Combating Web Spam with TrustRank

Zoltán Gyöngyi, Hector Garcia-Molina, Jan Petersen

Kristian Wolpers
Overview

• Problem of Web Spam
• State of the Art
• Idea
• TrustRank
• Experiments
• Related Work
Problem of Web Spam

What is Web Spam?
Web spam pages use a lot of tricks to get a higher rank in search engines than they deserve

e.g.:

• Keyword stuffing
• Meta tag stuffing
• etc
Problem of Web Spam

The trouble is ...

... it is too expansive, if human experts should verify every result to be non-spam
State of the art

Most search engines uses PageRank to assign global importance scores on all web pages. Therefore, PageRank models the web as a graph with web pages as knots $V$ and links as edges $E$ and then calculates there rank scores $r$ with a decay factor $\alpha$

$$r(p) = \alpha \cdot \sum_{q : (q, p) \in E} \frac{r(q)}{\omega(q)} + (1 - \alpha) \cdot \frac{1}{N}$$
Idea

• Use a similar calculation of trust like PageRank
• Use a manageable number of seed pages which can be verified by human experts
• Calculate the trust of all web pages based on this seed pages
TrustRank

TrustRank works in 4 Steps:

1. Evaluate seed-desirability of pages via inverse PageRank
2. Generate a corresponding order of the seeds
3. Select good seeds by human experts
4. Compute TrustRank scores of all web pages based on the seed pages
TrustRank

To compute the TrustRank scores they use a variation of PageRank:

$$\text{TrustRank} = \alpha \cdot T \cdot t^* + (1 - \alpha) \cdot d$$
TrustRank: Example

\[ s = [0.08, 0.13, 0.08, 0.10, 0.09, 0.06, 0.02] \]
\[ \sigma = [2, 4, 5, 1, 3, 6, 7] \]
\[ o = [1, 1, 0] \]
\[ t^* = [0, 0.18, 0.12, 0.15, 0.13, 0.05, 0.05] \]
Experiments

Data set:

• Complete set of pages crawled by AltaVista

• Regrouped Pages into nearly 31 million sites

• Aligned the pages via inverse PageRank and focused on the 25,000 top sites

• After a few more steps 178 sites remaining as good seeds and a sample of 1000 sites aligned in 20 buckets
Experiments

![Bar chart showing good and bad pages across experiments. The x-axis represents experiment numbers from 1 to 20, and the y-axis represents the number of good and bad pages. The chart indicates that certain experiments have significantly more bad pages compared to others.]
Related Work

• TrustRank builds on existing PageRank research
• The usage of custom distribution vectors comes from topic-sensitive PageRank