



4D2b - Web Search and Retrieval

Owen.Conlan@scss.tcd.ie



History of Web Search

- The Web is less than 20 years old!
 - In 1992 it was merely a collection of text docs
- Internet Search Engines
 - Archie - McGill University (1990)
 - Veronica - University of Nevada (1993)
- 1993-1996 the web grew from approx. 130 sites to 600,000
 - WWW Wanderer - MIT (1994)

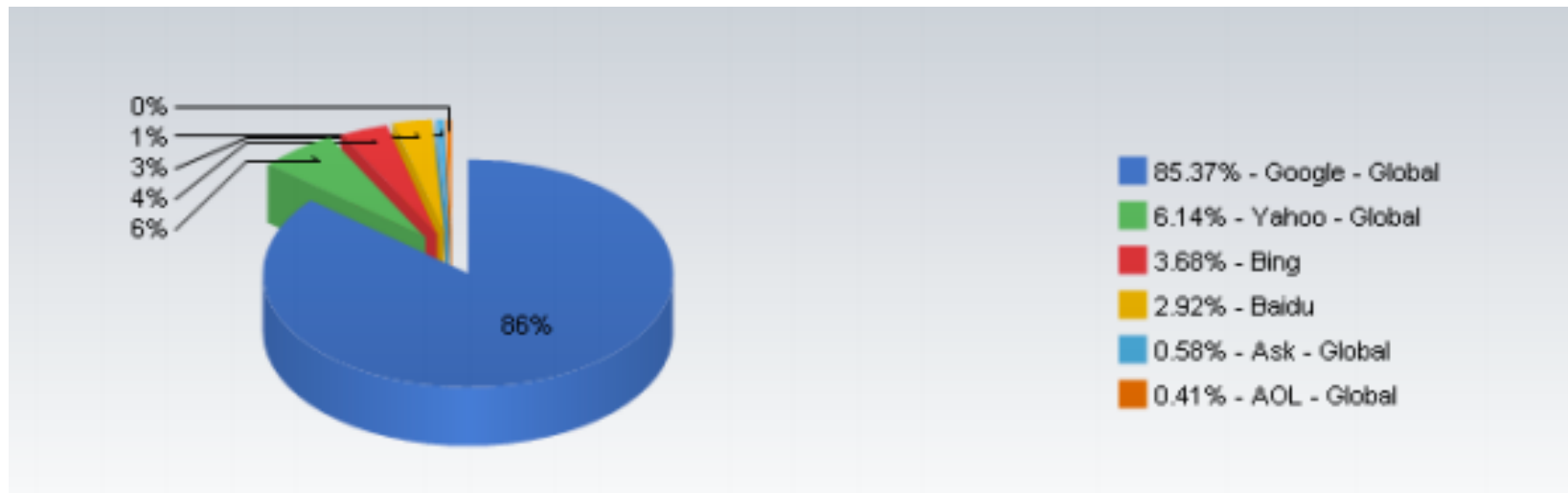


History of Web Search

- The first true Web Search Engines then began to appear
 - WebCrawler - Washington University (1994)
 - Lycos - Carnegie Mellon University (1994)
 - Altavista - Digital Equipment Corp. (1995)
- “Altavista wasn’t the first, but they were the first to do it in a way that was a significant improvement over the state of the art”
 - Dr. Gary Flake, Engineer, Microsoft Corp.



Search share as it stands...





As it stands...

- **224,749,695** web sites at last count! – stats from netcraft.com
- In July 2008 Google claimed to have found **1 Trillion** unique URLs on the web.
 - <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>
- **61 Billion** searches monthly worldwide – comscore 2007
- 50% of searches use two or three keywords, 20% use just one



Web Search Challenges

- WWW expanding faster than any current search engine can possibly index
- Many web pages are updated frequently or are dynamically generated which forces search engines to repeatedly revisit them
- Many dynamically generated sites are not indexable by search engines; this phenomenon is known as the *invisible web*.



Web Search Challenges

- The ordering of results is not always solely by relevance, but sometimes influenced by monetary contributions
 - Difficulty with Business Model
- Some sites use tricks to manipulate the search engine to improve their ranking for certain keywords; This is known as Search engine spamming



Hyperlink Analysis

- Adapted from the concept of Citation Analysis in Academia
- Analyses
 - Hyperlink and Anchor Text
 - Linking Page
 - Linked Page
- Numerous Algorithms Exist
 - HITS
 - PageRank



Enter Google

“Google's complex, automated methods **make human tampering with our results extremely difficult**. And though we do run relevant ads above and next to our results, Google does not sell placement within the results themselves (i.e., no one can buy a higher PageRank). A Google search is an easy, honest and objective way to find high-quality websites with information relevant to your search.”

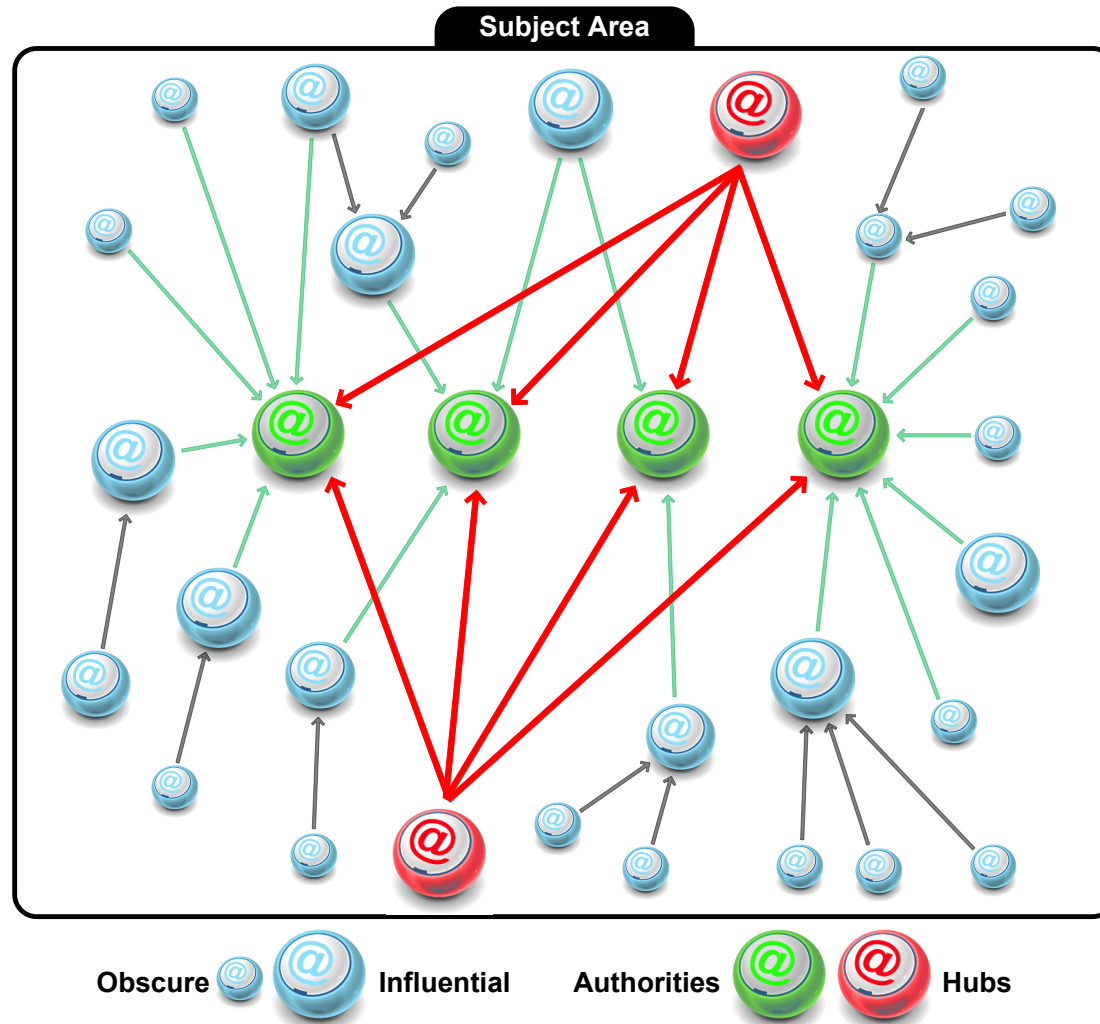


Hubs and Authorities

- Assumption
 - If a link exists from Page A \rightarrow Page B
 - Author of A recommends B
- A Page that is linked to by many other Influential Pages in a subject area is an Authority
- A Page that links to many Influential Pages in a subject area is a Hub
- A good Authority is linked to by many good Hubs
- A good Hub links to many good Authorities



Hubs and Authorities





Hyperlink Algorithms

- Recursive Algorithms which assume....
 - Quality of a Page is directly related to the quality of the Pages that link to it
 - Popularity of a Page is indicated by the number of Pages that link to it
 - Popular Pages are more likely to contain relevant Information than unpopular Pages
- “Link” measures are generated for each page and used in ranking



HITS Algorithm

- Hyperlink Induced Topic Selection – HITS
 - Popular Pages can become Hubs or Authorities
 - Quality of each Hub and Authority is calculated based on Pages that link to them
 - Each Hub and Authority is checked often to ensure that it has maintained its importance
 - Implemented in ask.com and teoma



HITS (Hypertext-Induced Topic Search)

1. Use query terms to retrieve a **root set** of pages (say 200).
2. Create a **base set** S by adding all the pages the root set links too (say 1000).
3. Associate non-negative **authority weights** a_p and **hub weights** h_p to each page



HITS (contd.)

- These weights can be updated as follows:

$$a_p = \sum_{q \in S | q \rightarrow p} h_q \quad h_p = \sum_{q \in S | q \leftarrow p} a_q$$

- We introduce an adjacency matrix **A**
 - $A(i,j) = 1$ if page i links to page j
 - The authority and hub weight vectors are
 - $h = \mathbf{A}.a; a = \mathbf{A}^T.h$



HITS (contd)

- $h = (\mathbf{AA}^T)^k h; a = (\mathbf{A}^T \mathbf{A})^k a$
- We can initialise h and a with random vals.
 - Or $h_i \leftarrow 1/|h|$
- According to linear algebra these two equations converge to the principle eigenvectors of \mathbf{AA}^T and $\mathbf{A}^T \mathbf{A}$ respectively.



PageRank Algorithm

- The basis of Google's Ranking System
 - Originally developed as BackRub
 - <http://backrub.tjtech.org/1997/backrub.htm>
- “simulates a random walk across the web and computes the score of a Page as the probability of reaching that Page”
- A Page has a high rank *if* the sum of the ranks of Pages pointing at it is high
 - Many Pages or Several Highly Ranked Pages



PageRank Algorithm

- Very complex algorithm but essentially...
 - Looks at the links on a Page
 - Analyses the Anchor Text around that link
 - Examines the popularity of Pages that link to that particular Page
 - Generates a PageRank score for every page in the Index
 - Over 100 factors incorporated in total



Hubs & Authorities

- Other things being equal, a web page with **a lot of links into it** is probably a better authority than one without.
 - A web page with a lot of links into it is an authority
 - A page with a lot of links out is a hub
- Then a web page with a **lot of links from a hub** is better than a web page with a lot of links from ordinary pages.



“Concept” Search

- Find pages on the topic but do not actually contain the keywords
- If hubs point at a page using a particular term then that page is probably relevant to that term



Google

- Google is HITS + Anchor Text
 - Details on a page can be augmented with text in Anchors of Links pointing to it
 - Font information can influence term weighting
 - Headings, Font Size, Bold,
- Emphasises high precision over recall
 - It might be said that Precision is real-valued rather than binary



Google vital stats (2001)

- 6000 Linux machines
 - 33 die every day
- 500TB of disk storage
- 1 Google day = 16.5 machine years
 - $(6,000/365)$
- 50 million queries per day
 - 1000 queries / sec
- 3 data replication centres



Google vital stats (2012)





Google vital stats (2012)

- Mostly speculation
 - Search is no longer Google's only business!
- Highly customised server configuration
- Tailored version of Linux
- 500 and 681 megawatts
- 20 to 100 petaflops in 2008?



Cheats

- Finding new and better ways to **scam search engines** has long been a popular pastime of webmasters.
 - Traditionally, search engines used keywords found in **metatags** and the body of html pages to index a web site.
 - So webmasters listed every possible keyword even remotely related to their site in **metatags** or in **invisible text**.
 - After a while most search engines became fairly useless because of all the junk sites listed in the search results.



Scamming Google

- First the “dumb m*****-f*****” controversy



For a while the top result for this query on Google

This is Google's of <http://www.georgewbushstore.com/>.
Google's cache is the snapshot that we took of the page as we crawled the web.
The page may have changed since that time. Click here for the [current page](#) without highlighting.

Google is not affiliated with the authors of this page nor responsible for its content.

These terms only appear in links pointing to this page: **dumb motherfucker**



georgewbushstore.com

enter
the online store



Welcome to the George W. Bush for President Political Materials website. You'll find great Bush-Cheney campaign materials, gift items and wearables.

**BE SURE TO LOOK FOR OUR SPECIAL VALUE ITEMS!
SOME OF OUR FAVORITE PRODUCTS NOW HAVE A
NEW LOW PRICE!**

This website is proudly presented and independently operated by the Spalding Group. For information on the Spalding Group, call 1-800-388-8755.

All materials have been sanctioned by Bush for President, Inc.



Scamming Google

- First the “dumb m*****-f*****” controversy
 - *Was this orchestrated?*
- Apparently “more evil than satan” brought back www.microsoft.com for a while so
 - “the engine was tweaked to fix it”



Future Research?

- Hubs, Authorities + Link Text
 - Very powerful indicators for ranking
- Is there room for further improvement?
 - Identify scammers (bogus link structures)
 - Networks of hubs and authorities are subgraphs
 - Are bogus subgraphs distinguishable from real graphs?