Web search

Dan Smith

University of East Anglia

Today's topics

- Web search
- Bibliometric analysis
- Link analysis
- PageRank

3

WEB SEARCH

4

Search engines let the web scale

- Content must be findable
- Search engines
  - let users find content
  - facilitate niche markets
  - facilitate very specialised interest groups
    - but erode common cultural experiences and views
  - enable "infinite product" stores
    - Amazon, …
- Search is the best advertising channel on the web

5

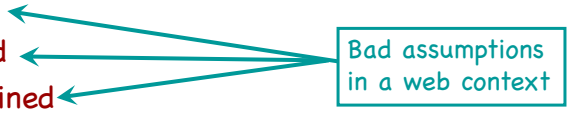
Basic IR assumptions

- Corpus: Fixed document collection
- Goal: Retrieve documents with content relevant to user's information need

6

Reminder: classic IR goal

- Classic relevance
  - For each query Q and document D in a given corpus assume there exists relevance score(Q, D)
    - score is averaged over users U and contexts C
  - Optimize score(Q, D) as opposed to Score(Q, D, U, C)
  - That is, usually:
    - Context ignored
    - Individuals ignored
    - Corpus predetermined



7

Web search

- Context is important
- User history is important (and available)
- Corpus is (effectively) infinite
  - Google estimated the Web has about 60TN pages (Nov. 2013)
  - Google's index is about 100TB
- Google (Apr 2013)
  - indexes approx. 45bn pages
  - 100 billion queries/month

8

User goals in search

- Informational c. 80%
  - advice
  - directed
  - undirected
  - question answering
- Transactional c. 10%
  - downloads
  - purchases
- Navigational c. 10%

Bernard J. Jansen, Danielle L. Booth, Amanda Spink:
Determining the informational, navigational, and transactional intent of Web queries.
Inf. Process. Manage. (IPM) 44(3):1251-1266 (2008)

9

Query characteristics

- Most queries are short
  - average c. 3 terms/query
- Most queries are imprecise
  - suggests users want to see a wide range of results
- Most queries are not modified
  - approximately half of all web search sessions have a single query (data c. 2002)
  - more recent results suggest over 75% of queries are not modified
- Few people use advanced search options

10

Viewing query results

- Users looked at one page of query results in about half the cases (1998-2002 data)
- More recent work shows that click-throughs decline very rapidly as the user scans the first few results
  - c. 85% of search results only scanned to the fold
  - but stabilise once they go below the fold (i.e. have to start scrolling)

11

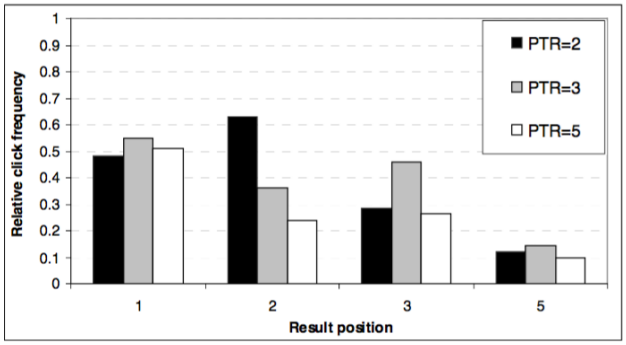
The importance of rank order

Laura A. Granka, Thorsten Joachims, Geri Gay: Eye-tracking analysis of user behavior in WWW search. SIGIR 2004:478-479

12

Relevance and position

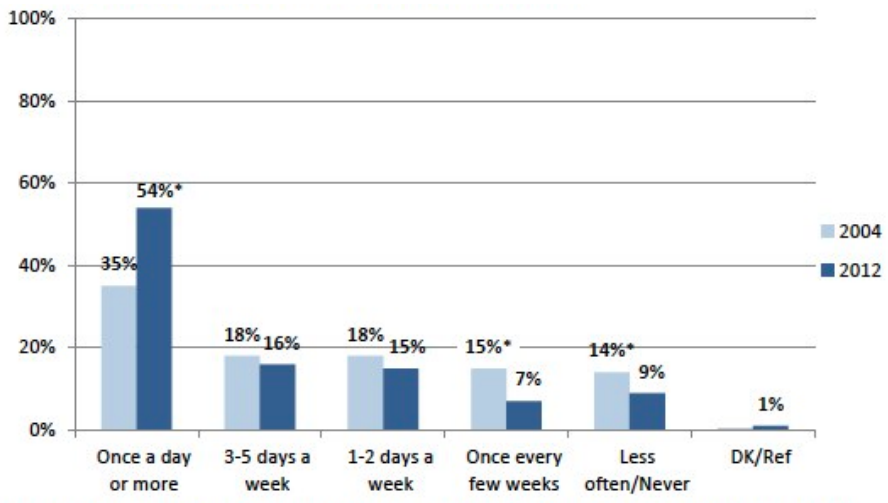
- Users are heavily biased in favour of highly ranked search results
  - experiment with top-ranked result in different positions (PTR) and the results selected by users



E. Agichtein, E. Brill, S. Dumais: Improving web search ranking by incorporating user behavior information. SIGIR 2006:19-26

13

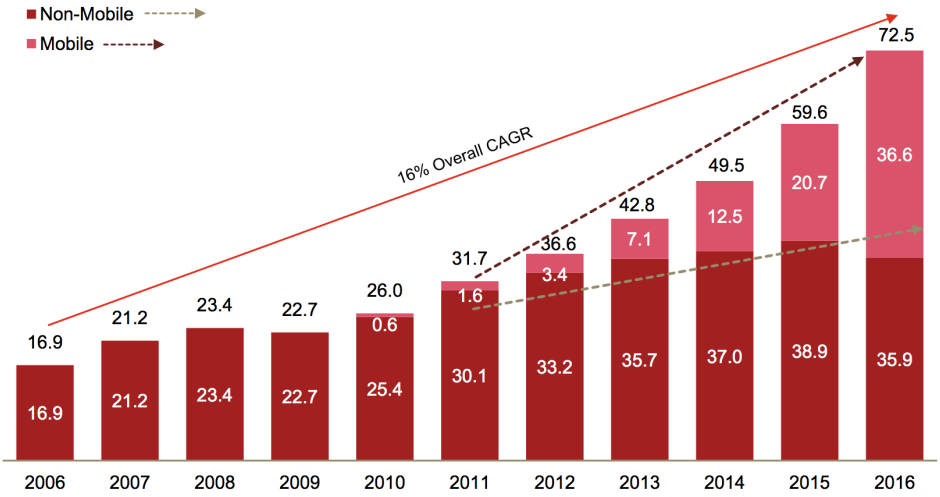
Search use …



Kristen Purcell, Joanna Brenner, Lee Rainie (2012) **Search Engine Use 2012**
http://pewinternet.org/Reports/2012/Search-Engine-Use-2012/Main-findings.aspx?view=all

14

Internet advertising revenues (USA)



IAB internet advertising revenue report : 2016 full year results
https://www.iab.com/insights/iab-internet-advertising-revenue-report-conducted-by-pricewaterhousecoopers-pwc-2/

15

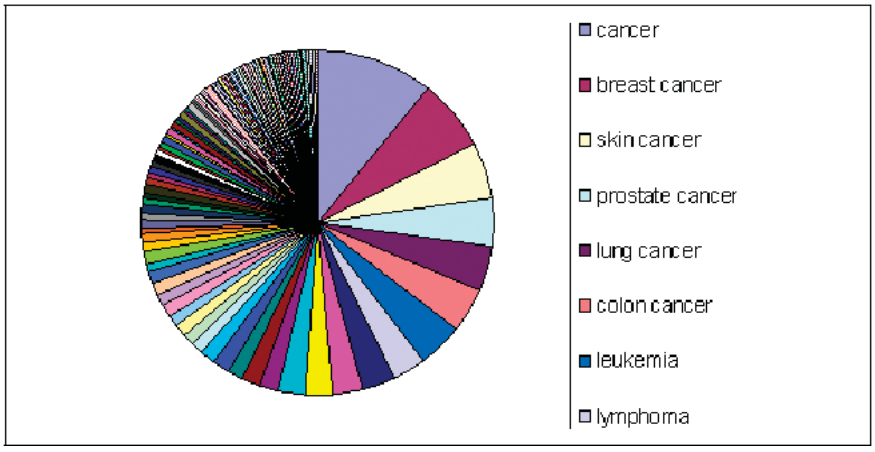
Internet advertising spend - UK

- Total digital advertising spend approx. £8bn in 2015
  - 13% annual growth
  - Display advertising growing at 25%
- Mobile advertising £2.15bn, 50% growth
  - accounts for most of the growth
  - 39% of display spend
  - 43% of video spend
  - 63% of social media spend
  - 74% of native/content ad spend

http://www.iabuk.net/about/press/archive/uk-digital-display-advertising-revenues-rise-275

16

Query distribution



Power law: few popular broad queries, many rare specific queries

17

User evaluation of search results (1)

- Quality of pages varies widely
  - relevance is not enough
- Readability
  - display correctly, fast
  - no annoyances: pop-ups, etc.
- Desirable content characteristics
  - trustworthy
  - new information
  - non-duplicate pages
  - current, well maintained,

18

User evaluation of search results (2)

- What matters
  - Precision at 1? Precision above the fold?
  - Comprehensiveness – must be able to deal with obscure queries
    - 15% of all queries are unique
  - Recall only matters when there are few matches
  - User perceptions may be unscientific, but are significant over a large aggregate

19

Answering the need behind the query

- Semantic analysis
  - Query language determination
    - Auto filtering
    - Different ranking
      (e.g. if query is in Japanese do not return English)
  - Hard and soft (partial) matches
    - Personalities (triggered on names)
    - Cities (travel info, maps)
    - Medical info (triggered on names and/or results)
    - Stock quotes, news (triggered on stock symbol)
    - Company info
    - …
  - Natural language reformulation
  - Integration with text analysis

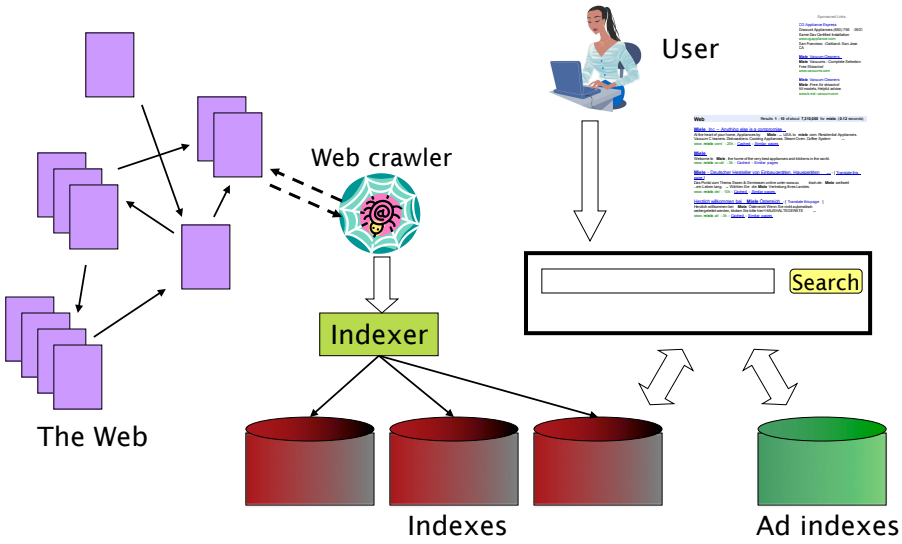
20

The need behind the query: context

- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search,)
  - implicit (use Google from France, use google.fr)
- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a "rough" generic ranking, but personalize later

21

Web search basics



22

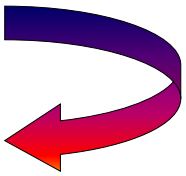
Search engine history: 1990s

- Keyword-based engines
  - Altavista, Excite, Infoseek, Inktomi,
- Paid placement ranking:
  - Goto.com (became Overture, became Yahoo!)
  - Search ranking depended on how much you paid
- Problem: easy to spam
  - By the late 1990s Altavista results were very heavily polluted by spam

23

Keyword spamming for the web

- Early search engines relied largely on *tf*idf* weighting
  - repeating terms increases the document's ranking
  - easy to do by keyword stuffing
- Keyword stuffing techniques
  - repetition in <meta> tags
  - text hidden by making it the same colour as the background
  - invisible to humans but visible to crawlers



Pure word density cannot be trusted as an IR signal for web search

24

Keyword spamming example

```
<html>
  <head>
  <meta name="keywords" content="mp3 free download mp3 free download mp3 free
download mp3 free download mp3 free download mp3 free download " />
  <title>UEALite stemmer: overview</title>
</head>

<body bgcolor="white">
  <font color="white" size="1pt">
  euromillions euromillions euromillions euromillions euromillions euromillions euromillions
  euromillions euromillions euromillions euromillions euromillions euromillions euromillions
  euromillions euromillions euromillions euromillions euromillions euromillions euromillions
  </font><br>
  <h3>UEALite Overview</h3>
    Similar to other stemmers, UEA-Lite operates on a set of rules
    which are used as steps…

</body>
</html>
```

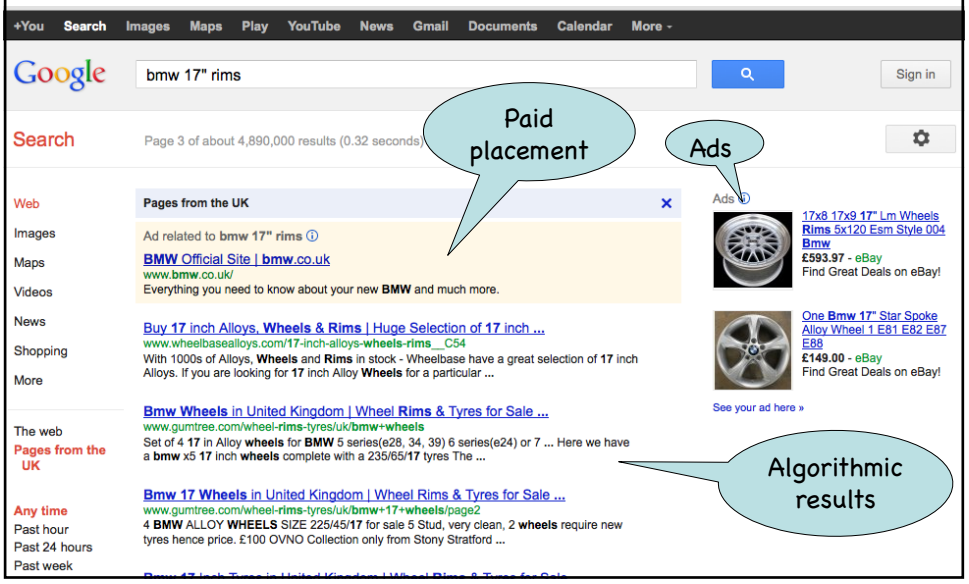
25

Search engine history: Google

- Google used link-based ranking (1998)
  - quickly became dominant
- Great user experience looking for a business model
  - Overture's annual revenues were nearly $1bn
- Result: in 2000 Google added paid placement independent of search results
  - AdWords is the main contributor to Google's advertising revenue, $43.7bn in 2012

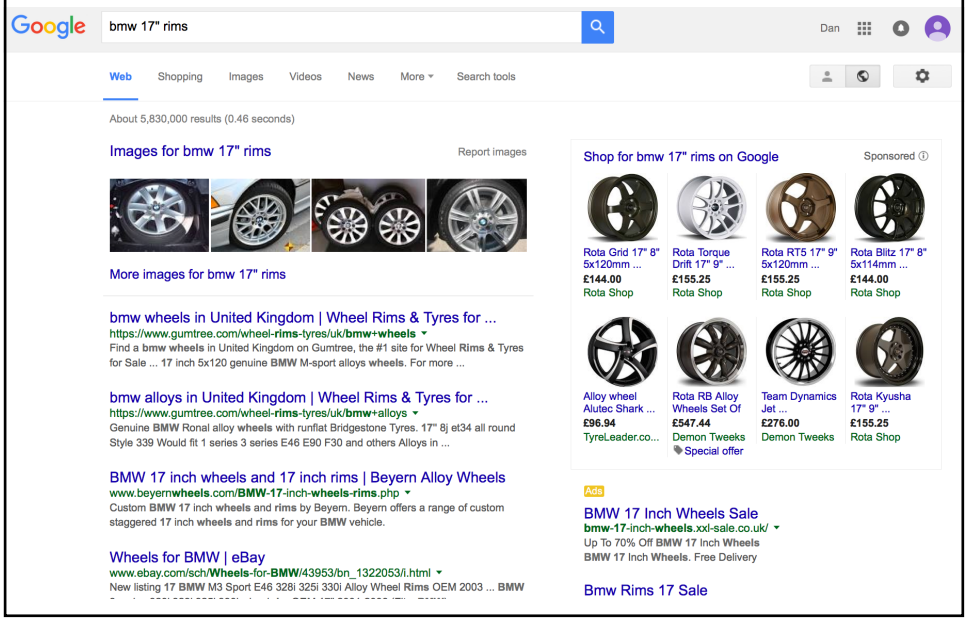
26

Google results (2012)



27

Google results (2015)



28

before the Web, before Google

BIBLIOMETRIC ANALYSIS

29

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 graphs
- Structure of the graph can provide interesting information about
  - similarity of documents
  - structure of information
- Graph structure is independent of content

30

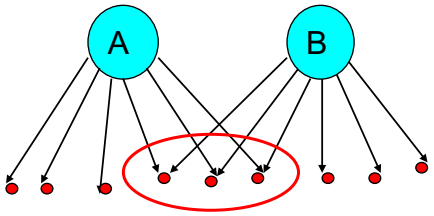
Impact Factor

- Developed by Garfield in 1972 to measure the importance (quality, influence) of scientific journals
- Measure of how often papers are cited by others
- Computed and published annually
  - Institute for Scientific Information (ISI)
  - Thompson Reuters (Web of Science)
- 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
  - the number of citations expected in a subject area

31

Bibliographic Coupling

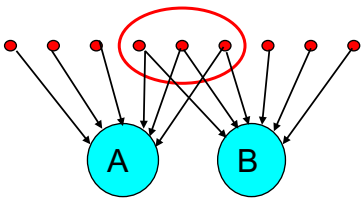
- Measure of similarity of documents introduced by Kessler in 1963.
- The bibliographic coupling of two documents *A* and *B* is the number of documents cited by *both A* and *B*.
- Size of the intersection of their bibliographies.
- Maybe want to normalize by size of bibliographies?



32

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* ?



33

LINK ANALYSIS

34

Citations vs. Links

- Web links are a bit different to citations:
  - Many links are navigational
  - Many pages with many incoming links are portals, not content providers
  - Not all links (or citations) are endorsements
  - Company websites don't point to their competitors
  - Citation of relevant academic literature is enforced by peer review
    - there's no universal peer review on the Web

35

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
  - issue: in-degree treats all links as equal
- Should links from pages that are themselves authoritative count more?

36

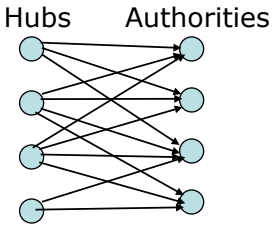
Hubs

- *Hubs* are pages that provide lots of links to relevant content pages (authorities)
- Hub pages for computing science
  http://dblp.org/search/index.php
  http://www.computer.org/csdl
  http://www.acm.org/
- Hub pages for IR
  http://www.webir.org/
  http://www-nlpir.nist.gov/projects/irlib/collection.html
  http://www-csli.stanford.edu/~hinrich/information-retrieval.html

37

Hubs and Authorities

- Hubs point to lots of authorities
- Authorities are pointed to by lots of hubs



38

PAGERANK

39

PageRank

- Link-analysis method used by Google (Brin and Page 1998)
- Key insights
  - Just measuring in-degree (citation count) ignores the authority of the source of a link
  - Do 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

40

Simplified PageRank

- Initial page rank equation for page *p*:

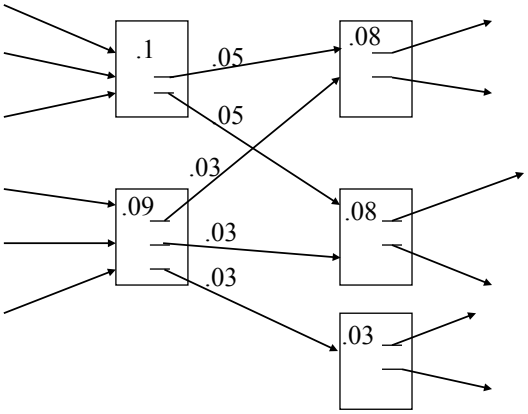
$$R(p) = c \sum_{q:q \to p} \frac{R(q)}{N_q}$$

  - N_q is the total number of out-links from page *q*
  - page *q* "gives" an equal fraction of its authority to all the pages it points to (e.g. *p*).
  - *c* is a normalizing constant set so that the rank of all pages always sums to 1

41

Simplified PageRank

- Can view it as a process of PageRank "flowing" from each page to the pages it cites



42

Initial algorithm

- Iterate rank-flowing process until convergence:

Let *S* be the total set of pages.
Initialize $\forall p \in S$: $R(p) = 1/|S|$
Until ranks do not change (much) (*convergence*)

For each $p \in S$: $R'(p) = \sum_{q:q \to p} \frac{R(q)}{N_q}$

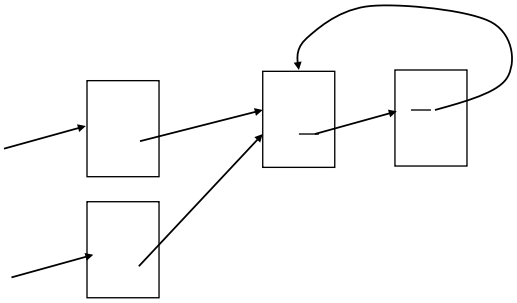
$$c = 1 / \sum_{p \in S} R'(p)$$

For each $p \in S$: $R(p) = cR'(p)$ (*normalize*)

43

Problem: rank sink

- 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 a cycle and can't get out
- Need to escape from the sink

44

Random Surfer interpretation

- PageRank can be seen as modeling a "random surfer" that starts on a random page and then at each point:
  - With probability $E(p)$ the surfer gets bored and 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

45

Rank source

- To escape from rank sinks, PR has a rank source E that replenishes the rank of each page, p, by a fixed amount $E(p)$ on each iteration

$$R(p) = c\left(\sum_{q:q\to p} \frac{R(q)}{N_q} + E(p)\right)$$

46

PageRank algorithm

Let S be the total set of pages.
Let $\forall p \in S$: $E(p) = \alpha/|S|$ (for some $0<\alpha<1$, e.g. 0.15)
Initialize $\forall p \in S$: $R(p) = 1/|S|$
Until ranks do not change (much) (*convergence*)

For each $p \in S$:

$$R'(p) = \sum_{q:q\to p} \frac{R(q)}{N_q} + E(p)$$

$$c = 1/\sum_{p\in S} R'(p)$$

For each $p \in S$: $R(p) = cR'(p)$ (*normalize*)

47

Speed of convergence

- Early experiments on Google used 322 million links
- PageRank algorithm converged 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

48

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

49

Google Ranking

- Complete Google ranking includes*
  - 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
  - Discussed at http://www.searchenginewatch.com

* based on the university version, pre-commercialisation

50

Google PageRank-biased crawling

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

51

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
- PageRank was the primary reason for Google's success

52

References

- Manning et al. 2008 Chapters 19 (up to 19.6) and 21 (up to 21.2.3)
- Wicker S., Karlsson K. (2017) Internet Advertising: Technology, Ethics, and a Serious Difference of Opinion, *CACM* 60(10), 70-79 https://dl.acm.org/citation.cfm?doid=3048384
  - a good account of Internet advertising networks