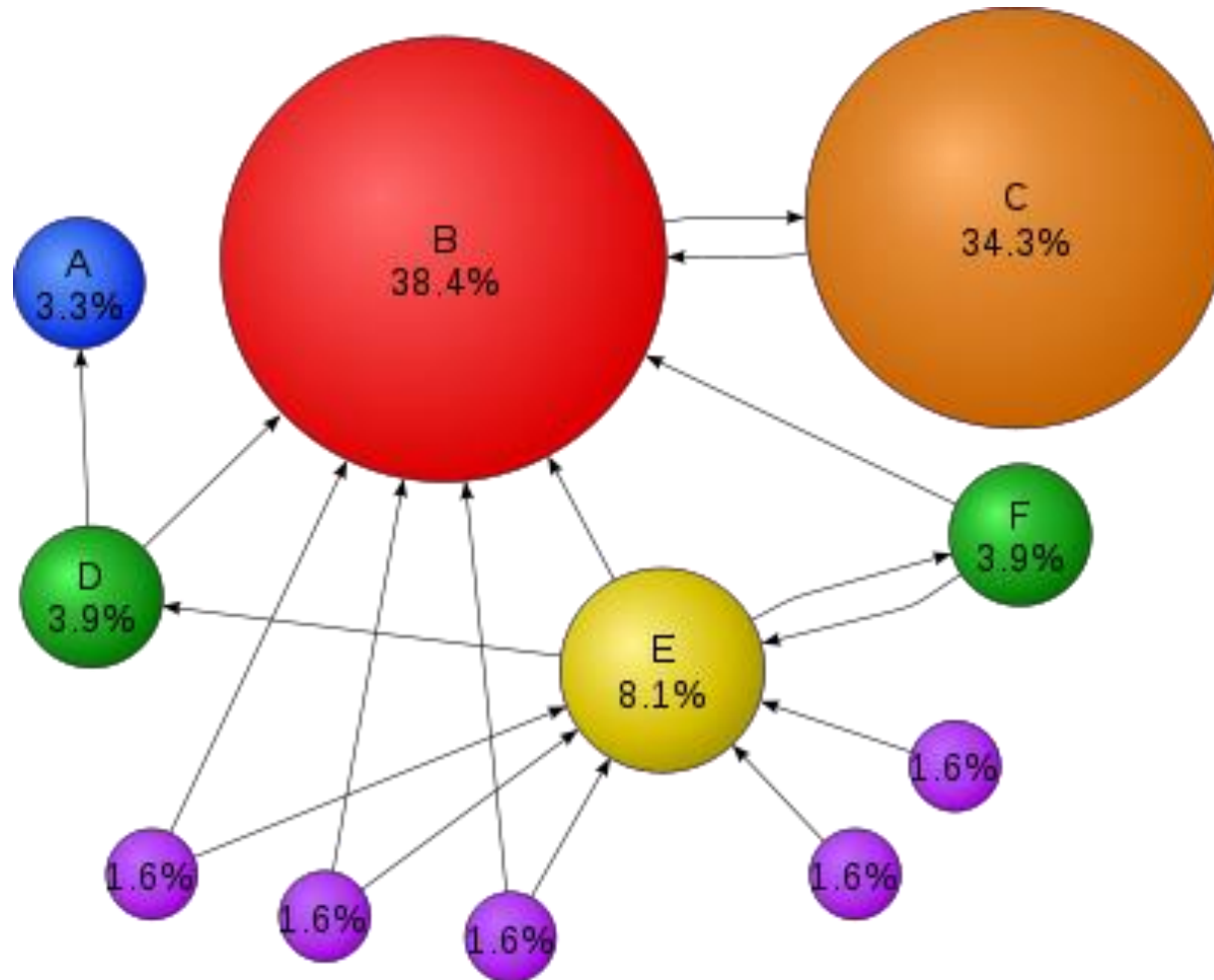


Lecture 18:
CS 5306 / INFO 5306:
Crowdsourcing and
Human Computation

Web Link Analysis (Wisdom of the Crowds)



(Not Discussing)

- Information retrieval
(term weighting, vector space representation, inverted indexing, etc.)
- Efficient web crawling
- Efficient real-time retrieval

Web Search: Prehistory

- Crawl the Web, generate an index of all pages
 - Which pages?
 - What content of each page?
 - (Not discussing this)
- Rank documents:
 - Based on the text content of a page
 - How many times does query appear?
 - How high up in page?
 - Based on display characteristics of the query
 - For example, is it in a heading, italicized, etc.

Link Analysis: Prehistory

- L. Katz. "A new status index derived from sociometric analysis", *Psychometrika* 18(1), 39-43, March 1953.
- Charles H. Hubbell. "An Input-Output Approach to Clique Identification", *Sociolmetry*, 28, 377-399, 1965.
- Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science* 178, 1972.
- G. Pinski and Francis Narin. "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics", *Information Processing and Management*. 12, 1976.
- Mark, D. M., "Network models in geomorphology," *Modeling in Geomorphologic Systems*, 1988
- T. Bray, "Measuring the Web". *Proceedings of the 5th Intl. WWW Conference*, 1996.
- Massimo Marchiori, "The quest for correct information on the Web: hyper search engines", *Computer Networks and ISDN Systems*, 29: 8-13, September 1997, Pages 1225-1235.

Hubs and Authorities

- J. Kleinberg. “Authoritative sources in a hyperlinked environment”. *Journal of the ACM* **46** (5): 604–632, 1999.

(Previously *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998, IBM technical report 1997.)

Hubs and Authorities

- For each web page v in a set of pages of interest (think: pages that contain your query):
 - $a(v)$ - the authority of v
 - $h(v)$ - the hubness of v
- $a(v)$: higher for “authorities” that are linked to by other pages
- $h(v)$: higher for “hubs” that link to other pages

Hubs and Authorities

$$a(v) \leftarrow \sum_{w \in \text{in}[v]} h(w)$$

$$h(v) \leftarrow \sum_{w \in \text{out}[v]} a(w)$$

Recursive, start with $a(v) = h(v) = 1$ for all v
Normalize values after each step

$a(v)$ and $h(v)$ converge (!)

Formulate as a linear algebra problem

Hubs and Authorities

A: Adjacency matrix

A_{ij} = link from i to j

$$\mathbf{h}(v) \leftarrow A \cdot \mathbf{a}(v)$$

$$\mathbf{a}(v) \leftarrow A^T \cdot \mathbf{h}(v)$$

Boils down to computing the eigenvectors of AA^T and $A^T A$

Known as the HITS algorithm (Hyperlink-Induced Topic Search)

PageRank

- S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems* 30: 107–117, 1998

PageRank

- Random Surfer model:
 - Users conduct a random walk of the Web graph, selecting a link at random from every page

$$S(V_i) = \sum_{j \in \text{in}(V_i)} \frac{S(V_j)}{|\text{out}(V_j)|}$$

- $S(V)$: Proportional to probability of landing at V

PageRank

- Problem:
 - Sinks get all the weight
- Solution:
 - Random walk with a probability of teleporting to another node at random

$$S(V_i) = (1-d) + d \sum_{j \in \text{In}(V_i)} \frac{S(V_j)}{|\text{Out}(V_j)|}$$

d – damping factor $\in [0,1]$ (usually 0.8-0.9)

PageRank

- Recursive
- $S(V)$ converges
- Formulate as linear algebra

Google Search Today

- Over 200 Factors
 - Previous searches
 - Previous page
 - Search history
 - Session history
 - Click history
 - Location
 - Time of day
 - Personal profile
 - Gmail
 - Social network
 - Images?
 - OS
 - Bandwidth of my connection
 - Bandwidth of website
 - Length of domain ownership
 - Trendiness (in news?)
 - Recency
 - Top-level domain (.edu, .gov, etc)
 - Trusted certificates
 - Lots of websites with unimportant content
 - Hosts of free websites
 - Legality (?)

Google Search Today

- Over 200 Factors
 - Frequency of query words in the page
 - Proximity of matching words to one another
 - Location of terms within the page
 - Location of terms within tags e.g. <title>, <h1>, link text, etc.
 - Word format characteristics (boldface, capitalized, etc)
 - Anchor text on pages pointing to this one
 - Frequency of terms on the page and in general
 - Click-through analysis: how often the page is clicked on
 - How “fresh” is the page

Google Search Today

- Over 200 Factors
 - Is page hosted by a provider with a high percentage of spam pages?
 - Is page hosted by a site with few pages?
 - Is page hosted by a free provider?
 - Distinctive link patterns
 - Are the links in content from “open” resources, like blog comments, guestbooks, etc.?
 - Are pages with links duplicates of others?
 - Does page have little original content?
 - Speed of server
 - Your search history
 - Your Google profile

Google Search Today

- How to weight factors?
- Machine learning to the rescue!
- Experimental infrastructure
 - “Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO”, Ron Kohavi and Randal M. Henne, KDD 2007

How Google Chooses Algorithm Updates

Precision Evaluations

The first phase is to get feedback from evaluators, people who evaluate search quality based on our guidelines. We show evaluators search results and ask them to rate the usefulness of the results for a given search.

Note: These ratings don't directly impact ranking.

Live Traffic Experiments

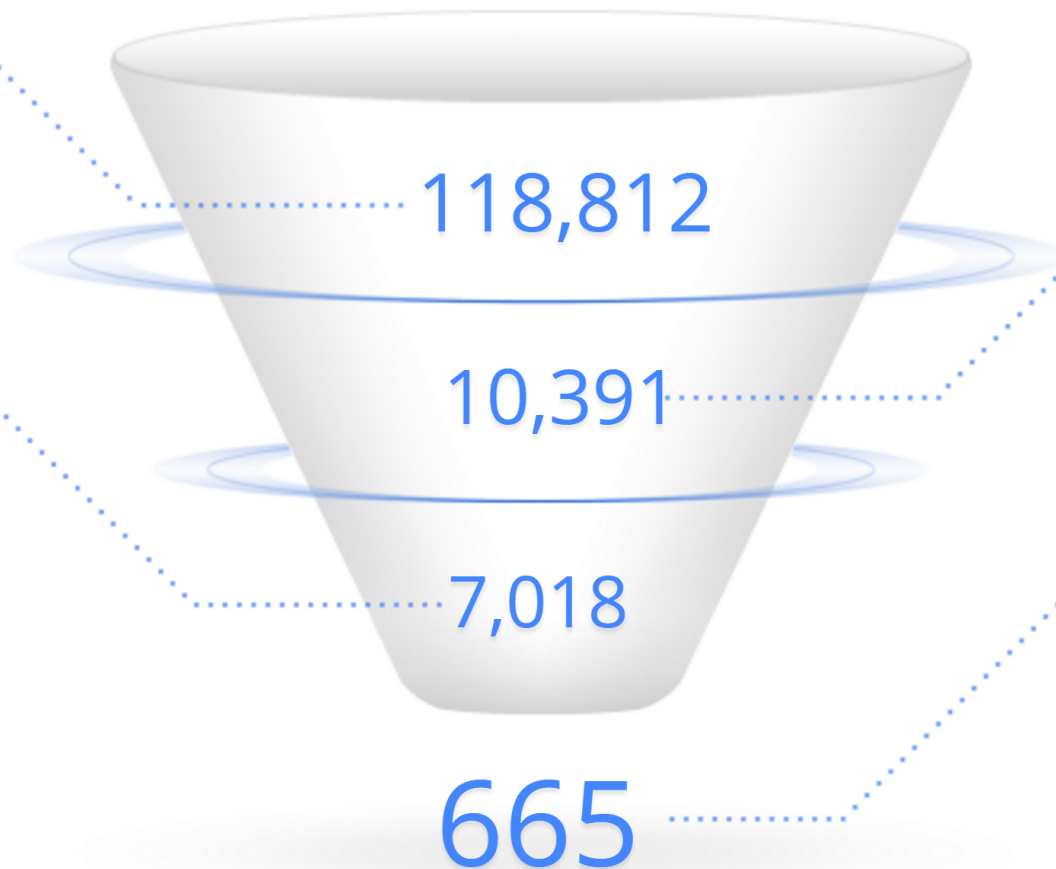
If the evaluators' feedback looks good, we move forward with a "live traffic experiment." In these experiments, we change search for a small percentage of real Google users and see how it changes the way they interact with the results. We carefully analyze the results to understand whether the change is an improvement to the search results. For example, do searchers click the new first result more often? If so, that's generally a good sign.

Side-by-Side Experiments

In a side-by-side experiment, we show evaluators two different sets of search results: one from the old algorithm and one from the new, and we ask them for details about which results they prefer.

Launches

Finally, our most experienced search engineers carefully review the data from all the different experiments and decide if the change is approved to launch. It sounds like a lot, but the process is well refined, so an engineer can go from idea to live on Google for a percentage of users in 24 hours. Based on all of this experimentation, evaluation and analysis, we launched 665 improvements to search in 2012.





OPTIMIZE YOUR WEBSITE

Follow these simple tips to help Google understand the content on your site. This information helps deliver great results to searchers (your future customers)!

1. LOOK GOOD IN THE SEARCH RESULTS

- A** Your page title is used as a suggestion for the title in Google's search results. Describe your business in a concise, informative phrase.
- B** Domain names are an important part of Google's search results. Choose a descriptive and easy-to-read domain name for your website. Sub-pages should also be easy to read. For example, use www.stasiabakery.com/custom-cakes instead of www.stasiabakery.com/prodid?12345.
- C** Meta descriptions are page summaries often used by Google and other search engines on the search results page. Write unique descriptions for each page in 160 characters or less.

A [Stasia's Bakery - The Best Bakery In New York](#)

B www.stasiabakery.com/

C A family owned bakery located in the heart of New York's cutest neighborhood specializing in custom cakes and pastries.

A [Home Page!](#)

B www.example.com/

C Welcome to the home page of our new store! About | News | How to find Us

2. HELP GOOGLE UNDERSTAND IMAGES

- A** Give your images short, descriptive file names.
- B** The "alt" attribute describes the image. This helps Google understand what's in the image.
- C** Write a short caption on the page below each image. Put important information in text rather than images.

A ``

B High quality prints of Van Gogh's Starry Night

C

3. UPDATE AND KEEP GOING



PROVIDE USEFUL CONTENT AND KEEP IT UP TO DATE

Your website is like a virtual storefront. You wouldn't leave a store unattended for 6 months, right? Keep your site fresh by starting a blog, announcing new products, sales, and special offers. Remember to put yourself in your customer's shoes and make sure you provide them with the information they need.

MORE INFORMATION & SUPPORT

Google's Webmaster Academy offers free step-by-step lessons and short instructional videos.

g.co/webmasteracademy

Looking for more advice on how to optimize your website? Read Google's SEO starter guide:

g.co/seoguide

Google's free Webmaster Tools helps you understand and improve your website, get timely alerts on problems, and find answers to questions.

google.com/webmasters



OPTIMIZE YOUR WEBSITE

Follow these simple tips to help Google understand the content on your site. This information helps deliver great results to searchers (your future customers)!

1. LOOK GOOD IN THE SEARCH RESULTS

- A** Your page title is used as a suggestion for the title in Google's search results. Describe your business in a concise, informative phrase.
- B** Domain names are an important part of Google's search results. Choose a descriptive and easy-to-read domain name for your website. Sub-pages should also be easy to read. For example, use www.stasiabakery.com/custom-cakes instead of www.stasiabakery.com/prodid?12345.
- C** Meta descriptions are page summaries often used by Google and other search engines on the search results page. Write unique descriptions for each page in 160 characters or less.

The diagram illustrates search results for two websites. The first result, for 'Stasia's Bakery', is enclosed in a green box with a green checkmark. Callout A points to the title 'Stasia's Bakery - The Best Bakery In New York', callout B points to the URL 'www.stasiabakery.com/', and callout C points to the meta description 'A family owned bakery located in the heart of New York's cutest neighborhood specializing in custom cakes and pastries.' The second result, for 'Home Page', is enclosed in a red box with a red minus sign. Callout A points to the title 'Home Page!', callout B points to the URL 'www.example.com/', and callout C points to the meta description 'Welcome to the home page of our new store! About | News | How to find Us'.

Search Engine Optimization (SEO)

important information in text rather than images.

High quality prints of Van Gogh's Starry Night

3. UPDATE AND KEEP GOING



PROVIDE USEFUL CONTENT AND KEEP IT UP TO DATE

Your website is like a virtual storefront. You wouldn't leave a store unattended for 6 months, right? Keep your site fresh by starting a blog, announcing new products, sales, and special offers. Remember to put yourself in your customer's shoes and make sure you provide them with the information they need.

MORE INFORMATION & SUPPORT

Google's Webmaster Academy offers free step-by-step lessons and short instructional videos.

g.co/webmasteracademy

Looking for more advice on how to optimize your website? Read Google's SEO starter guide:

g.co/seoguide

Google's free Webmaster Tools helps you understand and improve your website, get timely alerts on problems, and find answers to questions.

google.com/webmasters



**Search Engine Optimization
Starter Guide**



Search Engine Optimization (SEO)

- “White hat” SEO: Focus on “allowable” optimizations that are intended to steer sites to user-centered designs / that adhere to search engines’ rules
- “Black hat” SEO: Focus on search engine algorithm
 - Repeating keywords many many times
 - Invisible text
 - Non-genuine web pages with links to desired page

Google Bombing



Web

Results 1 - 10 of about 969,000 for [miserable failure](#). (0.06 seconds)

[Biography of President George W. Bush](#)

Biography of the president from the official White House web site.

www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)

[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)

[More results from www.whitehouse.gov »](#)

[Welcome to MichaelMoore.com!](#)

Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...

www.michaelmoore.com/ - 35k - [Sep 1, 2005](#) - [Cached](#) - [Similar pages](#)

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)

Web users manipulate a popular search engine so an unflattering description leads to the president's page.

news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - [Cached](#) - [Similar pages](#)

[Google's \(and Inktomi's\) Miserable Failure](#)

A search for **miserable failure** on Google brings up the official George W.

Bush biography from the US White House web site. Dismissed by Google as not a ...

searchenginewatch.com/sereport/article.php/3296101 - 45k - [Sep 1, 2005](#) - [Cached](#) - [Similar pages](#)

Google Bombing



Did you mean: [french military defeats](#)

No standard web pages containing all your search terms were found.

Your search - **french military victories** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Also, you can try [Google Answers](#) for expert help with your search.



Google Bombing



These Weapons of Mass Destruction cannot be displayed

The weapons you are looking for are currently unavailable. The country might be experiencing technical difficulties, or you may need to adjust your weapons inspectors mandate.

Please try the following:

- Click the  Regime change button, or try again later.
- If you are George Bush and typed the country's name in the address bar, make sure that it is spelled correctly. (IRAQ).
- To check your weapons inspector settings, click the **UN** menu, and then click **Weapons Inspector Options**. On the **Security Council** tab, click **Consensus**. The settings should match those provided by your government or NATO.
- If the Security Council has enabled it, The United States of America can examine your country and automatically discover Weapons of Mass Destruction. If you would like to use the CIA to try and discover them, click  [Detect weapons](#)
- Some countries require 128 thousand troops to liberate them. Click the **Panic** menu and then click **About US foreign policy** to determine what regime they will install.
- If you are an Old European Country trying to protect your interests, make sure your options are left wide open as long as possible. Click the **Tools** menu, and then click on

Google Bombing

Web Images Maps News Shopping Gmail more ▾

Google™ Search

Web Results 1 - 10 of about 252,000 for [dangerous cult](#). (0.06 se

[Scientology - Church of Scientology Official Site](#)
Living in a **Dangerous** Environment · Drug and Alcohol Problems · Personalities, Emot
and How to Deal with Others ...
[www.scientology.org/](#) - 73k - [Cached](#) - [Similar pages](#) - [Note this](#)

[The Most **Dangerous Cult** in The World by Laura Knight-Jadczyk](#)
There's a new religious **cult** in America. It's not composed of so-called "crazies" so mu
mainstream, middle to upper-middle class Americans. ...
[www.cassiopaea.org/cass/Laura-Knight-Jadczyk/fastest_growing_cult.htm](#) - 144k -
[Cached](#) - [Similar pages](#) - [Note this](#)

[Dangerous Cult Warning Signs](#)
If you, or a loved one, are in a **dangerous cult**, as determined by the above checklist,
must do everything you possibly can to remove the potential ...
[www.vistech.net/users/rsturge/cults.html](#) - 4k - [Cached](#) - [Similar pages](#) - [Note this](#)

[The Watchman Expositor: The Most **Dangerous Cult** in America](#)
However, when the world's final chapter is written, which will prove to be "THE most
dangerous cult in America?" One of the cults mentioned above? ...
[www.watchman.org/rektop/budcomp.htm](#) - 10k - [Cached](#) - [Similar pages](#) - [Note this](#)

Google Bombing

- “more evil than Satan himself”: microsoft.com (1999)
- “French military victories”: page with “Did you mean French military defeats?” (2003)
- “weapons of mass destruction” (2003)
- “miserable failure”: George Bush (2003)
- “waffles”: Al Gore (2004)
- “Jew”: Wikipedia article for “Jew” (2004)
- Amway Quixtar (2006)
- “liar”: Tony Blair (2005)
- “worst band in the world”: Creed (2006)
- “dangerous cult”: Scientology
- “murder”: Wikipedia article for abortion