting

- Instead of assigning an equal weight to each seed page,
  - assign a weight proportional to its quality / importance
  - use the normalized PageRank value of each seed page within the seed set
Seed Filtering

- Low quality pages may exist in topic directories
- Need to filter out these pages
- Use PageRank / TrustRank / Topical TrustRank for filtering
Finer Topics Hierarchy

- Most researchers use only top level topics.
- A finer topic hierarchy may be more accurate to categorize pages on the web.
- In Topical TrustRank, this has the effect of producing better partitions.
Data sets

- 20M pages from search.ch company
  - 35K sites
  - 3,589 labeled spam sites
  - dir.search.ch
- WebBase data for Jan, 2001
  - 65M pages
  - DMOZ RDF Jan, 2001
Initial comparison
Ranking

- Each page has three rankings:
  - PageRank, TrustRank and Topical TrustRank

- Pages are put into 20 buckets
  - Sum of values of pages within each bucket is 5%
Metrics

- Number of spam pages within top buckets
- Overall movement
  - The sum of movement for each spam page
Basic results on search.ch data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No. 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</td>
<td>42</td>
<td>4,620</td>
</tr>
<tr>
<td>Method</td>
<td>No. within top 10 buckets</td>
<td>Overall movement</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------</td>
<td>------------------</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>Quality bias</td>
<td>40</td>
<td>4,620</td>
</tr>
<tr>
<td>Two-level topic</td>
<td>37</td>
<td>4,604</td>
</tr>
<tr>
<td>Combination</td>
<td>33</td>
<td>4,617</td>
</tr>
</tbody>
</table>
Spam sites
Result 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%.

- For combination of ideas, the spam ratio is 32.9%.
Distribution of the 133 spam pages for WebBase data set
Conclusion

• Effective approach to demote spam
• Use of topical information
• Future work
  • explore partitioning strategies
  • lessons learned may be applied to Personalized Searching techniques
  • better techniques to combine trust scores