Search Engine Optimization (SEO)

Remainder of semester

- **Tues 11/19:** Lecture
- **Thurs 11/21:** Lab
- **Fri 11/22:** Assignment 4 due*
- **11/26, 11/28:** No class (Thanksgiving!)
- **Tues 12/3:** Lecture
- **Thurs 12/5:** Lab (work on final project)
  - OK if you work elsewhere but send me an email ahead of time
- **Tues 12/10:** Lab (in 107 Info West; work on final project)
  - OK if you work elsewhere but send me an email ahead of time
- **Thurs 12/12:** Lecture (in 107 Info West)
- **Sun 12/15:** Final project due via OnCourse
- **Tues 12/17:** Final exam (10:15-12:15)

Sponsored vs Organic Results

- **Sponsored**
  - Ranking determined by how much you pay
  - Listed in a few hours
  - Lower click rate; users trust links less
  - Can control ad content and ranking

- **Organic**
  - Ranking determined by Google’s algorithm
  - Takes weeks/months to build up ranking
  - Higher click rate; users trust links more
  - Similar conversion rate
  - Can indirectly influence content and ranking

Organic ranking matters (a lot)
Search Engine Optimization (SEO)

- Techniques for improving the ranking of your page
  - Many websites hire companies to do this, analogous to traditional marketing firms
  - Search marketing is a $20.7 billion business [Forrester/2011]
- 97% of companies target Google for SEO, 44% target Bing

“White hat” vs “Black hat” SEO

- White hat SEO techniques make changes to a site to increase its rank or ease of indexing
  - Also SEO campaigns try to increase number of links to a site, using marketing and other techniques
  - Tolerated (sometimes encouraged) by search engines
- Black hat techniques increase rank using illegal (or at least questionable) methods
  - Rely on some form of deception
  - Often involve creating fake links, fake content, etc.

SEO techniques

- Google (and Yahoo, Bing, …) do not publish algorithms, so SEOers glean information from other sources
  - Hints from patent applications, academic papers, interviews with Google engineers, etc.
  - Experiments
  - Rumors
  - Experiences of others

Make a site easy to crawl

- Shallow link structure
- Simple is best!
  - Simple text links, static URLs
  - Links in Flash, Javascript, Images, frames, etc. may confuse the search engine
  - URLs with dynamic parameters, or pages only accessible by a web form, may be difficult for the search engine to find

Boost ranking

- Choose a good domain name
  - Short is good, ideally with important keywords
- Make targeted keywords prominent
  - Put them in page title, in headings, in link text
  - Put them near the start of the page
  - Put them in image captions (ALT fields)
  - But don’t overdo – search engines penalize pages with gratuitously repeated words
- Write clear, concise, high-quality text
  - That differentiates from competing sites
A]rac2ng

- PageRank rewards pages linked to by high-quality sites

- Typical techniques:
  - Submit page to high-profile link sites, e.g. Yahoo Directory (charges $299/year), Wikipedia
  - Get mentioned in blogs, tweets, etc.
  - Put out a press release to get mentioned in news sites
  - Contact high-profile websites relevant to target keywords, and ask them to link to your site
Other factors

- Google’s records of clicks and visit times
  - i.e. make sure site is attractive for people, too
- Too many ads are bad
- Seemingly low-quality text, including text copied from other websites, is bad
- Speculation on many other factors:

Black hat techniques

- Use deception to try to increase a page’s ranking
  - Not necessarily illegal by itself, but often conducted through illegal means (e.g. plagiarism, hacking, etc.)
  - Search engines try to detect and punish sites using black-hat techniques

Keyword stuffing

- Filling a page with many irrelevant keywords
- Typically hide irrelevant words from the user
  - Text colored the same as the background
  - Covering up text with other web elements (images, etc)
  - HTML comments, “noscript” tags, “meta” tags, etc.
- Most search engines try to detect and penalize words that are invisible to the user
  - Or gratuitously repeated words, or words that appear not to be in a reasonable grammatical structure

Fake content

- Filling a page with fake content that appears real
  - Plagiarizing content from other sites
  - Using algorithmic techniques (from natural language processing) to automatically generate text

Cloaking

- A cloaked website maintains two versions, depending on the type of visitor
  - To human users, it presents the real site
  - To search engine crawlers, it presents a fake site

PageRank sculpting

- Recall that a page’s outgoing links help determine the PageRank of those pages
  - The page’s votes are divided among outgoing links
- PageRank Sculpting changes how the votes are divided
  - E.g., you might not want to waste votes on your “Terms of Service” or “Legal notices” links
  - Mark the links with HTML “nofollow” attribute
  - Use complicated link techniques (e.g. using Javascript) to confuse the indexer into missing the page
nofollow

- nofollow attribute tells Google not to use a link in computing PageRank
  
  `<a href="http://www.example.com/" rel="nofollow">Link text</a>`

Robots

- Use automated programs to visit Google, conduct a search, and click on your webpage in the results list
  
  - Can increase rank, because Google uses click-through rates to rank future queries

Link spamming

- Link spam are links created for the sole purpose of increasing a page’s PageRank
  
  - Create fake web pages, blog entries, blog posts, etc.
  
  - Typically created by automated programs
Warning: Don't use a tool like this! It violates IU policy and may be illegal.

Another technique: use people

- An estimated 40% of work on Mechanical Turk is related to spam [peyrotis10]
  - Solving CAPTCHAs
  - Creating fake blog posts
  - Adding fake blog comments, product reviews, etc.