Identifying Web Spam

Presented By: Rob Dye
Outline

• Introduction
  – Limitations of PageRank
• Simple Solutions and Less Simple Solutions
  – Content Filtering
  – BadRank
  – ParentRank
  – TrustRank
Introduction

• Limitations of PageRank
  – Only works if sites are honest
    • As many as 75 million of the 150 million web servers online today are believed to exist solely for the purpose of increasing the page rank of their targeted site. [Drost and Scheffer, 2005]
  – Link Farms
Nigritude Ultramarine?

• Total Nonsense Phrase
• 2004 SEO contest
  – http://www.nigritudeultramarines.com/
  – 0 hits before contest, 500,000 hits after contest
• Techniques used by participants
  – False Spam Reports
  – Cloaking
  – Link Farms
  – Page Count Inflation
  – Forum Spamming
  – Blog Spamming
How Can We Stop Spam Sites?

• Simple Solution 1: Manual Rankings
  – Works well but too time consuming
  – Not scalable

• Simple Solution 2: Spam Filter
  – Automatic and scalable
  – As with email spam filters, it does not always work.
PageRank 0 (PR0)

- A penalty applied to pages that use Link Farms
- Pages marked as “bad” get an automatic pageRank value of 0.
- Sometimes called BadRank
  - Measure to indicate how “bad” a page is
  - Link to other “bad” sites and your badness increases
  - Bad sites also propagate badness to sites that they link to
BadRank Formula

\[ BR(A) = E(A)(1 - d) + d \sum_{i=1}^{n} \frac{BR(T_i)}{C(T_i)} \]

- **BadRank**:  
  - \( E(A) \) = base “bad” value of page  
  - \( BR(T_i) \) = BadRank of page \( T_i \)  
  - \( C(T_i) \) = \# of inbound links of page \( T_i \)  
  - \( d \) = damping factor (same as in pageRank)
Combining PageRank and BadRank

• Several Possible Solutions:
  – Include BadRank in PageRank formula
  – Subtract BadRank from PageRank
  – Divide PageRank by BadRank
  – Normalize BadRank \([0-1]\) and multiply
  – Step Function
Calculating $E(A)$?

• We have looked at how to deal with badness, but how do we determine badness in the first place?

• Several different approaches:
  – ParentRank
  – TrustRank
  – Statistical Analysis
ParentRank

- Developed by Leigh University
- 2005 International World Wide Web Conference
- Step 1: Generate set of seeds that are known to be spam
- Step 2: Expand seed set
- Step 3: Combine badness ranking with page ranking
ParentRank

- Step 1: (Seed Step)
  - If a page has more than threshold links to domains that also link back to its domain, add it to seed list.
ParentRank

• Step 2: (Expansion Step)
  – If a page has more than \langle threshold\rangle links to a page in the existing spam set, add it to set.
  – Repeat until no new pages are added.
Parent Rank

• **Step 3:** (Combining Step)
  
  – Already discussed several ways to do this.
  
  – Badness is binary here, so none of these methods are applicable.
  
  – **Solution:** delete all links to or from “bad” pages so they will not influence ranks.
ParentRank Conclusions

• Pros:
  – Can tweak threshold to change sensitivity
• Cons:
  – Badness is binary, so no way for a page to be slightly bad.
• They claim better results than HITS when users were asked to rank relevance of results for ParentRank and HITS.
TrustRank

• Joint venture between Yahoo! and Stanford
• Not entirely automated as some human input is required.
Consulting the Oracle

• YAHOO = “Yet Another Hierarchical Officious Oracle”

• When the algorithm is unsure about badness of a page, the “oracle function” is called.
  – Oracle Function asks a human to judge page
    • 1 if good, 0 if bad
Consulting the Oracle

• The Oracle is slow and should be consulted as seldom as possible.
• How do we avoid consulting the oracle every time?
  – Approximate Isolation of Good Sets Rule
    • Good pages rarely link to bad ones
Trust Function

• Trust Function: $T(p)$
  – returns probability that page $p$ is trustworthy
  – Ideally, for 100 pages with $T(p) = 0.7$, if the oracle were called, it would give 1 as the response 70 times and 0 as the response 30 times.

• if $T(p) >$ threshold then page is trustworthy
• if $T(p) <$ threshold then oracle must be called
Trust Propagation

• How is trust determined without asking oracle?
• Pages ranked as trustworthy by an oracle distribute trust to pages they link to.
Selecting Order of Review

- Order matters. To reduce Oracle invocation, start with pages that have many links to pages with many links.
- Given reasonably connected graph, may not take too many oracle calls to calculate trust for all pages.
TrustRank Conclusions

• Shows improvements over PageRank but with the added expense of “oracle invocations”
Determining Badness with Content Analysis

• Spam is not distributed evenly across the web. It varies by:
  – website domain
  – language
  – word count in title
  – average document word length
  – compressibility
Spam Across Domains

![Graph showing the percentage of spam across different domains. The graph has bars for each domain: biz, us, com, de, net, uk, org, and edu. The biz domain shows the highest percentage of spam, followed by us. The other domains have much lower percentages.](image-url)
Spam Across Languages

![Bar chart showing the percentage of spam across different languages: French, German, English, Japanese, and Chinese. The chart indicates that French and German have the highest percentage of spam, followed by English, Japanese, and Chinese, respectively.](chart.png)
Spam by Word Count

The graph shows the fraction of pages against the number of words, indicating the probability of spam. As the number of words increases, the fraction of pages decreases, with a notable peak around 250 words, followed by a gradual decrease to a lower probability of spam for pages with a higher word count.
Spam by Average Word Length

• Due to Word Stuffing:
  – ex (musicDownload, freePictures)
Spam by Compression Ratio
Spam by Common Word Count

Graph showing the fraction of pages with the fraction of 500 most frequent words in the corpus that are common with text, plotted against the probability of spam.
Spam by Title Word Count
Spam by Visible Content Ratio
Spam By Content Analysis

• Lots of Metrics but none are foolproof and many are noisy

• Use Data Mining techniques
  – C4.5 classifier used in tests
  – Catches 83% of spam while mistakenly identifying 2% of legitimate content as spam.
  – Is this good enough?
    • By itself, probably not
    • Combined with link analysis, maybe
Conclusion

• Nothing is perfect.
  – “It is our hope that continued research on this front will make effective spam more expensive than genuine content.” – Fetterly, et al.
References

Questions?