


<isweb> ISWeb – Information Systems & Semantic Web
University of Koblenz • Landau

Chapter 4

Web Spam & Advertising

Sergej Sizov
Information Retrieval
Summer term 2008



7.1 Web Spam

..not just for email anymore

Users follow search results

- Money follows users... Spam follows money...

There is value in getting ranked high

- Funnel traffic from SEs to Amazon/eBay/...

Make a few bucks

- Funnel traffic from SEs to a Viagra seller

Make \$6 per sale

- Funnel traffic from SEs to a porn site

Make \$20-\$40 per new member

- Affiliate programs

Course Information Retrieval Summer Term 2008 Sergej Sizov Chapter 4 – Web Spam & Advertising 4-2

Web Spam: Motivation

Let's do the math..


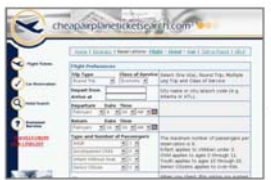
- Assume 500M searches/day on the web
- All search engines combined
- Assume 5% commercially viable

Much more if you include „adult-only“ queries

- Assume \$0.50 made per click (from 5c to \$40)
- \$12.5M/day or about \$4.5 Billion/year

Course Information Retrieval Summer Term 2008 Sergej Sizov Chapter 4 – Web Spam & Advertising 4-3

Spam: defeating IR

Keyword stuffing and cloaking
Crawlers declare that it is a SE spider
They dish us an "optimized" page
Users see a completely different page

But
easy to detect for SE :
just detect
keyword density

Course Information Retrieval Summer Term 2008 Sergej Sizov Chapter 4 – Web Spam & Advertising 4-4


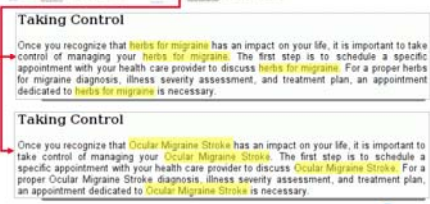
Spam: query flooding



easy to detect for SE :
just detect the page is not about the query

Course Information Retrieval Summer Term 2008 Sergej Sizov Chapter 4 – Web Spam & Advertising 4-5

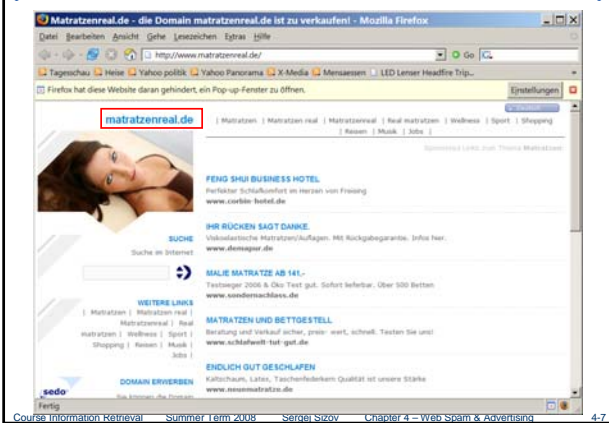
Spam: defeating IR/NLP

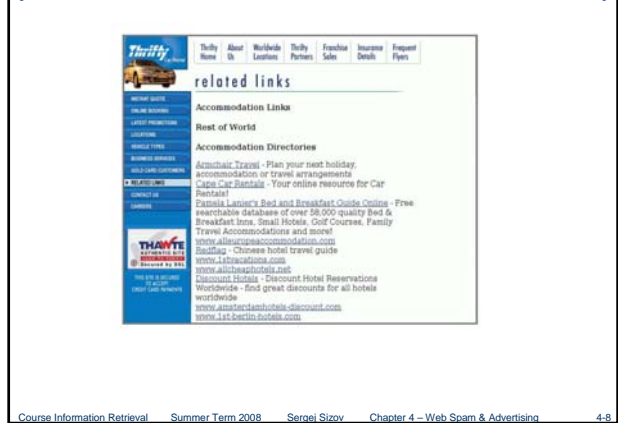
Ideally, links should help:
no one should link to these bad sites...

Course Information Retrieval Summer Term 2008 Sergej Sizov Chapter 4 – Web Spam & Advertising 4-6

Getting links: grabbing expired domains



Getting links: link exchange



Getting links: Mailing Lists



Getting Links: Guestbooks



Web Spam: Summary

Content spam:

- repeat words (boost tf)
- weave words/phrases into copied text
- manipulate anchor texts

Link spam:

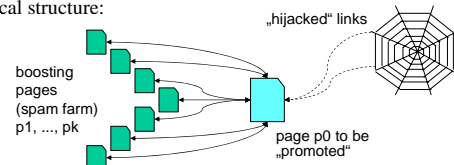
- copy links from Web dir. and distort
- create honeypot page and sneak in links
- infiltrate Web directory
- purchase expired domains
- generate posts to Blogs, message boards, etc.
- build & run spam farm (collusion) + form alliances

Hide/cloak the manipulation:

- masquerade href anchors
- use tiny anchor images with background color
- generate different dynamic pages to browsers and crawlers

Link Spam: General Scenario

Typical structure:



Web transfers to p0 the „hijacked“ score mass („leakage“)

$$\lambda = \sum_{q \in \text{IN}(p_0) - \{p_1, \dots, p_k\}} \text{PR}(q) / \text{outdegree}(q)$$

Theorem:

p0 obtains the following PR authority:

$$\text{PR}(p_0) = \frac{1}{1 - (1 - \epsilon)^2} \left((1 - \epsilon)\lambda + \frac{\epsilon((1 - \epsilon)k + 1)}{n} \right)$$

The above spam farm is optimal within some family of spam farms (e.g. letting hijacked links point to boosting pages).

Link Spam: Google bombs (1) – George W. Bush

The screenshot shows a Google search for "miserable failure" in Mozilla Firefox. The search results page displays several links, with the top result being "President of the United States, George W. Bush" from the official White House website. The search bar shows the query "miserable failure" and the results are sorted by relevance.

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-13

Link Spam: Google bombs (2) - Hommingberger Gepardenforelle

The screenshot shows a search for "c1-SEO-Wettbewerb" in Mozilla Firefox. The search results page displays a result for "c1-SEO-Wettbewerb" from the website "www.hommingberger.de". The page content includes a large image of a trout and text describing a competition for the best SEO website in the region.

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-14

Spam Countermeasures

Basic Ideas:

- compute negative propagation of blacklisted pages (BadRank)
- compute positive propagation of trusted pages (TrustRank)
- detect spam pages based on statistical anomalies
- inspect PR distribution in graph neighborhood (SpamRank)
- learn spam vs. ham based on page and page-context features
- spam mass estimation (fraction of PR that is undeserved)
- probabilistic models for link-based authority (overcome the discontinuity from 0 outlinks to 1 outlink)

BadRank and TrustRank

BadRank:

start with explicit set B of blacklisted pages
define random-jump vector r by setting $r_i = 1/|B|$ if $i \in B$ and 0 else
propagate BadRank mass to predecessors

$$BR(p) = \beta r_p + (1 - \beta) \sum_{q \in \text{OUT}(p)} BR(q) / \text{indegree}(q)$$

TrustRank:

start with explicit set T of trusted pages with trust values t_i
define random-jump vector r by setting $r_i = t_i$ if $i \in T$ and 0 else
propagate TrustRank mass to successors

$$TR(q) = \tau r_q + (1 - \tau) \sum_{p \in \text{IN}(q)} TR(p) / \text{outdegree}(p)$$

Problems:

maintenance of explicit lists is difficult
difficult to understand (& guarantee) effects

7.2 Web Advertising

Banner ads (1995-2001)

- Initial form of web advertising
- Popular websites charged X\$ for every 1000 "impressions" of ad
 - Called "CPM" rate
 - Modeled similar to TV, magazine ads
- Untargeted to demographically targeted
- Low clickthrough rates
 - low ROI for advertisers

Performance-based advertising

Introduced by Overture around 2000

- Advertisers "bid" on search keywords
- When someone searches for that keyword, the highest bidder's ad is shown
- Advertiser is charged only if the ad is clicked on

Similar model adopted by Google with some changes around 2002

- Called "Adwords"

Ads vs. Search Results

Web Results 1 - 10 of about 2,230,000 for geico. (0.04 sec)

GEICO Car Insurance. Get an auto insurance quote and save today...
 GEICO auto insurance, online car insurance quote, motorcycle insurance quote, online insurance sales and service from a leading insurance company.
[www.geico.com/~21k - Sep 22, 2005 - Cached - Similar pages](#)
[Auto Insurance](#) - [Buy Auto Insurance](#)
[Contact Us](#) - [Make a Payment](#)
[More Results from www.geico.com »](#)

GEICO Google Settle Trademark Dispute
 The case was resolved out of court, so advertisers are still left without legal guidance on use of trademarks within ads or as keywords.
[www.clickz.com/news/article.php?9547356-44k - Cached - Similar pages](#)

Google and GEICO settle AdWords dispute | The Register
 Google and car insurance firm GEICO have settled a trade mark dispute over ... Car insurance firm GEICO sued both Google and Yahoo! subsidiary Overture W ...
[www.theregister.co.uk/2005/09/09/google_geico_settlement/#.21k - Cached - Similar pages](#)

GEICO v. Google
 ... involving a lawsuit filed by Government Employees Insurance Company (GEICO), GEICO has filed suit against two major Internet search engine operators, ...
[www.consumeraffairs.com/news/04/geico_google.html - 15k - Cached - Similar pages](#)

Sponsored Links

Great Car Insurance Rates
 Simplify Buying Insurance at Safeco
 See Your Rate with an Instant Quote
[www.Safeco.com](#)

Free Insurance Quotes
 Fill out one simple form to get multiple quotes from local agents.
[www.HometownQuotes.com](#)

5 Free Quotes. 1 Form.
 Get 5 Free Quotes in Minutes!
 You Have Nothing To Lose. It's Free
[aggressoftware.com/insurance/Missouri](#)

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-19

Web Advertising: Questions

Performance-based advertising works!

- Multi-billion-dollar industry

Interesting problems

- What ads to show for a search?
- If I'm an advertiser, which search terms should I bid on and how much to bid?

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-20

Adwords problem

A stream of queries arrives at the search engine

- q1, q2, ...

Several advertisers bid on each query

When query q_i arrives, search engine must pick a subset of advertisers whose ads are shown

Goal: maximize search engine's revenues

Clearly we need an online algorithm!

Simplest algorithm is greedy..

.. the greedy algorithm is actually optimal!

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-21

Advertising: Justification (1)

Each ad has a different likelihood of being clicked

- Advertiser 1 bids \$2, click probability = 0.1
- Advertiser 2 bids \$1, click probability = 0.5

Clickthrough rate measured historically

Simple solution

- Instead of raw bids, use the "expected revenue per click"

Each advertiser has a limited budget

- Search engine guarantees that the advertiser will not be charged more than their daily budget

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-22

Advertising: Simplified Model

Assume all bids are 0 or 1

Each advertiser has the same budget B

One advertiser per query

Let's try the greedy algorithm

- Arbitrarily pick an eligible advertiser for each keyword

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-23

Bad scenario for greedy

Two advertisers A and B

A bids on query x, B bids on x and y

Both have budgets of \$4

Query stream: xxxxyyy

- Worst case greedy choice: BBBB__
- Optimal: AAAABBBB
- Competitive ratio = 1/2

.. formal analysis shows this is the worst case

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-24

BALANCE algorithm [MSVV]

[Mehta, Saberi, Vazirani, and Vazirani]

For each query, pick the advertiser with the largest unspent budget

- Break ties arbitrarily

Two advertisers A and B

A bids on query x, B bids on x and y

Both have budgets of \$4

Query stream: xxxxyyy

BALANCE choice: ABABBB_

- Optimal: AAAABBBB

Competitive ratio = $\frac{3}{4}$

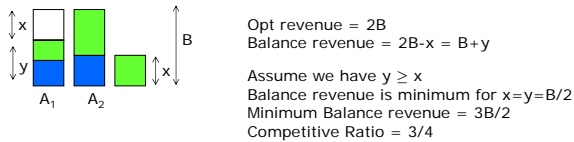
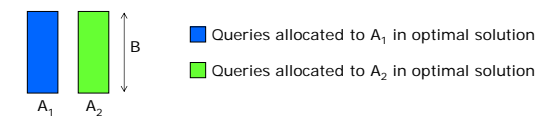
Analyzing BALANCE

Consider simple case: two advertisers, A_1 and A_2 , each with budget B (assume $B \gg 1$)

Assume optimal solution exhausts both advertisers' budgets
BALANCE must exhaust at least one advertiser's budget

- If not, we can allocate more queries
- Assume BALANCE exhausts A_2 's budget

Analyzing Balance



General Result

In the general case, worst competitive ratio of BALANCE is $1 - 1/e = \text{approx. } 0.63$

Interestingly, no online algorithm has a better competitive ratio
Won't go through the details here, but let's see the worst case that gives this ratio

Worst case for BALANCE

N advertisers, each with budget $B \gg N \gg 1$

NB queries appear in N rounds of B queries each

Round 1 queries: bidders A_1, A_2, \dots, A_N

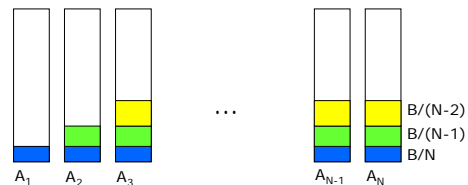
Round 2 queries: bidders A_2, A_3, \dots, A_N

Round i queries: bidders A_i, \dots, A_N

Optimum allocation: allocate round i queries to A_i

- Optimum revenue NB

BALANCE allocation



After k rounds, sum of allocations to each of bins A_k, \dots, A_N is $S_k = S_{k+1} = \dots = S_N = \sum_{1 \leq i \leq k} B/(N-i+1)$

If we find the smallest k such that $S_k \geq B$, then after k rounds we cannot allocate any queries to any advertiser

BALANCE analysis

$B/1 \quad B/2 \quad B/3 \quad \dots \quad B/(N-k+1) \quad \dots \quad B/(N-1) \quad B/N$
 $\leftarrow S_1 \rightarrow$
 $\leftarrow S_2 \rightarrow$
 $S_k = B$

$1/1 \quad 1/2 \quad 1/3 \quad \dots \quad 1/(N-k+1) \quad \dots \quad 1/(N-1) \quad 1/N$
 $\leftarrow S_1 \rightarrow$
 $\leftarrow S_2 \rightarrow$
 $S_k = 1$

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-31

BALANCE analysis

Fact: $H_n = \sum_{1 \leq i \leq n} 1/i = \text{approx. } \log(n)$ for large n

- Result due to Euler

$1/1 \quad 1/2 \quad 1/3 \quad \dots \quad 1/(N-k+1) \quad \dots \quad 1/(N-1) \quad 1/N$
 $\leftarrow \log(N) \rightarrow$
 $\leftarrow \log(N)-1 \rightarrow \leftarrow S_k = 1 \rightarrow$

$S_k = 1$ implies $H_{N-k} = \log(N)-1 = \log(N/e)$
 $N-k = N/e$
 $k = N(1-1/e)$

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-32

BALANCE analysis

So after the first $N(1-1/e)$ rounds, we cannot allocate a query to any advertiser

Revenue = $BN(1-1/e)$

Competitive ratio = $1-1/e$

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-33

General version of problem

Arbitrary bids, budgets

Consider query q , advertiser i

- Bid = x_i
- Budget = b_i

BALANCE can be terrible

- Consider two advertisers A_1 and A_2
- A_1 : $x_1 = 1, b_1 = 110$
- A_2 : $x_2 = 10, b_2 = 100$

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-34

Generalized BALANCE

Arbitrary bids; consider query q , bidder i

- Bid = x_i
- Budget = b_i
- Amount spent so far = m_i
- Fraction of budget left over $f_i = 1 - m_i/b_i$
- Define $\psi_i(q) = x_i(1 - e^{-f_i})$

Allocate query q to bidder i with largest value of $\psi_i(q)$

Same competitive ratio $(1-1/e)$

Course Information Retrieval Summer Term 2008 Sergei Sizov Chapter 4 – Web Spam & Advertising 4-35