SEO and spam

Dan Smith

Overview

- Google
- Search engine optimisation
- Spam and adversarial IR

3

GOOGLE

4

Web search engines

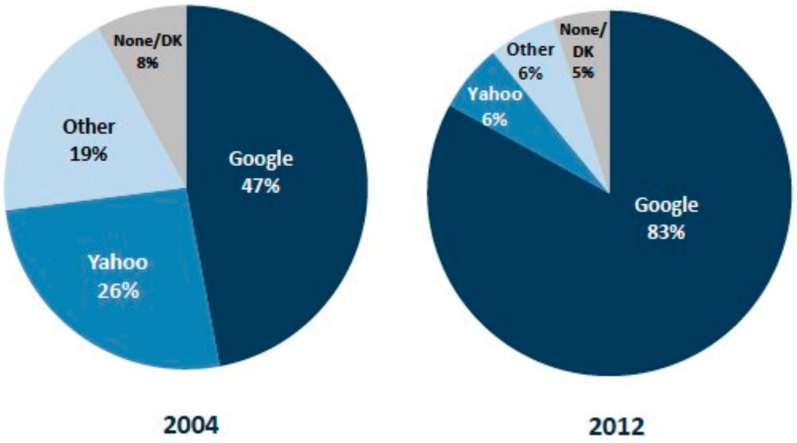
- Google has approx. 90% of the world's web search traffic
  - UK 88%
  - USA 76%
- Other search engines
  - Bing the biggest competitor in USA, globally 3.37%
  - Baidu important in China, globally 0.79%
  - Yahoo! declining, globally 3.43%

http://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

5

The rise of Google

% of search users who answered the question: Which search engine do you use MOST OFTEN?

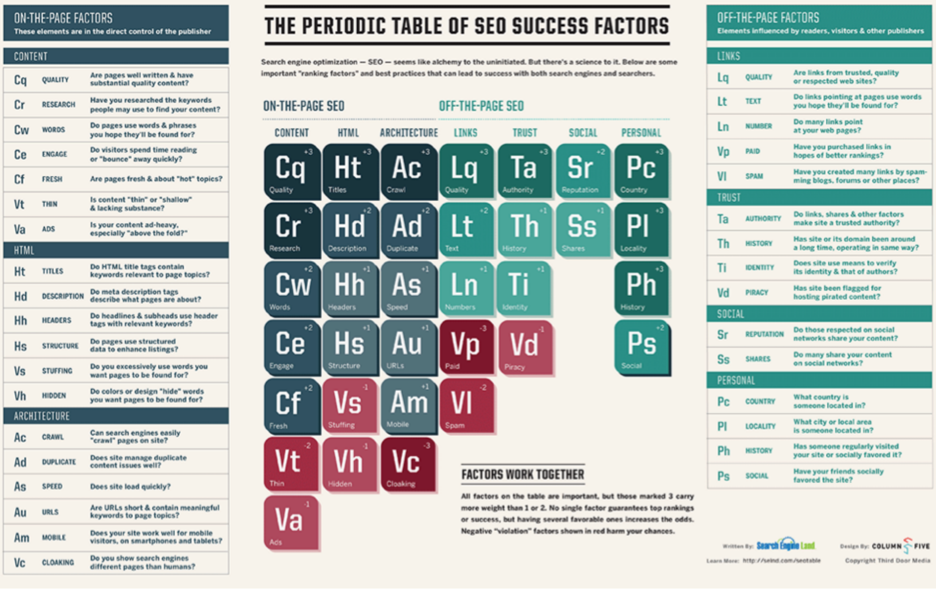


6

Google now: Knowledge Graph

- Major development aimed to give better answers to conversational and question searches
- Based on a large graph of (mostly) named entities with links between them
- Graph is built from search history data
  - e.g. people who search for Impressionists look at pages about Claude Monet

Google ranking factors



ON-THE-PAGE FACTORS

These elements are in the direct control of the publisher

CONTENT		
Cq	QUALITY	Are pages well written & have substantial quality content?
Cr	RESEARCH	Have you researched the keywords people may use to find your content?
Cw	WORDS	Do pages use words & phrases you hope they'll be found for?
Ce	ENGAGE	Do visitors spend time reading or "bounce" away quickly?
Cf	FRESH	Are pages fresh & about "hot" topics?
Vt	THIN	Is content "thin" or "shallow" & lacking substance?
Va	ADS	Is your content ad-heavy, especially "above the fold?"
HTML		
Ht	TITLES	Do HTML title tags contain keywords relevant to page topics?
Hd	DESCRIPTION	Do meta description tags describe what pages are about?
Hh	HEADERS	Do headlines & subheads use header tags with relevant keywords?
Hs	STRUCTURE	Do pages use structured data to enhance listings?
Vs	STUFFING	Do you excessively use words you want pages to be found for?
Vh	HIDDEN	Do colors or design "hide" words you want pages to be found for?
ARCHITECTURE		
Ac	CRAWL	Can search engines easily "crawl" pages on site?
Ad	DUPLICATE	Does site manage duplicate content issues well?
As	SPEED	Does site load quickly?
Au	URLS	Are URLs short & contain meaningful keywords to page topics?
Am	MOBILE	Does your site work well for mobile visitors, on smartphones and tablets?
Vc	CLOAKING	Do you show search engines different pages than humans?

THE PERIODIC TABLE OF SEO SUCCESS FACTORS

Search engine optimization — SEO — seems like alchemy to the uninitiated. But there's a science to it. Below are some important "ranking factors" and best practices that can lead to success with both search engines and searchers.

ON-THE-PAGE SEO			OFF-THE-PAGE SEO			
CONTENT	HTML	ARCHITECTURE	LINKS	TRUST	SOCIAL	PERSONAL
Cq +3 Quality	Ht +3 Titles	Ac +3 Crawl	Lq +3 Quality	Ta +3 Authority	Sr +2 Reputation	Pc +3 Country
Cr +3 Research	Hd +2 Description	Ad +2 Duplicate	Lt +2 Text	Th +1 History	Ss +1 Shares	Pl +3 Locality
Cw +2 Words	Hh +1 Headers	As +1 Speed	Ln +1 Numbers	Ti +1 Identity		Ph +3 History
Ce +2 Engage	Hs +1 Structure	Au +1 URLs	Vp -3 Paid	Vd -1 Piracy		Ps +2 Social
Cf +2 Fresh	Vs -1 Stuffing	Am +1 Mobile	Vl -2 Spam			
Vt -2 Thin	Vh -1 Hidden	Vc -3 Cloaking				
Va -1 Ads						

FACTORS WORK TOGETHER

All factors on the table are important, but those marked 3 carry more weight than 1 or 2. No single factor guarantees top rankings or success, but having several favorable ones increases the odds. Negative "violation" factors shown in red harm your chances.

OFF-THE-PAGE FACTORS

Elements influenced by readers, visitors & other publishers

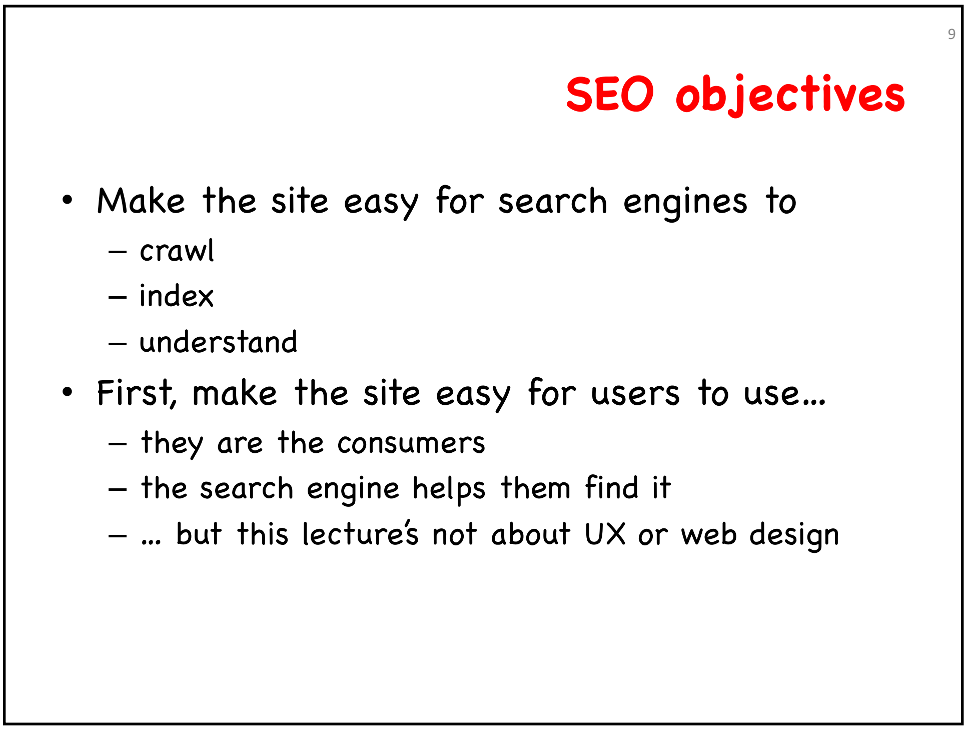
LINKS		
Lq	QUALITY	Are links from trusted, quality or respected web sites?
Lt	TEXT	Do links pointing at pages use words you hope they'll be found for?
Ln	NUMBER	Do many links point at your web pages?
Vp	PAID	Have you purchased links in hopes of better rankings?
Vl	SPAM	Have you created many links by spamming blogs, forums or other places?
TRUST		
Ta	AUTHORITY	Do links, shares & other factors make site a trusted authority?
Th	HISTORY	Has site or its domain been around a long time, operating in same way?
Ti	IDENTITY	Does site use means to verify its identity & that of authors?
Vd	PIRACY	Has site been flagged for hosting pirated content?
SOCIAL		
Sr	REPUTATION	Do those respected on social networks share your content?
Ss	SHARES	Do many share your content on social networks?
PERSONAL		
Pc	COUNTRY	What country is someone located in?
Pl	LOCALITY	What city or local area is someone located in?
Ph	HISTORY	Has someone regularly visited your site or socially favored it?
Ps	SOCIAL	Have your friends socially favored the site?

Written By: Search Engine Land Design By: COLUMN FIVE

Learn More: http://selnd.com/seotable Copyright Third Door Media

How to make search engines like your pages

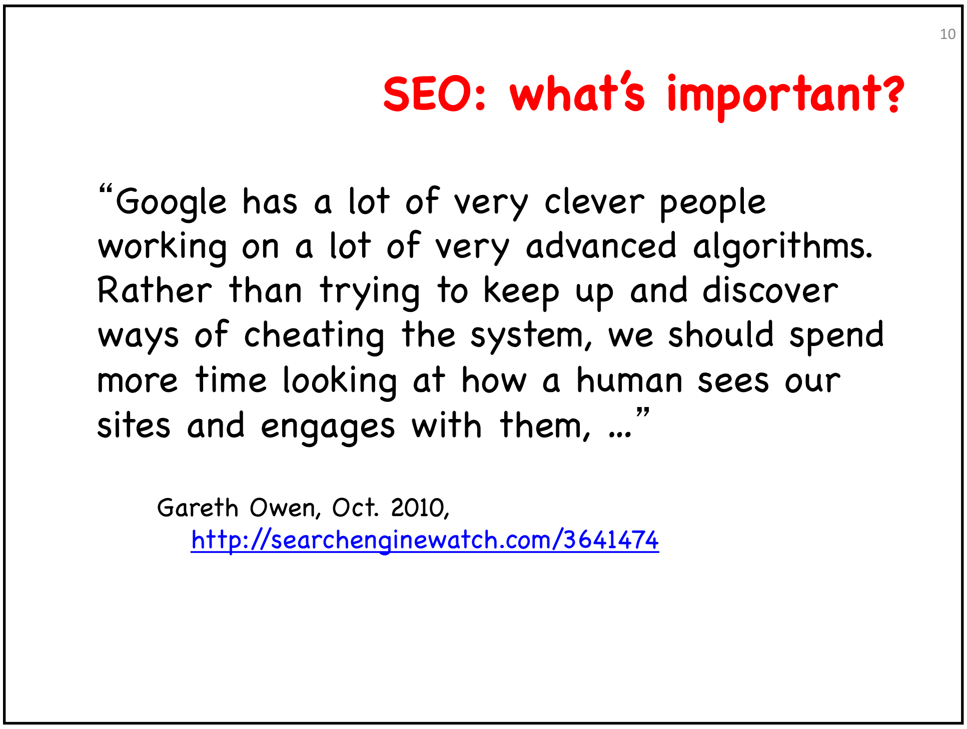
SEARCH ENGINE OPTIMISATION



9

SEO objectives

- Make the site easy for search engines to
  - crawl
  - index
  - understand
- First, make the site easy for users to use…
  - they are the consumers
  - the search engine helps them find it
  - … but this lecture's not about UX or web design



10

SEO: what's important?

"Google has a lot of very clever people working on a lot of very advanced algorithms. Rather than trying to keep up and discover ways of cheating the system, we should spend more time looking at how a human sees our sites and engages with them, …"

Gareth Owen, Oct. 2010,
http://searchenginewatch.com/3641474

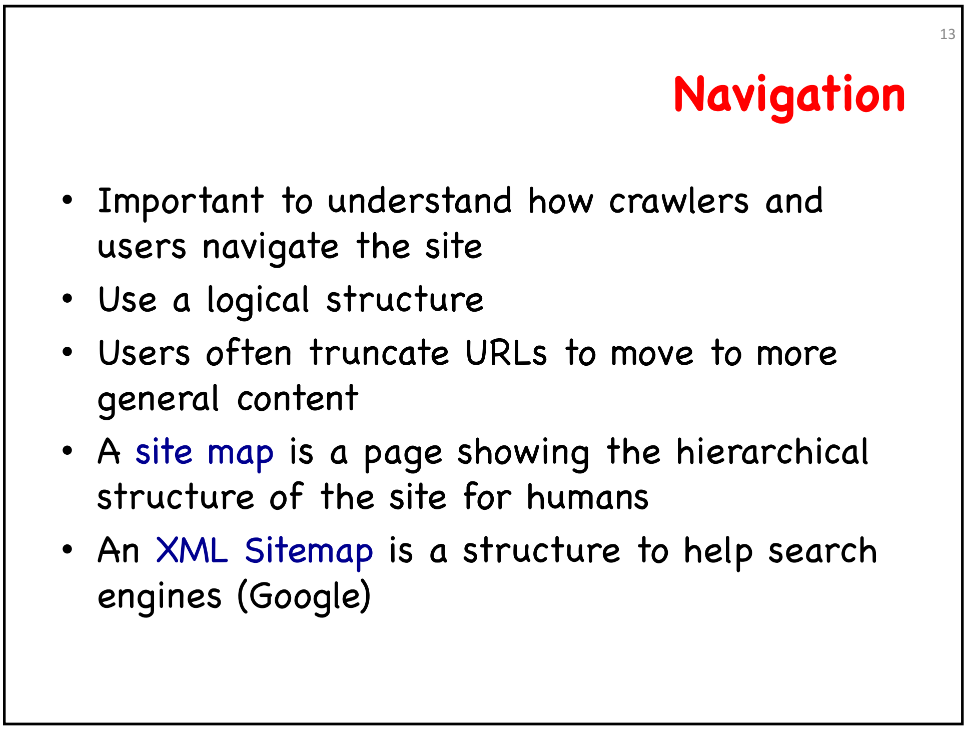
Title and Meta tags

- Use brief, descriptive titles
  - each page (or small group of pages) should have a separate title
  - it's what people see in the search results
- Use the description meta tag to provide a concise summary of the page
  - it may be used as the snippet for the page

URLs

- Crawlers use URLs to navigate the site
- Users use URLs to infer the site structure
- Other sites may use URLs as the displayed link text
- URLs are displayed in search results

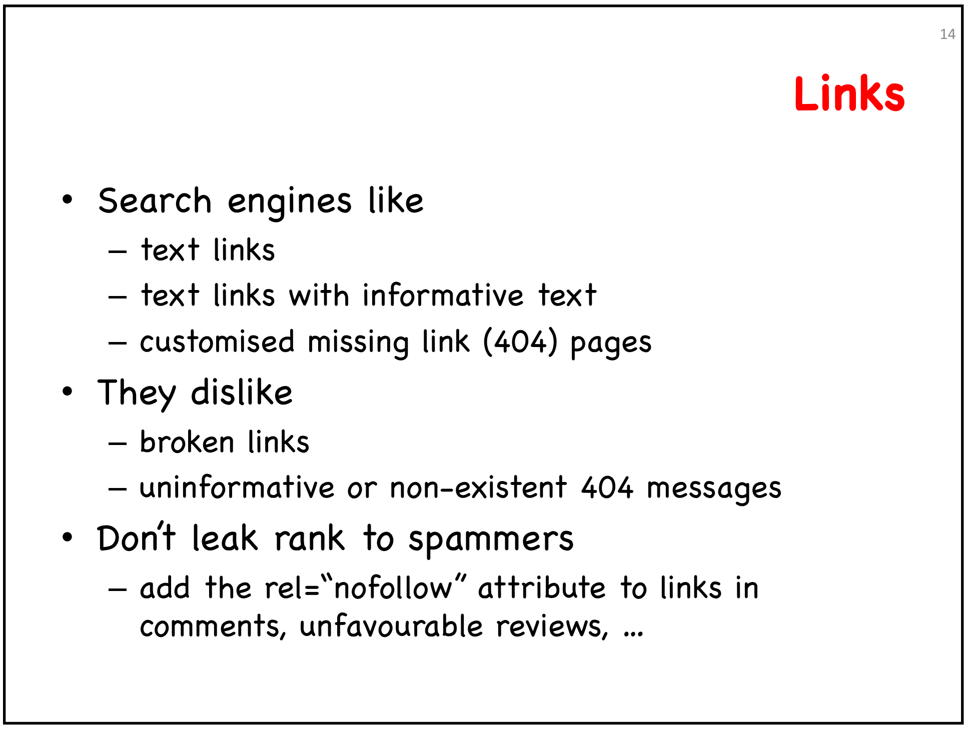
- Search engines prefer shorter URLs with recognisable words
  - dislike long numeric or mixed identifiers, session identifiers, …
  - use a single URL per page, as multiple URLs dilute the rank



13

Navigation

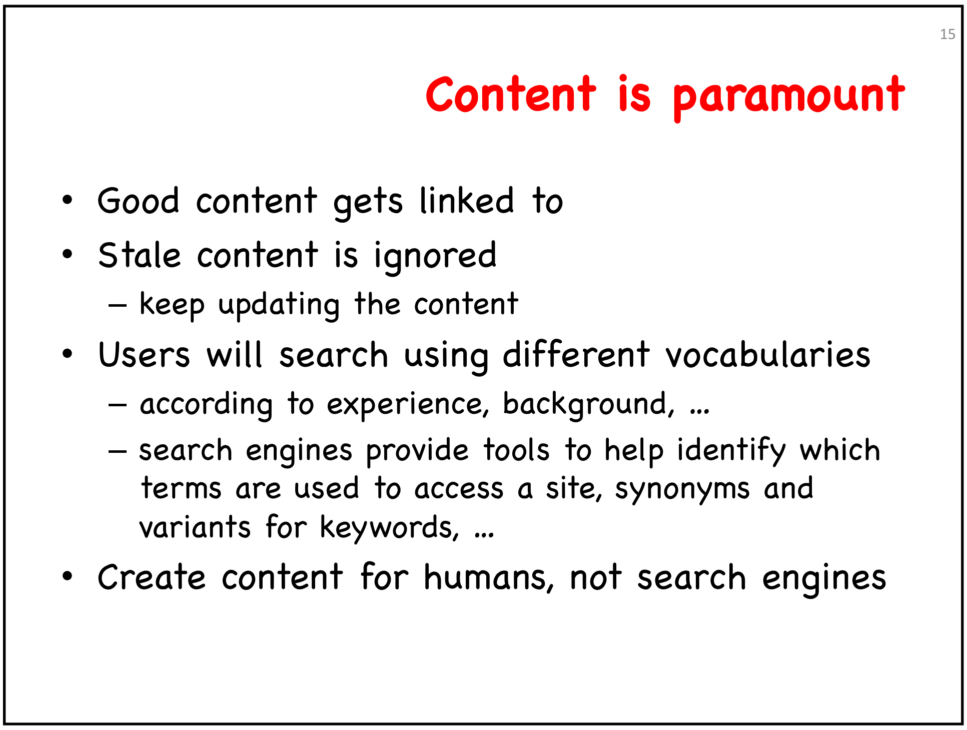
- Important to understand how crawlers and users navigate the site
- Use a logical structure
- Users often truncate URLs to move to more general content
- A site map is a page showing the hierarchical structure of the site for humans
- An XML Sitemap is a structure to help search engines (Google)



14

Links

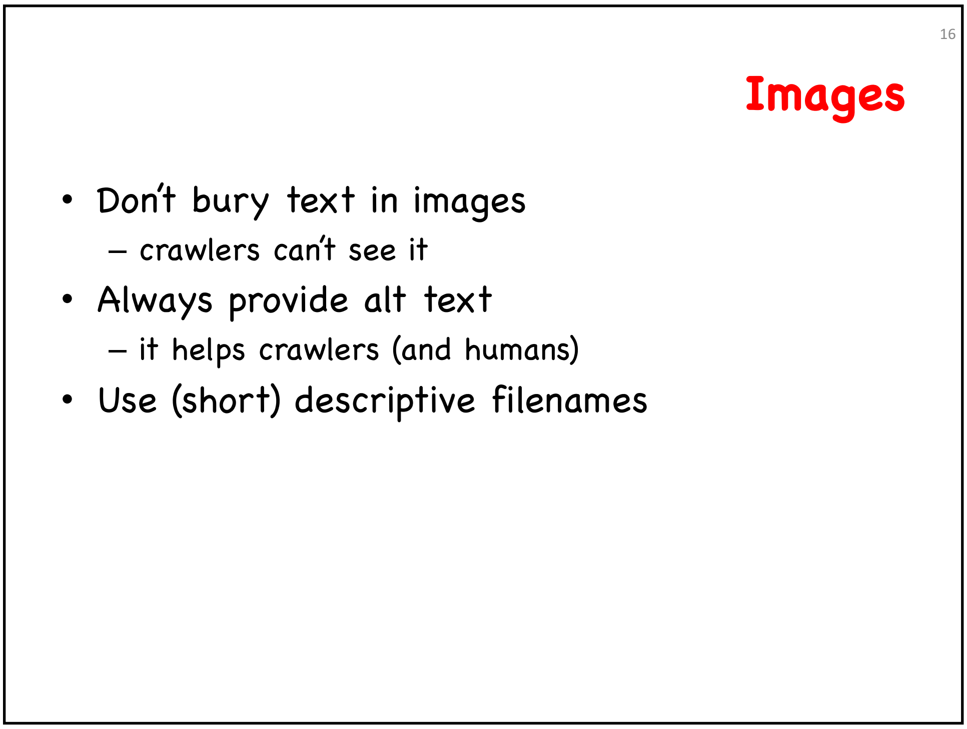
- Search engines like
  - text links
  - text links with informative text
  - customised missing link (404) pages
- They dislike
  - broken links
  - uninformative or non-existent 404 messages
- Don't leak rank to spammers
  - add the rel="nofollow" attribute to links in comments, unfavourable reviews, …



15

Content is paramount

- Good content gets linked to
- Stale content is ignored
  - keep updating the content
- Users will search using different vocabularies
  - according to experience, background, …
  - search engines provide tools to help identify which terms are used to access a site, synonyms and variants for keywords, …
- Create content for humans, not search engines



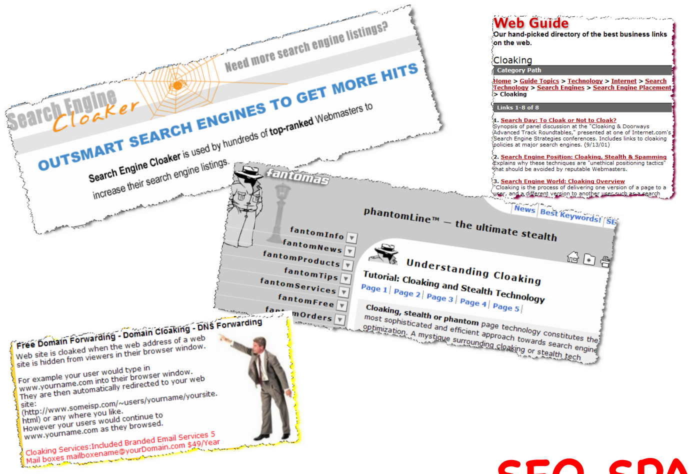
16

Images

- Don’t bury text in images
  - crawlers can’t see it
- Always provide alt text
  - it helps crawlers (and humans)
- Use (short) descriptive filenames

Page structure

- Use heading tags (h1-h6) appropriately
  - to reflect the structure and importance of text
- Use headings sparingly
  - avoid lots of very short sections
- Use the site's robots.txt file to keep crawlers away from content you don't want indexed
  - but not as a security measure



OUTSMART SEARCH ENGINES TO GET MORE HITS

Search Engine Cloaker is used by hundreds of **top-ranked** Webmasters to increase their search engine listings.

Web Guide

Our hand-picked directory of the best business links on the web.

Cloaking

Category Path

Home > Guide Topics > Technology > Internet > Search Technology > Search Engines > Search Engine Placement > Cloaking

Links 1-8 of 8

1. Search Day: To Cloak or Not to Cloak?
Synopsis of panel discussion at the "Cloaking & Doorways Advanced Track Roundtables," presented at one of Internet.com's Search Engine Strategies conferences. Includes links to cloaking policies at major search engines. (9/13/01)

2. Search Engine Position: Cloaking, Stealth & Spamming
Explains why these techniques are "unethical positioning tactics" that should be avoided by reputable Webmasters.

3. Search Engine World: Cloaking Overview
"Cloaking is the process of delivering one version of a page to a user, and a different version to another user such as a search

phantomLine™ – the ultimate stealth

Understanding Cloaking

Tutorial: Cloaking and Stealth Technology

Page 1 | Page 2 | Page 3 | Page 4 | Page 5 |

Cloaking, stealth or phantom page technology constitutes the most sophisticated and efficient approach towards search engine optimization. A mystique surrounding cloaking or stealth tech

Free Domain Forwarding - Domain Cloaking - DNS Forwarding

Web site is cloaked when the web address of a web site is hidden from viewers in their browser window.

For example your user would type in www.yourname.com into their browser window. They are then automatically redirected to your web site:
(http://www.someisp.com/~users/yourname/yoursite.html) or any where you like.
However your users would continue to www.yourname.com as they browsed.

Cloaking Services:Included Branded Email Services 5 Mail boxes mailboxename@yourDomain.com $49/Year

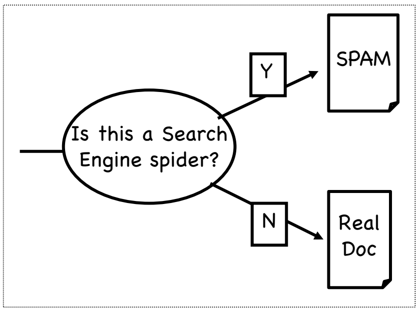
SEO SPAMMING

Search engine optimization (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - e.g., Webmaster world (www.webmasterworld.com)
    - Search engine specific tricks
    - Discussions about academic papers ☺

Cloaking

- Serve different content to crawlers and humans
  - sites suspected of cloaking get removed from search results
- DNS cloaking: Switch IP address

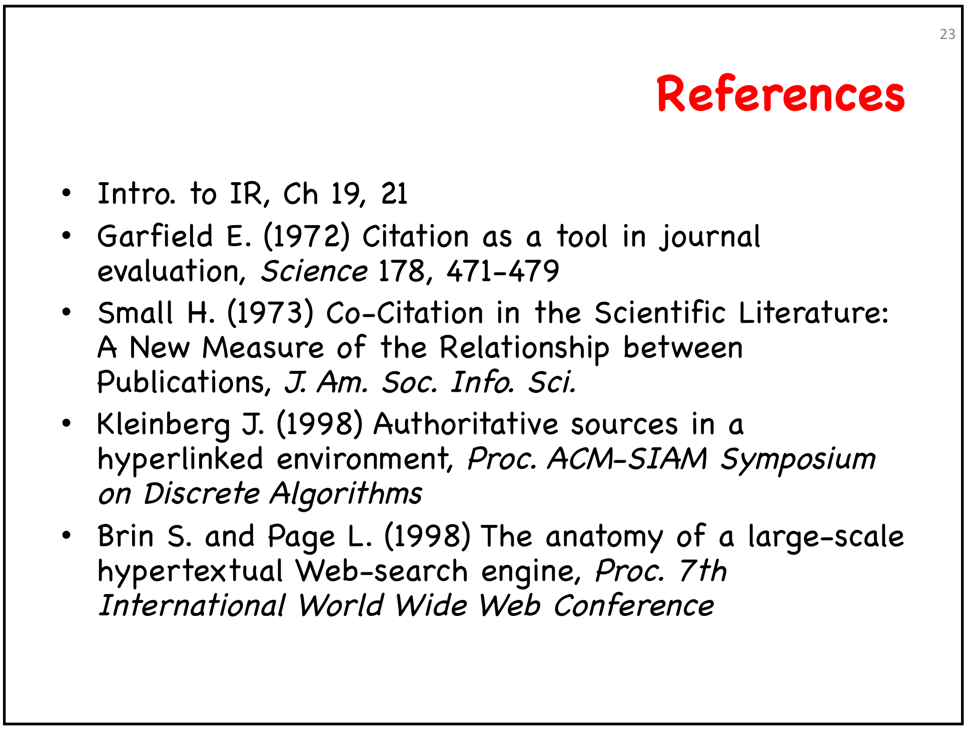


The war against spam

- Quality signals – prefer pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)
- Spam recognition
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, …
  - For images: flesh tone detectors, source text analysis, …
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect pattern detection

More on spam

- Web search engines have policies on SEO practices they tolerate/block
  - http://help.yahoo.com/help/us/ysearch/index.html
  - http://www.google.com/intl/en/webmasters/
- Adversarial IR: the (technical) battle between SEOs and web search engines
- Research http://airweb.cse.lehigh.edu/



References

- Intro. to IR, Ch 19, 21
- Garfield E. (1972) Citation as a tool in journal evaluation, *Science* 178, 471-479
- Small H. (1973) Co-Citation in the Scientific Literature: A New Measure of the Relationship between Publications, *J. Am. Soc. Info. Sci.*
- Kleinberg J. (1998) Authoritative sources in a hyperlinked environment, *Proc. ACM-SIAM Symposium on Discrete Algorithms*
- Brin S. and Page L. (1998) The anatomy of a large-scale hypertextual Web-search engine, *Proc. 7th International World Wide Web Conference*