Detecting Spam Web Pages

Marc Najork
Microsoft Research, Silicon Valley
About me

- 1989-1993: UIUC (home of NCSA Mosaic)
- 1993-2001: Digital Equipment/Compaq
  - Started working on web search in 1997
  - Mercator web crawler (used by AltaVista)
- 2001-now: Microsoft Research
  - Measuring web evolution
  - Link-based ranking (algorithms and infrastructure)
  - Web spam detection
About MSR Silicon Valley

- One of six MSR labs (founded in 2001)
- Located in Mountain View
- About 30 full-time researchers
- Areas
  - Algorithms & Theory
  - Distributed Systems
  - Security & Privacy
  - Software Tools
  - Web Search & Data Mining
- We’re growing & hiring!
There’s gold in those hills

- E-Commerce is big business
  - Total US e-Commerce sales in 2004: $69.2 billion (1.9% of total US sales) (US Census Bureau)
  - Grow rate: 7.8% per year (well ahead of GDP growth)
  - Forrester Research predicts that online US B2C sales (incl. auctions & travel) will grow to $329 by 2010 (13% of all US retail sales)
Search engines direct traffic

- Significant amount of traffic results from Search Engine (SE) referrals
  - E.g. Jacob Nielsen’s site “HyperTextNow” receives one third of its traffic through SE referrals

- Only sites that are highly placed in SE results (for some queries) benefit from SE referrals
Ways to increase SE referrals

- Buy keyword-based advertisements
- Improve the ranking of your pages
  - Provide genuinely better content, or
  - “Game” the system

- “Search Engine Optimization” is a thriving business
  - Some SEOs are ethical
  - Some are not …
Web spam
(you know it when you see it)
Defining web spam

- Working Definition
  - Spam web page: A page created for the sole purpose of attracting search engine referrals (to this page or some other “target” page)

- Ultimately a judgment call
  - Some web pages are borderline useless
  - Some pages look fine in isolation, but in context are clearly “spam”
Why web spam is bad

● Bad for users
  ● Makes it harder to satisfy information need
  ● Leads to frustrating search experience

● Bad for search engines
  ● Burns crawling bandwidth
  ● Pollutes corpus (infinite number of spam pages!)
  ● Distorts ranking of results
Detecting Web Spam

- Spam detection: A classification problem
  - Given salient features, decide whether a web page (or website) is spam

- Can use automatic classifiers
  - Plethora of existing algorithms (Bayes, C4.5, SVM, …)
  - Use data sets tagged by human judges to train and evaluate classifiers (this is expensive!)

- But what are the “salient features”?
  - Need to understand spamming techniques to decide on features
  - Finding the right features is “alchemy”, not science
  - Spammers adapt – it’s an arms race!
Taxonomy of web spam techniques

- “Keyword stuffing”
- “Link spam”
- “Cloaking”
Keyword stuffing

- Search engines return pages that contain query terms
  - (Certain caveats and provisos apply ...)
- One way to get more SE referrals: Create pages containing popular query terms (“keyword stuffing”)
- Three variants:
  - Hand-crafted pages (ignored in this talk)
  - Completely synthetic pages
  - Assembling pages from “repurposed” content
Examples of synthetic content

Monetization

Random words

Well-formed sentences stitched together

Links to keep crawlers going
Examples of synthetic content

Creative ideas for valentine's day gifts and christening gift idea including holiday office party ideas


Someone's wedding site!
Features identifying synthetic content

- Average word length
  - The mean word length for English prose is about 5 characters

- Word frequency distribution
  - Certain words ("the", "a", …) appear more often than others

- N-gram frequency distribution
  - Some words are more likely to occur next to each other than others

- Grammatical well-formedness
  - Alas, natural-language parsing is expensive
Really good synthetic content

“Nigritude Ultramarine”: An SEO competition

Links to keep crawlers going

Grammatically well-formed but meaningless sentences

“Nigritude Ultramarine” is a synthetic content example that is designed to keep crawlers going with grammatically well-formed but meaningless sentences.
Content “repurposing”

- Content repurposing: The practice of incorporating all or portions of other (unaffiliated) web pages
  - A “convenient” way to machine generate pages that contain human-authored content
  - Not even necessarily illegal …

- Two flavors:
  - Incorporate large portions of a single page
  - Incorporate snippets of multiple pages
Example of page-level content “repurposing”
Example of phrase-level content “repurposing”
Techniques for detecting content repurposing

- Single-page flavor: Cluster pages into equivalence classes of very similar pages
  - If most pages on a site are very similar to pages on other sites, raise a red flag
  - (There are legitimate replicated sites; e.g. mirrors of Linux man pages)

- Many-snippets flavor: Test if page consists mostly of phrases that also occur somewhere else
  - Computationally hard problem
  - Have probabilistic technique that makes it tractable
Detour: Link-based ranking

- Most search engines use hyperlink information for ranking
- Basic idea: Peer endorsement
  - Web page authors endorse their peers by linking to them
- Prototypical link-based ranking algorithm: PageRank
  - Page is important if linked to (endorsed) by many other pages
  - More so if other pages are themselves important
Link spam

- Link spam: Inflating the rank of a page by creating nepotistic links to it
  - From own sites: Link farms
  - From partner sites: Link exchanges
  - From unaffiliated sites (e.g. blogs, guest books, web forums, etc.)
- The more links, the better
  - Generate links automatically
  - Use scripts to post to blogs
  - Synthesize entire web sites
  - Synthesize many web sites (DNS spam)
- The more important the linking page, the better
  - Buy expired highly-ranked domains
  - Post links to high-quality blogs
Link farms and link exchanges
The trade in expired domains

Looking to drive targeted, long-lasting use of expired domain names?

Every day, tens of thousands of previously registered domains are vacated and once again become available for registration. Many of these previously registered domains have web sites and carry a high link count. Search engine listed domain names attract visitors. Unique visitors, on the other hand, convert to profit.

Domain-Retriever works by querying the given list of expired, on-hand or soon to be expired domains. DR also extracts essential domain data in ranking and availability.

Stop wasting time and money on users on overused mailing lists. Start generating revenue through the use of expired domain names.

The world's top traffic providers generate expired domains. Click here to learn how it can save you money!
Web forum and blog spam

Dr. B.'s Blog
A blog of classroom activities and discussions. A place where rhetoric rocks!

User login
Username:
Password:
Create new account
Request new password

Links
- Archived blog posts
- Academic page
- Home Playlist
- Hearing Marathon Pictures
- C&W 2004 Pictures
- WPA 2005: Alaska Pictures

Who's online
There are currently 1 user and 29 guests online.

Online users:
- db.

Navigation
- Accesories
- books
- content
- search
- categories
- vaccination

Buttons and Such
Technical Glossary

Gaming Presentations
- Conference
- Game Theory
- Just Another Angry Newscast
- Research and Writing: This is what's

Rennies looks like

Found out today that my papers were accepted for both the GLS Conference and the Feminisms and Rhetorics conference. The OLS paper is almost done and is looking at rhetorical representations of race and the FUR paper is on rhetorical representations of gender.

I'm excited. GLS gives me the chance to present my work to other folks who are working in the same area. Two days of papers on video games. I'll try to justify myself. Trackback URL for this post:
http://www.english.purdue.edu/blogs/blog/4121

Roulette
- roulette
- from roulette on Fri, 10/9/2005 - 9:11pm
- You are invited to check out the sites in the field of blackjack internet casino
- riverbelle online casino
- from riverbelle casino on Sat, 10/8/2005 - 7:29pm
- You are invited to check out some relevant information in the field of video poker games on cd from canada
- premio online
- from premio online on Fri, 10/7/2005 - 5:58pm
- You can also check the pages on poker download video poker portals
- buy cheap online xanaz
- from buy cheap online xanaz on Thu, 10/6/2005 - 9:03am
- You can also take a look at the pages on doctor online prescription
- ganar dinero internet
- from ganar dinero internet on Mon, 10/3/2005 - 8:10am
- In your free time, check out the sites in the field of online video poker game
- discover card
- from discover card on Sat, 10/1/2005 - 4:35pm
- In your free time, check some information on home loans mortgage rates bad credit loans
- rebel strip poker downloaded
- from rebel strip poker downloaded on Fri, 9/29/2005 - 3:36pm
- Please check out some relevant information dedicated to Free Poker Tournaments
- more
Features identifying link spam

- Large number of links from low-ranked pages
- Discrepancy between number of links (peer endorsement) and number of visitors (user endorsement)
- Links mostly from affiliated pages
  - Same web site; same domain
  - Same IP address
  - Same owner (according to WHOIS record)
- Evidence that linking pages are machine-generated
- …
Cloaking

- Cloaking: The practice of sending different content to search engines than to users

- Techniques:
  - Recognize page request is from search engine (based on “user-agent” info or IP address)
  - Make some text invisible (i.e. black on black)
  - Use CSS to hide text
  - Use JavaScript to rewrite page
  - Use “meta-refresh” to redirect user to other page

- Hard (but not impossible) for SE to detect
How well does web spam detection work?

- Experiment done at MSR-SVC:
  - (joint work with Fetterly, Manasse, Ntoulas)
  - using a number of the features described earlier
  - fed into C4.5 decision-tree classifier
  - corpus of about 100 million web pages
  - judged set of 17170 pages (2364 spam, 14806 non-spam)
  - 10-fold cross-validation

- Our results are **not** indicative of spam detection effectiveness of MSN Search!
How well does web spam detection work?

- Confusion matrix:

<table>
<thead>
<tr>
<th>classified as</th>
<th>spam</th>
<th>non-spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>1,918</td>
<td>446</td>
</tr>
<tr>
<td>non-spam</td>
<td>367</td>
<td>14,439</td>
</tr>
</tbody>
</table>

- Expressed as precision-recall matrix:

<table>
<thead>
<tr>
<th>class</th>
<th>recall</th>
<th>precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>81.1%</td>
<td>83.9%</td>
</tr>
<tr>
<td>non-spam</td>
<td>97.5%</td>
<td>97.0%</td>
</tr>
</tbody>
</table>
Questions

http://research.microsoft.com/aboutmsr/labs/siliconvalley/