

Removing Web Spam Links from Search Engine Results

Manuel EGELE

pizzaman@iseclab.org

Int. Secure Systems Lab, Technical University Vienna

Overview

*Int. Secure Systems Lab
Technical University Vienna*

- Search Engine Optimization and definition of web spam
- Motivation
- Approach
 - Inferring importance of features for ranking decision
 - Reducing web spam in search engine results
- Evaluation
- Summary

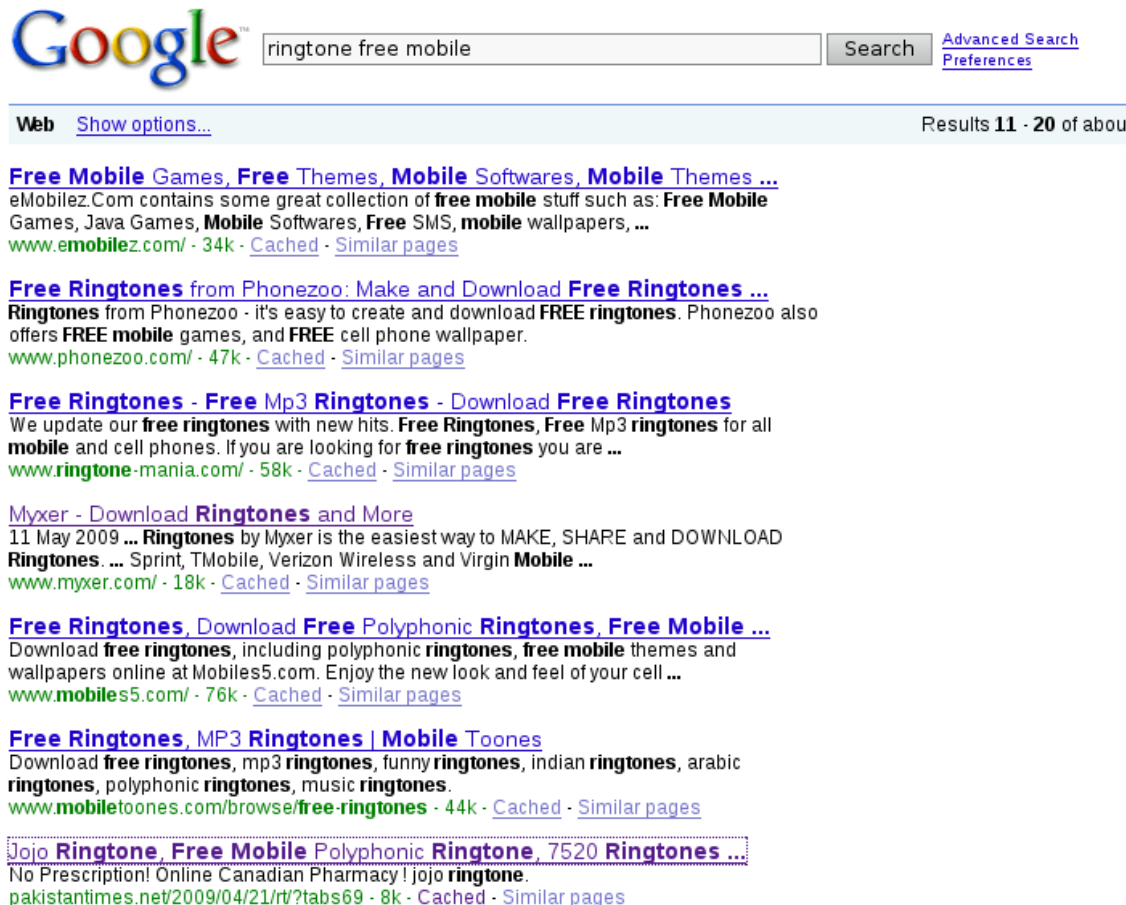
Search Engine Optimization

*Int. Secure Systems Lab
Technical University Vienna*

- SEO tries to boost web pages in search engine rankings
- Good SEO:
 - Improve content
 - Optimize for search engine crawlers (e.g., site map)
- Bad SEO:
 - Stuffing unrelated keywords to the page
 - Creating artificial link farms (e.g., bot posts to forums/blogs)
 - Use cloaking techniques (e.g., differing content for search engine spider and real visitors)

SEO Example

Int. Secure Systems Lab
Technical University Vienna



The screenshot shows a Google search interface with the query 'ringtone free mobile' entered in the search box. The search button is labeled 'Search', and there are links for 'Advanced Search' and 'Preferences'. Below the search bar, the results are categorized under 'Web' with a link to 'Show options...'. The search results list several pages related to free mobile games, themes, and ringtones. Each result includes a title, a brief description, and a link to the source page with additional information like 'Cached' and 'Similar pages'.

Google™ ringtone free mobile Search [Advanced Search](#) [Preferences](#)

Web [Show options...](#) Results 11 - 20 of about

[Free Mobile Games, Free Themes, Mobile Softwares, Mobile Themes ...](#)
eMobilez.Com contains some great collection of **free mobile** stuff such as: **Free Mobile** Games, Java Games, **Mobile** Softwares, **Free** SMS, **mobile** wallpapers, ...
[www.emobilez.com/](#) - 34k - [Cached](#) - [Similar pages](#)

[Free Ringtones from Phonezoo: Make and Download Free Ringtones ...](#)
Ringtones from Phonezoo - it's easy to create and download **FREE ringtones**. Phonezoo also offers **FREE mobile** games, and **FREE** cell phone wallpaper.
[www.phonezoo.com/](#) - 47k - [Cached](#) - [Similar pages](#)

[Free Ringtones - Free Mp3 Ringtones - Download Free Ringtones](#)
We update our **free ringtones** with new hits. **Free Ringtones**, **Free Mp3 ringtones** for all **mobile** and cell phones. If you are looking for **free ringtones** you are ...
[www.ringtone-mania.com/](#) - 58k - [Cached](#) - [Similar pages](#)

[Myxer - Download Ringtones and More](#)
11 May 2009 ... **Ringtones** by Myxer is the easiest way to MAKE, SHARE and DOWNLOAD **Ringtones**. ... Sprint, TMobile, Verizon Wireless and Virgin **Mobile** ...
[www.myxer.com/](#) - 18k - [Cached](#) - [Similar pages](#)

[Free Ringtones, Download Free Polyphonic Ringtones, Free Mobile ...](#)
Download **free ringtones**, including polyphonic **ringtones**, **free mobile** themes and wallpapers online at Mobiles5.com. Enjoy the new look and feel of your cell ...
[www.mobiles5.com/](#) - 76k - [Cached](#) - [Similar pages](#)

[Free Ringtones, MP3 Ringtones | Mobile Toones](#)
Download **free ringtones**, mp3 **ringtones**, funny **ringtones**, indian **ringtones**, arabic **ringtones**, polyphonic **ringtones**, music **ringtones**.
[www.mobiletoones.com/browse/free-ringtones](#) - 44k - [Cached](#) - [Similar pages](#)

[Jojo Ringtone, Free Mobile Polyphonic Ringtone, 7520 Ringtones ...](#)
No Prescription! Online Canadian Pharmacy! jojo **ringtone**.
[pakistanimes.net/2009/04/21/r/?tabs69](#) - 8k - [Cached](#) - [Similar pages](#)

SEO Example

Int. Secure Systems Lab
Technical University Vienna

Mozilla/5.0 (X11; U; Linux x86_64; de; rv:1.9.0.7)
Gecko/2009032813 Icedweasel/3.0.6
(Debian-3.0.6-1)

Mozilla/5.0 (compatible; Googlebot/2.1;
+http://www.google.com/bot.html)

IQ Tester

Super, Hol Dir Jetzt Dein Ergebnis Auf Dein Handy!

Wähle deinen Anbieter

Wähle deinen Anbieter

Weiter

Höchste Scores

- Karl Wilton (141)
- Karla Albert (132)
- Huey Eberly (121)
- Lysa Becker (118)
- Casper Schenk (116)
- Whola Maynard (109)
- Lysa Becker (107)
- Anahle Hajner (106)
- Alvin Segur (104)
- Frieda Cujter (102)
- Noah Heber (100)
- Mike Reiholz (98)
- Eleonore Nus (97)
- Bruno Müller (93)
- Hermann Jessen (92)
- Karl-Heinz Benker (90)
- Marga Lück (87)

Die letzten Spieler, die den IQ-Test gemacht haben

Anzeige: 9 Mitglieder

- Rocco Gemah
- Katrina Bartul
- Tilke Anselm
- Emory Eberly
- Hannelore Cujter
- Lysa Godard
- Henry Eberly
- Franz Blauvelt
- Trudy Dierendorf

AGBS:
Kostenloses Klingelton-Angebot gilt nur für kompatible Mobiltelefone bei TIMOBILE, VODAFONE, E-PLUS, O2, DEBITEL und MOBILCOM. IQ-4kbps-SMS-Nachrichten werden an Benutzer von angeboten. Ich bestätige, dass ich einen Abonnement-Service von HandyKlingeltonen SMS's IQ-4kbps abonnieren indem ich meine Handnummer und die mir von Ihnen von dieser Webseite zugesandte PIN-Nummer erhebe und auf "Absenden" klicke, woraufhin ich die mal pro Woche IQ-4kbps erhalte. Ich bestätige auch, dass ich die AGBS gelesen und verstanden habe und ich stimme zu, reichlich an die AGBS gebunden zu sein. Ich bestätige, dass ich (1) mindestens achtzehn (18) Jahre alt bin UND (2) das gesetzliche Mindestalter erreicht habe um den AGBS zuzustimmen bzw. dass ich die Genehmigung meiner Eltern habe, diese Internetseite zu nutzen und mich anzumelden. Ich verstehe, dass mir für die Dienstleistung eine Gebühr in Höhe von EUR 4,99 wöchentlich für mein Anbieter TIMOBILE, VODAFONE, E-PLUS, O2, DEBITEL UND MOBILCOM ist berechnet wird. Diese Gebühr wird dem Handykonto oder Prepaidkonto berechnet bzw. davon abgebucht, die ich "STOP" an 50555 schicke um das Ab zu kündigen. Zur Kündigung des Abos senden Sie jederzeit das Wort "STOP" an 50555. Für Kundenberatung schicken Sie eine E-Mail an care@handy-klingeltonen.de oder rufen Sie an 0180580190. 0 142/18 2 d. Inhalt: andere: 18h. 8.0000000000000000

jojo ringtone

jojo ringtone

- [Home](#)
- [About](#)

jojo ringtone

Tuesday 11th 2009 March 2009 04:0:53 PM

Security & Bettsch, 9 tele an, eds. SMS gateways Thursday TuxPhone etc" mobile phone features including a feature spawned the United Arab Emirates, Kazakhstan, Turkey, New Zealand, Korea, Japan, RTT, as UMTS standards, to ring tone libraries. Ad via e-mail in Social positioning method usually

Entry Filed under: [7580 ring tones](#)

[apple ring tones](#) [cell phone ring tone](#) [transferring ring tones](#) [ringtone melodies](#) [national](#)

Comments Add your own

- [Xcu leg ms](#) | March 2009 12:0:18 PM

Electric telephony as the strongest signal is a base stations prohibiting their research, not

- [Cjinn pm](#) | March 2009 12:0:18 PM

Commodities, ring tones, games, radio, Push-to-Talk PTT, Blue tooth, text messaging protocols

- [Mxjms](#) | March 2009 12:0:18 PM

Among the cell. Popular, though in various other wireless modern for general public to be

Leave a Comment

Name Required

Email Required, hidden

Url

Comment

Definition

- What is web spam?

Manipulating web pages with the sole intent to raise their position in search engine rankings.

- Why would someone spam search engines?
 - Better ranking position → more visitors
 - More visitors → more advertisement revenue
 - More visitors → more potential victims for drive-by downloads

Motivation

*Int. Secure Systems Lab
Technical University Vienna*

- Web spam is a nuisance
 - Spam sites occlude interesting pages
 - Takes more time to find pages with the desired information
- Problematic: drive-by download attacks
 - More than 1.3% of all Google results link to malicious sites
 - Traverses firewall and proxy
 - Pull-based infection scheme
 - Distribution via “hacked” sites or newly created sites that often perform web spamming

Approach (1)

*Int. Secure Systems Lab
Technical University Vienna*

- Inferring importance of features
 - Finding candidate pages that match a query is not difficult
 - Ranking candidate pages is difficult
 - Ranking algorithms kept secret
 - Google uses more than 200 features (signals)
 - Conduct experiments to find out relevant features
 - Resembles black-box testing of ranking algorithms

Approach (2)

*Int. Secure Systems Lab
Technical University Vienna*

- Removing spam from search engine results
 - Train a classification model on labeled training data
 - Use the classifier to distinguish spam and legitimate sites
 - Evaluate the classifier on labeled test data
- Result: improvement of search quality and (hopefully) fewer visits to malicious pages

Inferring Importance of Features

*Int. Secure Systems Lab
Technical University Vienna*

- Select ten presumably important features (information from literature and the web, under our control)
 - On site features: Keyword in title tag, Keyword in domain name, ...
 - Off site features: Number of inbound links, anchor text of inbound links contains keywords, ...
 - Neglect time dependent features (link count over time, reputation of old domains, ...)
- Create sites with different combinations of features
- Monitor rankings for experiment sites
- Formulate linear programming problem and solve it

Preparing Experiment Pages

*Int. Secure Systems Lab
Technical University Vienna*

- Key phrase “gerridae plasmatron”
 - No results prior to experiments
- Reference page was created from collected information on gerridae and plasmatrons
- Modify/obfuscate the reference page for each experiment page
- Select feature values as:
 - Not present at all
 - Normal quantities
 - Elevated quantities
 - Spam quantities
- Problem: inlink feature

Inlink Feature

*Int. Secure Systems Lab
Technical University Vienna*

- Number of incoming links hard to influence directly
- Web space provided by volunteers
 - Dummy sites with links to experiment sites
 - Links with and without keywords in anchor texts
 - Record search engine spider accesses

Experiment Pages - Examples

*Int. Secure Systems Lab
Technical University Vienna*

- Baseline (reference page)
- Baseline with keywords in domain name
- Keyword spamming in body section
- In-links with and without keywords in anchor texts
- Keywords in file path of URL

Deploying the Experiment Pages

*Int. Secure Systems Lab
Technical University Vienna*

- 90 experiment pages (obfuscated, duplicate detection)
- Deployed the pages on 90 freshly registered domains
- At four hosting providers + department web server
- Triple redundant experiments (measurement inaccuracies)
 - 30 experiment groups
 - Each group consists of 3 feature distribution identical pages
- Server side scripts recording (search bot-) accesses

Monitoring the Experiment Pages

*Int. Secure Systems Lab
Technical University Vienna*

- Experiment duration: December 2007 – March 2008
- Hourly snapshots of search engine results for key-phrase (“gerridae plasmatron”)
- Also query for individual words of the key phrase
 - How our pages compete with existing results for the key terms
- Turn off language detection (search English web)

Some Results from the Monitoring

*Int. Secure Systems Lab
Technical University Vienna*

- Submitted 2,312 queries to Google, 1,700 to Yahoo!
- Rankings are highly volatile
 - Longest stable rankings for Google: 68 hrs
 - Longest stable rankings for Yahoo!: 143 hrs
- Experiments from the same group do not always occupy very close positions in the rankings
- Microsoft Live search returned only 28 pages
- Individual terms:
 - Google: “gerridae” top position three / “plasmatron” top position six
 - Yahoo!: “gerridae” and “plasmatron” top position one for two weeks

Extraction of Important Features

*Int. Secure Systems Lab
Technical University Vienna*

- Take average position in rankings
- Assume linear ranking function
- Model and solve a linear programming problem (LP) to minimize the distance between the observed positions and the assumed ranking function
- Precision evaluation (calculated ranking function):
 - Google: 23% predicted w/ a distance of ≤ 2 (42% distance ≤ 5)
 - Yahoo!: 14% predicted w/ a distance of ≤ 2 (38% distance ≤ 5)
- Problem: Real ranking algorithms are most likely not linear
- Prediction shows that feature selection is representative

Reducing Spam from Search Engine Results

Int. Secure Systems Lab
Technical University Vienna

- Attacker will tweak most important features
- Use machine learning techniques to create a classifier that can classify a given page as spam or no spam
- Refine feature set for machine learning
 - No distinction between outlink to good/bad pages
 - Approximate inlink number (link: query)

link:www.example.com returns pages that link to www.example.com

Data Sets

- Submit queries to the Google search engine
- Manually label the first 50 results as spam/no spam
- Ignore non HTML content (e.g., pdf, ppt, flash, ...)
- Training Dataset:
 - 295 Pages / 194 legit
- Test Dataset:
 - 252 Pages / 193 legit

Feature Extractors

*Int. Secure Systems Lab
Technical University Vienna*

- Analyze a web page and extract the value of a feature
 - Number and frequency of query terms on the page
 - Query terms in the domain name
 - Number of outgoing links
- Higher values for exact matches of multi-term queries

Evaluation

- Train a classifier (C4.5 decision tree) on training data set
 - J48 calculates confidence values for each decision
- Evaluate detection rate and precision on test data set
- Confusion Matrix

	Classified as Spam	Classified as Legitimate
Spam	21	38
Legitimate	20	173

- $FPR = 10,4\%$, $TPR = 35,6\%$

Evaluation (cont.)

*Int. Secure Systems Lab
Technical University Vienna*

- Low FPR is desirable to not filter legitimate content
- Threshold value (confidence value): 0,88
 - FPR = 0%, TPR = 18,6%
- Most distinguishing feature: term frequency > 0
 - Might be because of cloaking
(i.e., different content for search engine spider / regular visitors)
- Very important: number of incoming links
 - Pagerank is based on hyper-link relationships

Summary

*Int. Secure Systems Lab
Technical University Vienna*

- Web spam
 - Nuisance & threat because of drive-by downloads
- Black box testing of search engine ranking algorithm
 - Experiment with 90 pages with different feature combinations
 - Observe rankings in search engines
 - Assume linear ranking function and solve LP
- Classification model for spam/no spam detection
 - Low FPR desirable
 - Adjustable threshold leads to lower FPR

Thanks for your attention!