The effects of Web Spam on
The Evolution of Search Engines

CS315-Web Search and Mining

Have you ever used the Web...

- to get informed?
- to help you make decisions?
- Financial
- Medical
- Political
- Religious...

- The Web is huge
  - 1 trillion (!) static pages publicly available,
  - and growing every day
  - Much larger, if you count the "deep web"
  - Infinite, if you count pages created on-the-fly

We depend on search engines to find information

The Web has Spam...

Search results steroid drug HGH (human growth hormone)

Any controversial issue will be spammed

Search results for mental disease ADHD
(attention-deficit/hyperactivity disorder)
Political issues will be spammed...

... you like it or not!

A Brief History of Search Engines

Why is there Web Spam?

Political issues will be spammed

Search results for Senatorial candidate John N. Kennedy, 2008 USA Elections

Famous search results for "miserable failure"

Why is there Web Spam?

Web Spam:
- Attempt to modify the web (its structure and contents), and thus influence search engine results in ways beneficial to web spammers

A Brief History of Search Engines

1st Generation (ca 1994):
- AltaVista, Excite, Infoseek...
- Ranking based on Content:
  - Pure Information Retrieval

2nd Generation (ca 1996):
- Lycos
- Ranking based on Content + Structure
- Site Popularity

3rd Generation (ca 1998):
- Google, Teoma, Yahoo
- Ranking based on Content + Structure + Value
- Page Reputation

In the Works:
- Ranking based on "the user's need behind the query"
1st Generation: Content Similarity

- **Content Similarity Ranking**: The more rare words two documents share, the more similar they are.
- Documents are treated as "bags of words" (no effort to "understand" the contents).
- Similarity is measured by vector angles.
- Query Results are ranked by sorting the angles between query and documents.
- How To Spam?

1st Generation: How to Spam

- **“Keyword stuffing”:** Add keywords, text, to increase content similarity.

2nd Generation: Add Popularity

- A hyperlink from a page in site A to some page in site B is considered a **popularity vote** from site A to site B.
- Rank similar documents according to popularity.
- How To Spam?

2nd Generation: How to Spam

- **Create “Link Farms”:** Heavily interconnected owned sites spam popularity.

3rd Generation: Add Reputation...

- The reputation “PageRank” of a page $P_i$ is:
  $$P_i = (1 - d) + d \sum_{j=1}^{n} \frac{P_j}{C_j}$$

- Idea similar to academic co-citations
- Beautiful Math behind it
- $PR$ = principal eigenvector of the web’s link matrix
- $PR$ equivalent to the chance of randomly surfing to the page
- HITS algorithm tries to recognize “authorities” and “hubs”

3rd Generation: How to Spam

- Organize Mutual Admiration Societies: “link farms” of irrelevant reputable sites

Resource Partner Additions

- Organize Mutual Admiration Societies: “link farms” of irrelevant reputable sites

How To Spam?

- Organize Mutual Admiration Societies: “link farms” of irrelevant reputable sites
- Advertisement Consultants
- Conferences

Mutual Admiration Societies via Link Exchange

An Industry is Born

- “Search Engine Optimization” Companies
- Advertisement Consultants
- Conferences
3rd Generation: Reputation & Anchor Text

- Anchor text tells you what the reputation is about

- How To Spam?

  ![Diagram](image)

  - Page A
  - Page B

  ![Anchor](image)

  - Joe`s computer hardware links
  - Compaq
  - HP
  - IBM

  - Big Blue today announced record profits for the quarter

  - www.ibm.com

  - Armonk, NY-based computer giant IBM announced today

“Google-bombs” spam Anchor Text...

- Business weapons
  - “more evil than satan”

- Political weapon in pre-election season
  - “miserable failure”
  - “waffles”
  - “Clay Shaw” (+ 50 Republicans)

- Misinformation
  - Promote steroids
  - Discredit ADHD research

- Activism / online protest
  - “Egypt”
  - “Jew”

- Other uses we do not know?
  - “views expressed by the sites in your results are not in any way endorsed by Google…”

... mostly for political purposes

The Multiplier: Google Bomb The Elections

Now that you know how Chris Bowers and joined the effort by adding the anchor text "IBM" in the sentences for your hit's website, opening a anchor few minutes and all your blogging software suddenly the effect of adding the links to your blog template.

Here are some tips you can use:

1. Each time you post a link, you are telling google that you are voting for that link to be relevant in the next page on which people will search for the same.
2. You want to get a lot of your links, not a few automatically create an artificial network of your pages and directories. You can think of the anchor text as the votes for your link. The more votes you get for your link, the greater the chance it will be considered relevant.

But there is a limit to this technique, the limit to your blog template. That will occur in every single blog from your blog having a similar anchor length results in the same relevance. If you have been using this technique for a year, the anchor text may not reach the critical mass where the anchor text can be considered relevant.

The process of adding a link to your template is simple. All you need is a search engine query for the anchor text you want to link to. Then you need to search for the anchor text in your own blog.

Activists openly collaborating to Google-bomb search results of political opponents in 2006

Search Engines vs Web Spam

- Search Engine's Action
  - 1st Generation: Similarity
    - Content
  - 2nd Generation: Popularity
    - Content + Structure
  - 3rd Generation: Reputation
    - Content + Structure + Anchor Text
  - 4th Generation (in the Works)
    - Ranking based on the user's "need behind the query"

- Web Spammers Reaction
  - Add keywords so as to increase content similarity
  - Create "link farms" of heavily interconnected sites
  - Organize "mutual admiration societies" of irrelevant reputable sites
  - Googlebombs

Is there a pattern on how to spam?

Can you guess what they will do?
And Now For Something Completely(?) Different

**Propaganda:**
- Attempt to modify human behavior, and thus influence people’s actions in ways beneficial to propagandists.

**Theory of Propaganda**
- Developed by the Institute for Propaganda Analysis 1938-42.

**Propagandistic Techniques (and ways of detecting propaganda):**
- **Word games** - associate good/bad concept with social entity
- **Glittering Generalities** — Name Calling
- **Transfer** - use special privileges (e.g., office) to breach trust
- **Testimonial** - famous non-experts’ claims
- **Plain Folk** - people like us think this way
- **Bandwagon** - everybody’s doing it, jump on the wagon
- **Card Stacking** - use of bad logic

Societal Trust is (also) a Graph

<table>
<thead>
<tr>
<th>Graph Theory</th>
<th>Web Graph</th>
<th>Trust Social Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node weight computation</td>
<td>web page or site rank (accord. to a search engine)</td>
<td>social entity reputation (accord. to a person)</td>
</tr>
<tr>
<td>Arc semantics</td>
<td>hyperlink</td>
<td>consensus (e.g., 2 recommendations)</td>
</tr>
<tr>
<td>weight range</td>
<td>“vote of confidence” degree of confidence</td>
<td>computed continuously</td>
</tr>
<tr>
<td></td>
<td>[0...1]</td>
<td>computed on demand</td>
</tr>
</tbody>
</table>

**Web Spam:**
- Attempt to modify the Web Graph, and thus influence users through search engine results in ways beneficial to web spammers.

Web Spammers as Propagandists

- Web Spammers can be seen as employing propagandistic techniques in order to modify the Web Graph.

*There is a pattern on how to spam!*

<table>
<thead>
<tr>
<th>S.E.’s</th>
<th>Ranking</th>
<th>Spawning</th>
<th>Propaganda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Gen</td>
<td>Doc Similarity</td>
<td>keyword stuffing</td>
<td>glittering generalities</td>
</tr>
<tr>
<td>2nd Gen</td>
<td>+ Site popularity</td>
<td>+ link farms</td>
<td>+ bandwagon</td>
</tr>
<tr>
<td>3rd Gen</td>
<td>+ Page reputation</td>
<td>+ mutual admiration societies</td>
<td>+ testimonials</td>
</tr>
<tr>
<td></td>
<td>+ anchor text</td>
<td>+ link bombs</td>
<td>+ card stacking</td>
</tr>
</tbody>
</table>

Propaganda in Graph Terms

- **Word Games**
  - Name Calling
  - Glittering Generalities
- **Transfer**
- **Testimonial**
- **Plain Folk**
- **Bandwagon**
- **Card Stacking**
- **Modify Node weights**
  - Decrease node weight
  - Increase node weight
- **Modify Arcs**
  - Modify Arc content + keep weights
  - Modify Arcs & generate nodes