

# CS47300 Web Information Search and Management

Search Engine Optimization

*Prof. Chris Clifton*

11 November 2020



**PURDUE**  
UNIVERSITY

Department of Computer Science

## What is Search Engine Optimization?

- 90% of search engine “clickthroughs” are on the first page (top 10)
- Goal: get your site/page into the top 10
  - For some queries
  - given by people you want to reach
  - who are looking for what you have
- How?
  - ~~Pay for placement~~
  - Design of site

## Design of site?

- Proper words
  - In the proper places
- Proper structure
  - Links to and from
- *But remember, people need to use the site too*

7

## How it's done: Google SEO Starter Guide

- Unique, accurate page titles
  - Describe page content briefly
- “description” meta tag
  - Possible page summary for viewing result
  - And others metadata (e.g., “alt” text for images)
- Descriptive URLs
  - Informative to user
  - Easier for others to link to!
  - One URL for the page
- Easy structure to crawl
  - Or submit Sitemap
  - Meaningful anchor text
- Easy to read
  - Appropriate use of headings
- Avoid indexing the wrong things
  - Robots.txt file
  - `<a href="http://..." rel="nofollow">spam page</a>`
- Go Viral!
  - Promote site in social media

8

## How it's done: SEO.com

- Keyword Research
  - What words do people use when you want them to find you?
- Competitive Analysis
  - How to distinguish from your competitors?
- Link Building
  - “We employ a wide range of methods and techniques to attract valuable links, and the right links, to build your rankings.”
- Website optimization Services and Content Development/Copywriting
  - Search engines index new and unique content
  - final product needs to be more than just content for search engines
- Online Public Relations/Press Release Optimization

9

## Debatable Approaches

- “Doorway” pages loaded with keywords
  - Often “invisible” text
  - Does this help?
    - *Think TFIDF*
    - *What do [search engines really use](#)?*
- Gratuitous cross-site linking
  - I’ll link to your <basket weaving> course if you’ll link to my <information retrieval> course
  - *Increases PageRank!*

10

## SEO or Web Spam?

- SEO goal: get users to site
- Search engine goal: give users what they want
- *What happens when these conflict?*
- [Web spam](#)
  - Spam sites attempt to game their way to the top of search results through techniques like repeating keywords over and over, buying links that pass PageRank or putting invisible text on the screen.
  - Algorithms to detect/reduce rank of pages doing this
  - Manual analysis
  - Google: >400,000 notifications of actions per month

11

## Web Spam Taxonomy

*Courtesy Ullman, Gyongyi and Garcia-Molina*

- Boosting techniques
  - Techniques for achieving high relevance/importance for a web page
- Hiding techniques
  - Techniques to hide the use of boosting
    - From humans and web crawlers



*Dr. Hector Garcia-Molina*  
1953-2019

## Boosting techniques

---

- Term spamming
  - Manipulating the text of web pages in order to appear relevant to queries
- Link spamming
  - Creating link structures that boost page rank or hubs and authorities scores

## Term Spamming

---

- Repetition
  - of one or a few specific terms e.g., free, cheap, viagra
  - Goal is to subvert TF.IDF ranking schemes
- Dumping
  - of a large number of unrelated terms
  - e.g., copy entire dictionaries
- Weaving
  - Copy legitimate pages and insert spam terms at random positions
- Phrase Stitching
  - Glue together sentences and phrases from different sources

## Term spam targets

---

- Body of web page
- Title
- URL
- HTML meta tags
- Anchor text

## Link spam

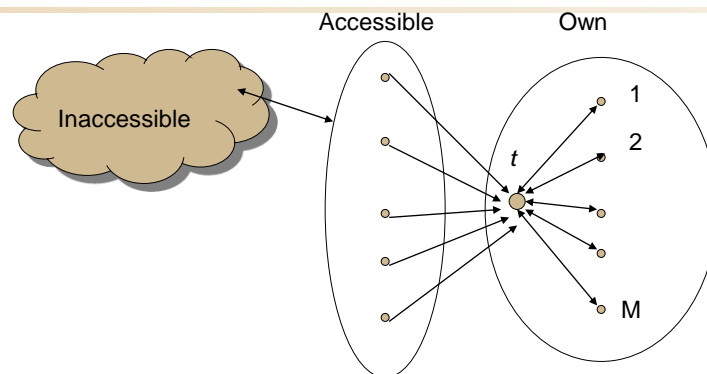
---

- Three kinds of web pages from a spammer's point of view
  - Inaccessible pages
  - Accessible pages
    - e.g., web log comments pages
    - spammer can post links to his pages
  - Own pages
    - Completely controlled by spammer
    - May span multiple domain names

# Link Farms

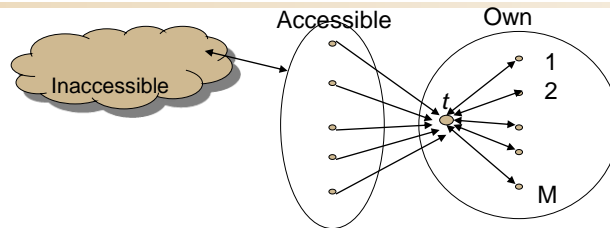
- Spammer's goal
  - Maximize the page rank of target page  $t$
- Technique
  - Get as many links from accessible pages as possible to target page  $t$
  - Construct "link farm" to get page rank multiplier effect

# Link Farms



One of the most common and effective organizations for a link farm

## Analysis



Suppose rank contributed by accessible pages =  $x$

Let page rank of target page =  $y$

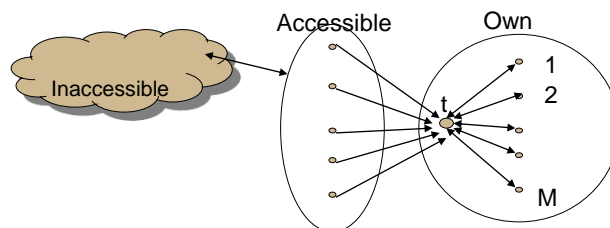
Rank of each “farm” page =  $\beta y/M + (1-\beta)/N$

$y = x + \beta M[\beta y/M + (1-\beta)/N] + (1-\beta)/N$

$= x + \beta^2 y + \beta(1-\beta)M/N + \boxed{(1-\beta)/N}$  Very small; ignore

$y = x/(1-\beta^2) + cM/N$  where  $c = \beta/(1+\beta)$

## Analysis



- $y = x/(1-\beta^2) + cM/N$  where  $c = \beta/(1+\beta)$
- For  $\beta = 0.85$ ,  $1/(1-\beta^2) = 3.6$ 
  - Multiplier effect for “acquired” page rank
  - By making  $M$  large, we can make  $y$  as large as we want



## Hiding techniques

---

- Content hiding
  - Use same color for text and page background
- Cloaking
  - Return different page to crawlers and browsers
- Redirection
  - Alternative to cloaking
  - Redirects are followed by browsers but not crawlers

## Detecting Spam

---

- Term spamming
  - Analyze text using statistical methods e.g., Naïve Bayes classifiers
  - Similar to email spam filtering
  - Also useful: detecting approximate duplicate pages
- Link spamming
  - Open research area
  - One approach: TrustRank

## TrustRank idea

---

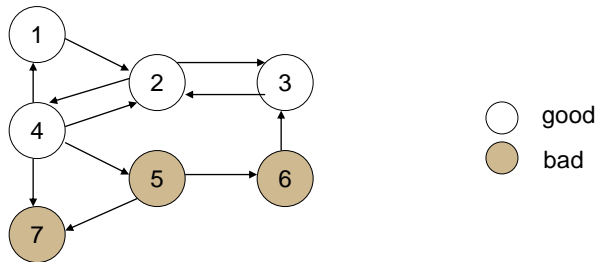
- Basic principle: approximate isolation
  - It is rare for a “good” page to point to a “bad” (spam) page
- Sample a set of “seed pages” from the web
- Have an oracle (human) identify the good pages and the spam pages in the seed set
  - Expensive task, so must make seed set as small as possible

## Trust Propagation

---

- Call the subset of seed pages that are identified as “good” the “trusted pages”
- Set trust of each trusted page to 1
- Propagate trust through links
  - Each page gets a trust value between 0 and 1
  - Use a threshold value and mark all pages below the trust threshold as spam

## Example



## Rules for trust propagation

- Trust attenuation
  - The degree of trust conferred by a trusted page decreases with distance
- Trust splitting
  - The larger the number of outlinks from a page, the less scrutiny the page author gives each outlink
  - Trust is “split” across outlinks

## Simple model

- Suppose trust of page  $p$  is  $t(p)$ 
  - Set of outlinks  $O(p)$
- For each  $O(p)$ ,  $p$  confers the trust
  - $\beta t(p)/|O(p)|$  for  $0 < \beta < 1$
- Trust is additive
  - Trust of  $p$  is the sum of the trust conferred on  $p$  by all its inlinked pages
- Note similarity to Topic-Specific Page Rank
  - Within a scaling factor, trust rank = biased page rank with trusted pages as teleport set

## Picking the seed set

- Two conflicting considerations
  - Human has to inspect each seed page, so seed set must be as small as possible
  - Must ensure every “good page” gets adequate trust rank, so need make all good pages reachable from seed set by short paths

## Approaches to picking seed set

---

- Suppose we want to pick a seed set of  $k$  pages
- PageRank
  - Pick the top  $k$  pages by page rank
  - Assume high page rank pages are close to other highly ranked pages
  - We care more about high page rank “good” pages

## Inverse page rank

---

- Pick the pages with the maximum number of outlinks
- Can make it recursive
  - Pick pages that link to pages with many outlinks
- Formalize as “inverse page rank”
  - Construct graph  $G'$  by reversing each edge in web graph  $G$
  - Page Rank in  $G'$  is inverse page rank in  $G$
- Pick top  $k$  pages by inverse page rank