Web Search

Introduction

The World Wide Web

- Developed by Tim Berners-Lee in 1990 at CERN to organize research documents available on the Internet.
- Combined idea of documents available by FTP with the idea of *hypertext* to link documents.
- Developed initial HTTP network protocol, URLs, HTML, and first “web server.”

Web Pre-History

- Ted Nelson developed idea of hypertext in 1965.
- Doug Engelbart invented the mouse and built the first implementation of hypertext in the late 1960’s at SRI.
- ARPANET was developed in the early 1970’s.
- The basic technology was in place in the 1970’s; but it took the PC revolution and widespread networking to inspire the web and make it practical.

Web Browser History

- Early browsers were developed in 1992 (Erwise, ViolaWWW).
- In 1993, Marc Andreessen and Eric Bina at UIUC NCSA developed the Mosaic browser and distributed it widely.
- Andreessen joined with James Clark (Stanford Prof. and Silicon Graphics founder) to form Mosaic Communications Inc. in 1994 (which became Netscape to avoid conflict with UIUC).
- Microsoft licensed the original Mosaic from UIUC and used it to build Internet Explorer in 1995.

Search Engine Early History

- By late 1980's many files were available by anonymous FTP.
- In 1990, Alan Emtage of McGill Univ. developed Archie (short for "archives")
  - Assembled lists of files available on many FTP servers.
  - Allowed regex search of these file names.
- In 1993, Veronica and Jughead were developed to search names of text files available through Gopher servers.

Web Search History

- In 1993, early web robots (spiders) were built to collect URL’s:
  - Wanderer
  - ALIWEB (Archie-Like Index of the WEB)
  - WWW Worm (indexed URL’s and titles for regex search)
- In 1994, Stanford grad students David Filo and Jerry Yang started manually collecting popular web sites into a topical hierarchy called Yahoo.

Web Search History (cont.)

- In early 1994, Brian Pinkerton developed WebCrawler as a class project at U Wash. (eventually became part of Excite and AOL).
- A few months later, Fuzzy Maudlin, a grad student at CMU developed Lycos. First to use a standard IR system as developed for the DARPA Tipster project. First to index a large set of pages.
- In late 1995, DEC developed Altavista. Used a large farm of Alpha machines to quickly process large numbers of queries. Supported boolean operators, phrases, and "reverse pointer" queries.

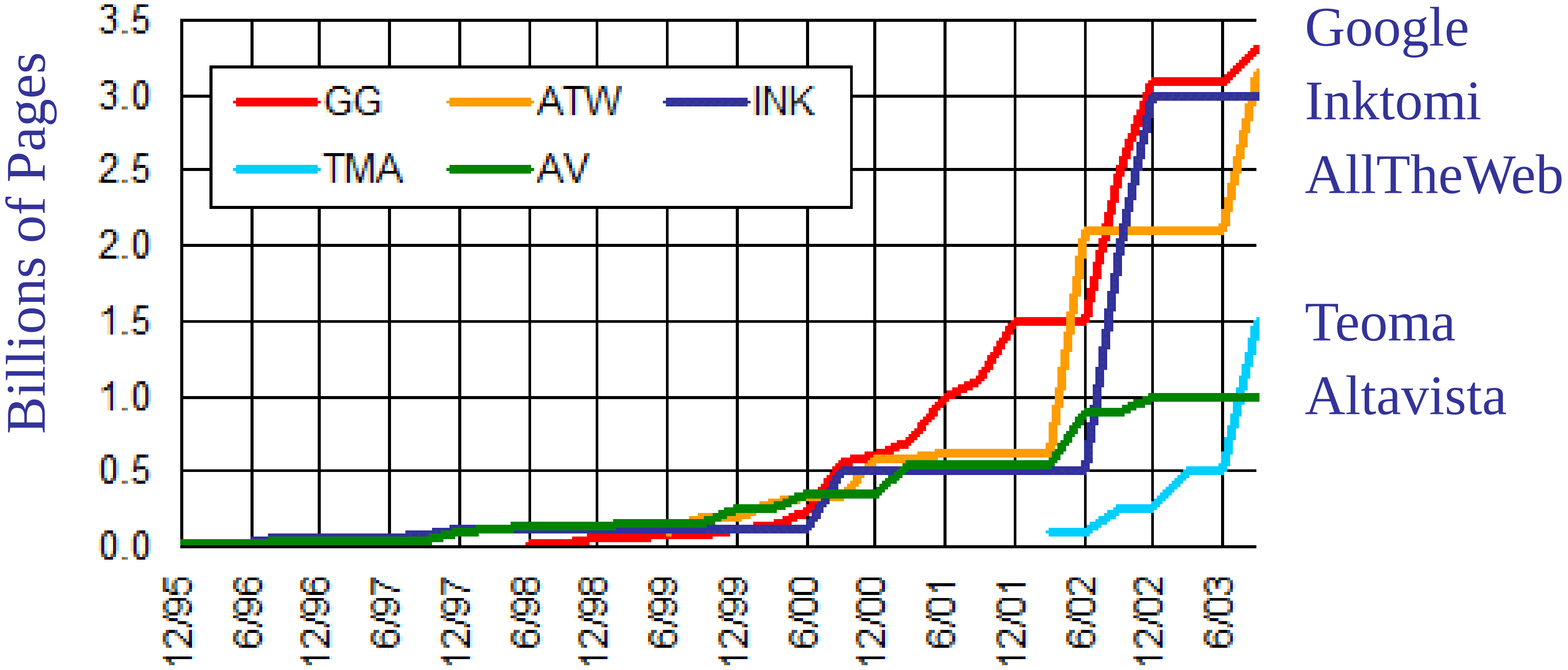
Web Search History (cont.)

- In 1998, Larry Page and Sergey Brin, Ph.D. students at Stanford, started Google. Main advance is use of *link analysis* to rank results partially based on authority.
- Microsoft lauched MSN Search in 1998 based on Inktomi (started from UC Berkeley in 1996), changed to Live Search in 2007, and Bing in 2009.

Web Challenges for IR

- Distributed Data: Documents spread over millions of different web servers.
- Volatile Data: Many documents change or disappear rapidly (e.g. dead links).
- Large Volume: Billions of separate documents.
- Unstructured and Redundant Data: No uniform structure, HTML errors, up to 30% (near) duplicate documents.
- Quality of Data: No editorial control, false information, poor quality writing, typos, etc.
- Heterogeneous Data: Multiple media types (images, video, VRML), languages, character sets, etc.

Growth of Web Pages Indexed



SearchEngineWatch

Assuming 20KB per page,
1 billion pages is about 20 terabytes of data.

“Small World” (Scale-Free) Graphs

- Social networks and six degrees of separation.
  - Stanley Milgram Experiment
- Power law distribution of in and out degrees.
- Distinct from purely random graphs.
- “Rich get richer” generation of graphs (preferential attachment).
- Kevin Bacon game.
  - Oracle of Bacon
- Erdos number.
- Networks in biochemistry, roads, telecommunications, Internet, etc are “small word”

Manual Hierarchical Web Taxonomies

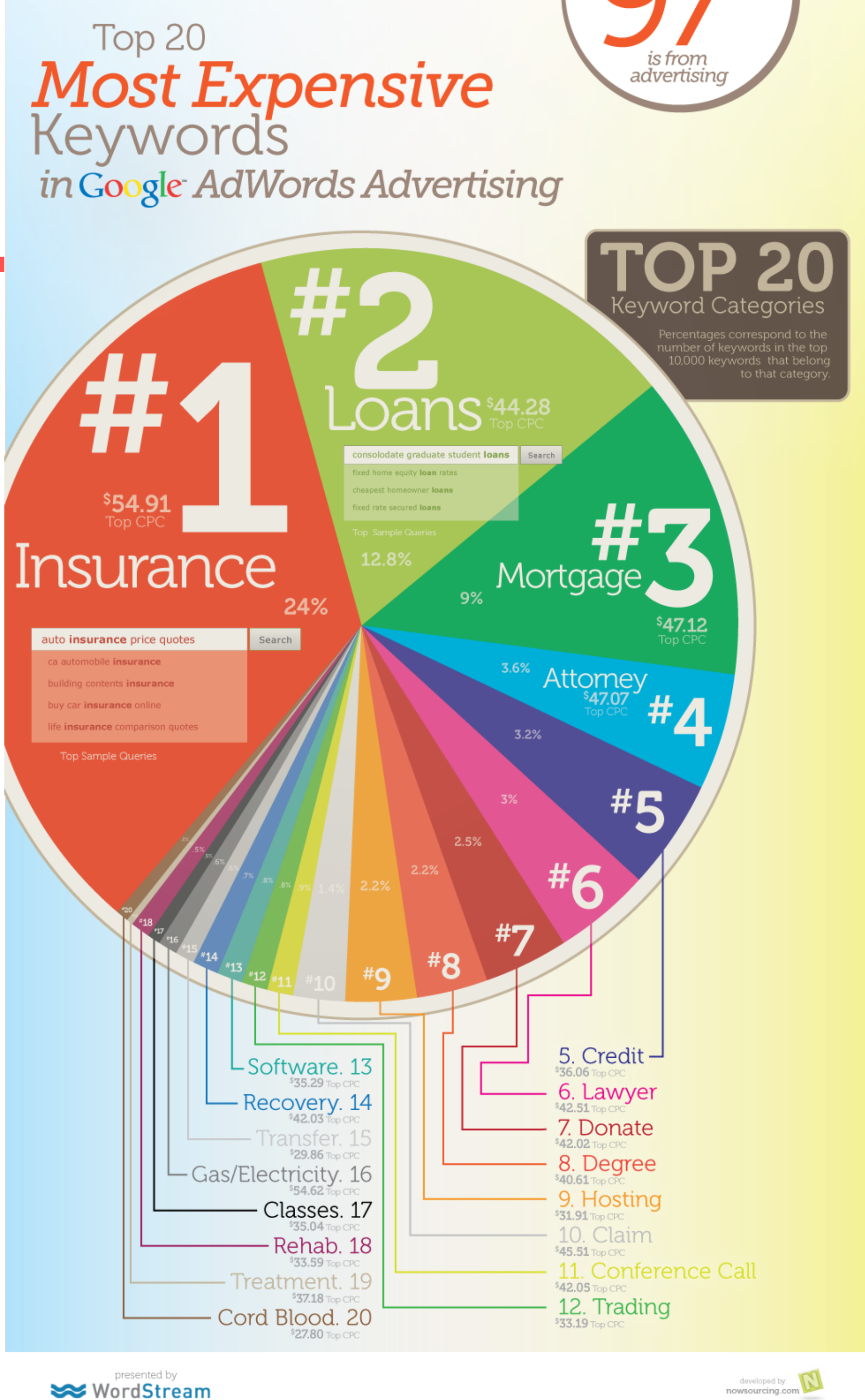
- Yahoo approach of using human editors to assemble a large hierarchically structured directory of web pages (closed in 2014).
- Open Directory Project is a similar approach based on the distributed labor of volunteer editors (“net-citizens provide the collective brain”). Used by most other search engines. Started by Netscape.
  - http://www.dmoz.org/

Business Models for Web Search

- Advertisers pay for banner ads on the site that do not depend on a user’s query.
  - CPM: Cost Per Mille (thousand impressions). Pay for each ad display.
  - CPC: Cost Per Click. Pay only when user clicks on ad.
  - CTR: Click Through Rate. Fraction of ad impressions that result in clicks throughs. CPC = CPM / (CTR * 1000)
  - CPA: Cost Per Action (Acquisition). Pay only when user actually makes a purchase on target site.
- Advertisers bid for “keywords”. Ads for highest bidders displayed when user query contains a purchased keyword.
  - PPC: Pay Per Click. CPC for bid word ads (e.g. Google AdWords).

History of Business Models

- Initially, banner ads paid thru CPM were the norm.
- GoTo Inc. formed in 1997 and originates and patents bidding and PPC business model.
- Google introduces AdWords in fall 2000.
- GoTo renamed Overture in Oct. 2001.
- Overture sues Google for use of PPC in Apr. 2002.
- Overture acquired by Yahoo in Oct. 2003.
- Google settles with Overture/Yahoo for 2.7 million shares of Class A common stock in Aug. 2004.



Top 20
Most Expensive
Keywords
in Google AdWords Advertising
97
is from advertising
TOP 20
Keyword Categories
Percentages correspond to the number of keywords in the top 10,000 keywords that belong to that category.
#1
$54.91 Top CPC
Insurance
24%
auto insurance price quotes
Search
ca automobile insurance
building contents insurance
buy car insurance online
life insurance comparison quotes
Top Sample Queries
#2
Loans
$44.28 Top CPC
consolodate graduate student loans
Search
fixed home equity loan rates
cheapest homeowner loans
fixed rate secured loans
Top Sample Queries
12.8%
#3
Mortgage
9%
$47.12 Top CPC
Attorney
3.6%
$47.07 Top CPC
#4
3.2%
#5
3%
#6
2.5%
#7
2.2%
#8
2.2%
#9
1.4%
#10
.9%
#11
.8%
#12
.8%
#13
.7%
#14
.6%
#15
.5%
#16
.5%
#17
#18
#20
5. Credit
$36.06 Top CPC
6. Lawyer
$42.51 Top CPC
7. Donate
$42.02 Top CPC
8. Degree
$40.61 Top CPC
9. Hosting
$31.91 Top CPC
10. Claim
$45.51 Top CPC
11. Conference Call
$42.05 Top CPC
12. Trading
$33.19 Top CPC
Software. 13
$35.29 Top CPC
Recovery. 14
$42.03 Top CPC
Transfer. 15
$29.86 Top CPC
Gas/Electricity. 16
$54.62 Top CPC
Classes. 17
$35.04 Top CPC
Rehab. 18
$33.59 Top CPC
Treatment. 19
$37.18 Top CPC
Cord Blood. 20
$27.80 Top CPC
presented by
WordStream
developed by
nowsourcing.com

Affiliates Programs

- If you have a website, you can generate income by becoming an *affiliate* by agreeing to post ads relevant to the topic of your site.
- If users click on your impression of an ad, you get some percentage of the CPC or PPC income that is generated.
- Google introduces AdSense affiliates program in 2003.

Web Search

Advances &
Link Analysis

Meta-Search Engines

- Search engine that passes query to several other search engines and integrate results.
  - Submit queries to host sites.
  - Parse resulting HTML pages to extract search results.
  - Integrate multiple rankings into a “consensus” ranking.
  - Present integrated results to user.
- Examples:
  - Metacrawler
  - SavvySearch
  - Dogpile

HTML Structure & Feature Weighting

- Weight tokens under particular HTML tags more heavily:
  - <TITLE> tokens (Google seems to like title matches)
  - <H1>,<H2>… tokens
  - <META> keyword tokens
- Parse page into conceptual sections (e.g. navigation links vs. page content) and weight tokens differently based on section.

Bibliometrics: Citation Analysis

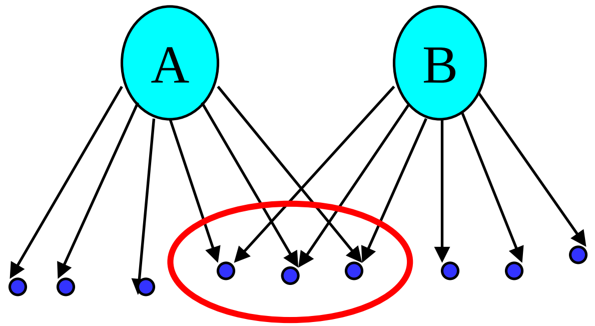
- Many standard documents include *bibliographies* (or *references*), explicit *citations* to other previously published documents.
- Using citations as links, standard corpora can be viewed as a graph.
- The structure of this graph, independent of content, can provide interesting information about the similarity of documents and the structure of information.
- CF corpus includes citation information.

Impact Factor

- Developed by Garfield in 1972 to measure the importance (quality, influence) of scientific journals.
- Measure of how often papers in the journal are cited by other scientists.
- Computed and published annually by the Institute for Scientific Information (ISI).
- The *impact factor* of a journal *J* in year *Y* is the average number of citations (from indexed documents published in year *Y*) to a paper published in *J* in year *Y*–1 or *Y*–2.
- Does not account for the quality of the citing article.

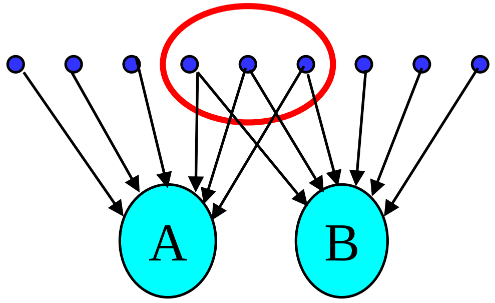
Bibliographic Coupling

- Measure of similarity of documents introduced by Kessler in 1963.
- The bibliographic coupling of two documents *A* and *B* is the number of documents cited by *both A* and *B*.
- Size of the intersection of their bibliographies.
- Maybe want to normalize by size of bibliographies?



Co-Citation

- An alternate citation-based measure of similarity introduced by Small in 1973.
- Number of documents that cite both *A* and *B*.
- Maybe want to normalize by total number of documents citing either *A* or *B* ?



Citations vs. Links

- Web links are a bit different than citations:
  - Many links are navigational.
  - Many pages with high in-degree are portals not content providers.
  - Not all links are endorsements.
  - Company websites don’t point to their competitors.
  - Citations to relevant literature is enforced by peer-review.

Authorities

- *Authorities* are pages that are recognized as providing significant, trustworthy, and useful information on a topic.
- *In-degree* (number of pointers to a page) is one simple measure of authority.
- However in-degree treats all links as equal.
- Should links from pages that are themselves authoritative count more?

Hubs

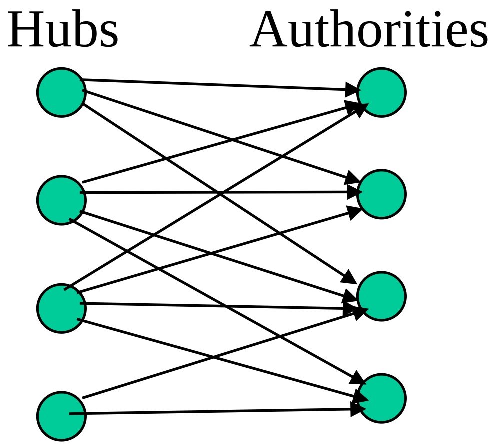
- *Hubs* are index pages that provide lots of useful links to relevant content pages (topic authorities).
- Hub pages for IR are included in the course home page:
  - http://www.cs.utexas.edu/users/mooney/ir-course

HITS

- Algorithm developed by Kleinberg in 1998.
- Attempts to computationally determine hubs and authorities on a particular topic through analysis of a relevant subgraph of the web.
- Based on mutually recursive facts:
  - Hubs point to lots of authorities.
  - Authorities are pointed to by lots of hubs.

Hubs and Authorities

- Together they tend to form a bipartite graph:

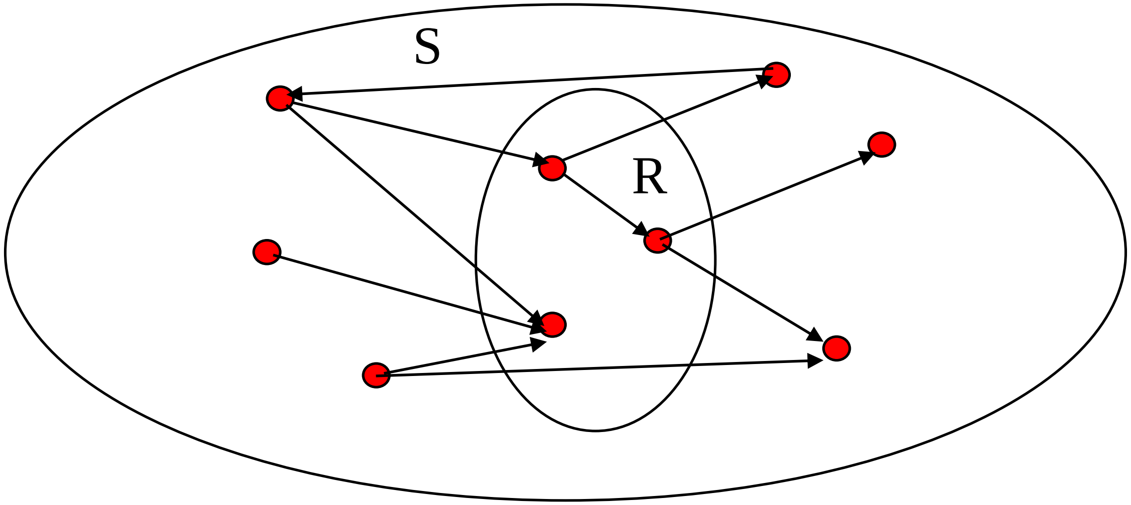


HITS Algorithm

- Computes hubs and authorities for a particular topic specified by a normal query.
- First determines a set of relevant pages for the query called the *base* set *S*.
- Analyze the link structure of the web subgraph defined by *S* to find authority and hub pages in this set.

Constructing a Base Subgraph

- For a specific query *Q*, let the set of documents returned by a standard search engine (e.g. VSR) be called the *root* set *R*.
- Initialize *S* to *R*.
- Add to *S* all pages pointed to by any page in *R*.
- Add to *S* all pages that point to any page in *R*.



Base Limitations

- To limit computational expense:
  - Limit number of root pages to the top 200 pages retrieved for the query.
  - Limit number of "back-pointer" pages to a random set of at most 50 pages returned by a "reverse link" query.
- To eliminate purely navigational links:
  - Eliminate links between two pages on the same host.
- To eliminate "non-authority-conveying" links:
  - Allow only *m* ($m \cong 4$–8) pages from a given host as pointers to any individual page.

Authorities and In-Degree

- Even within the base set *S* for a given query, the nodes with highest in-degree are not necessarily authorities (may just be generally popular pages like Yahoo or Amazon).
- True authority pages are pointed to by a number of hubs (i.e. pages that point to lots of authorities).

Results

- Authorities for query: "Java"
  - java.sun.com
  - comp.lang.java FAQ
- Authorities for query "search engine"
  - Yahoo.com
  - Excite.com
  - Lycos.com
  - Altavista.com
- Authorities for query "Gates"
  - Microsoft.com
  - roadahead.com

Result Comments

- In most cases, the final authorities were not in the initial root set generated using Altavista.
- Authorities were brought in from linked and reverse-linked pages and then HITS computed their high authority score.

Finding Similar Pages Using Link Structure

- Given a page, *P*, let *R* (the root set) be *t* (e.g. 200) pages that point to *P*.
- Grow a base set *S* from *R*.
- Run HITS on *S*.
- Return the best authorities in *S* as the best similar-pages for *P*.
- Finds authorities in the "link neighbor-hood" of *P*.

Similar Page Results

- Given "honda.com"
  - toyota.com
  - ford.com
  - bmwusa.com
  - saturncars.com
  - nissanmotors.com
  - audi.com
  - volvocars.com

HITS for Clustering

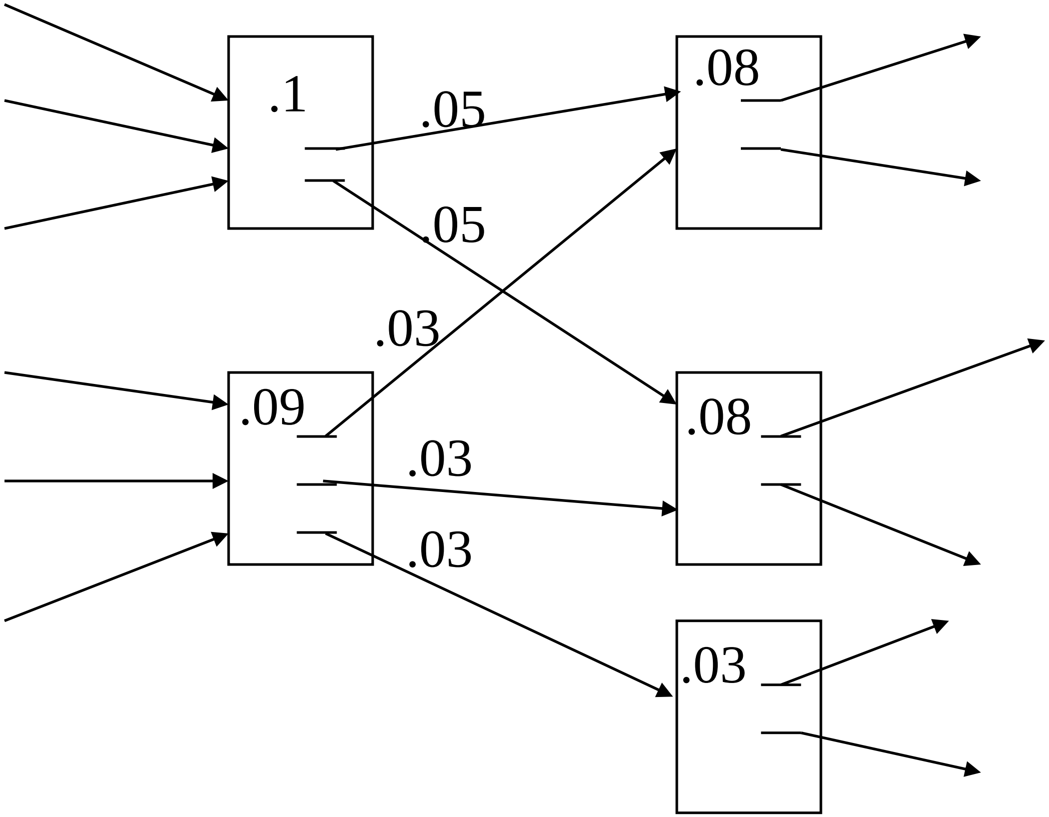
- An ambiguous query can result in the principal eigenvector only covering one of the possible meanings.
- Non-principal eigenvectors may contain hubs & authorities for other meanings.
- Example: “jaguar”:
  - Atari video game (principal eigenvector)
  - NFL Football team (2nd non-princ. eigenvector)
  - Automobile (3rd non-princ. eigenvector)

PageRank

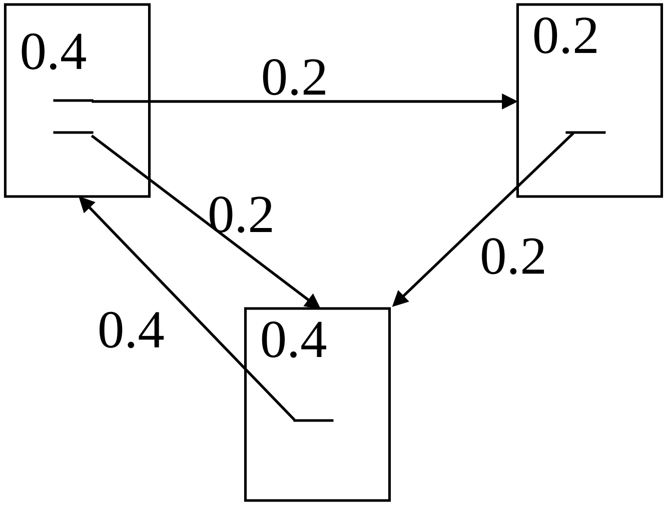
- Alternative link-analysis method used by Google (Brin & Page, 1998).
- Does not attempt to capture the distinction between hubs and authorities.
- Ranks pages just by authority.
- Applied to the entire web rather than a local neighborhood of pages surrounding the results of a query.

Initial PageRank Idea (cont.)

- Can view it as a process of PageRank "flowing" from pages to the pages they cite.

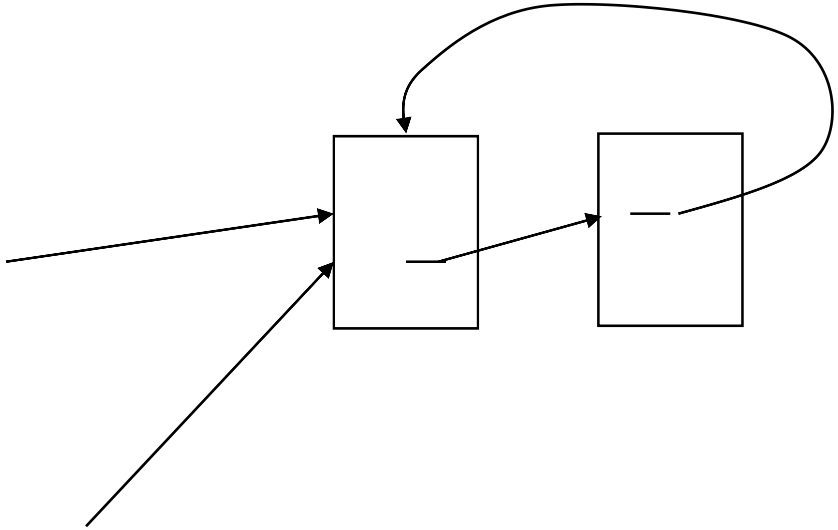


Sample Stable Fixpoint



Problem with Initial Idea

- A group of pages that only point to themselves but are pointed to by other pages act as a "rank sink" and absorb all the rank in the system.



Rank flows into
cycle and can't get out

Random Surfer Model

- PageRank can be seen as modeling a "random surfer" that starts on a random page and then at each point:
  - With probability *E*(*p*) randomly jumps to page *p*.
  - Otherwise, randomly follows a link on the current page.
- *R*(*p*) models the probability that this random surfer will be on page *p* at any given time.
- "E jumps" are needed to prevent the random surfer from getting "trapped" in web sinks with no outgoing links.

Speed of Convergence

- Early experiments on Google used 322 million links.
- PageRank algorithm converged (within small tolerance) in about 52 iterations.
- Number of iterations required for convergence is empirically O(log *n*) (where *n* is the number of links).
- Therefore calculation is quite efficient.

Simple Title Search with PageRank

- Use simple Boolean search to search web-page titles and rank the retrieved pages by their PageRank.
- Sample search for “university”:
  - Altavista returned a random set of pages with “university” in the title (seemed to prefer short URLs).
  - Primitive Google returned the home pages of top universities.

Google Ranking

- Complete Google ranking includes (based on university publications prior to commercialization).
  - Vector-space similarity component.
  - Keyword proximity component.
  - HTML-tag weight component (e.g. title preference).
  - PageRank component.
- Details of current commercial ranking functions are trade secrets.

Personalized PageRank

- PageRank can be biased (personalized) by changing **E** to a non-uniform distribution.
- Restrict “random jumps” to a set of specified relevant pages.
- For example, let $E(p) = 0$ except for one’s own home page, for which $E(p) = \alpha$
- This results in a bias towards pages that are closer in the web graph to your own homepage.

Google PageRank-Biased Spidering

- Use PageRank to direct (focus) a spider on "important" pages.
- Compute page-rank using the current set of crawled pages.
- Order the spider's search queue based on current estimated PageRank.

Link Analysis Conclusions

- Link analysis uses information about the structure of the web graph to aid search.
- It is one of the major innovations in web search.
- It was one of the primary reasons for Google's initial success.