Web Search

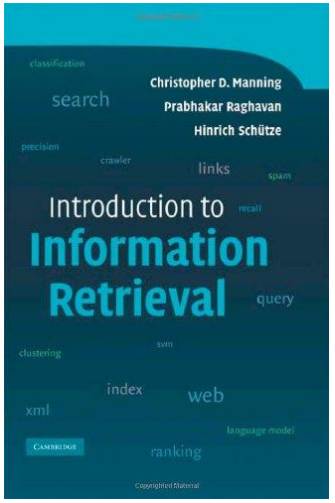
Allan Hanbury

Announcement

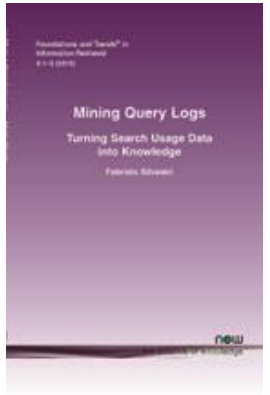
- Music and Image Retrieval Lecture on 12 December 2019
- Usual time and place

Material covered

- Manning Book
  - **Chapter 19: Web Search Basics**
  - **Chapter 20: Web Crawling and Indexing**
  - **Chapter 21: Link Analysis**



- Manning Book slides have been widely used and adapted for this lecture



- F. Silvestri, Mining Query Logs: Turning Search Usage Data into Knowledge, now Publishers, 2009
  http://ferryas.lecturer.pens.ac.id/Others/PA/web%20search%20mining.pdf

Outline

① The Web: What makes it Unique?

② Indexing the Web

③ Ranking on the Web

④ Query Log Analysis

Outline

① The Web: What makes it Unique?

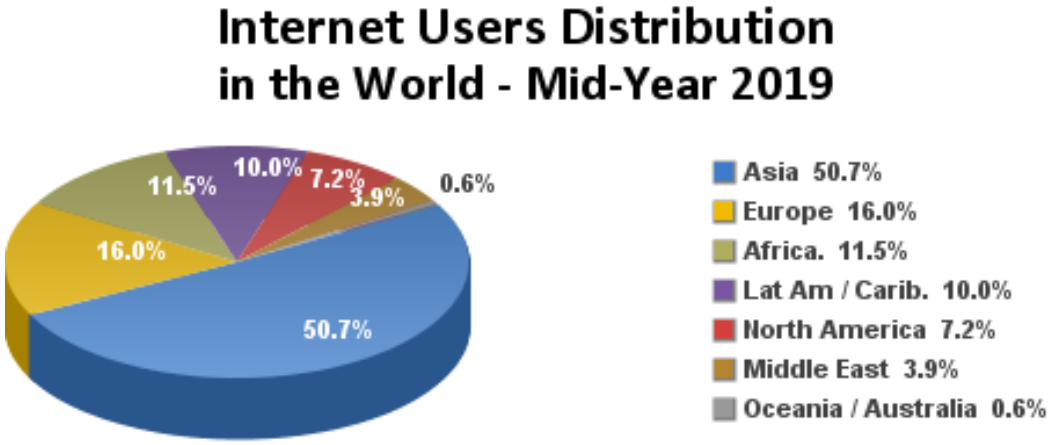
② Indexing the Web

③ Ranking on the Web

④ Query Log Analysis

Internet Today

4.5 billion people are connected



Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 4,536,248,808 Internet users in June 30, 2019
Copyright © 2019, Miniwatts Marketing Group

http://www.internetworldstats.com/stats.htm

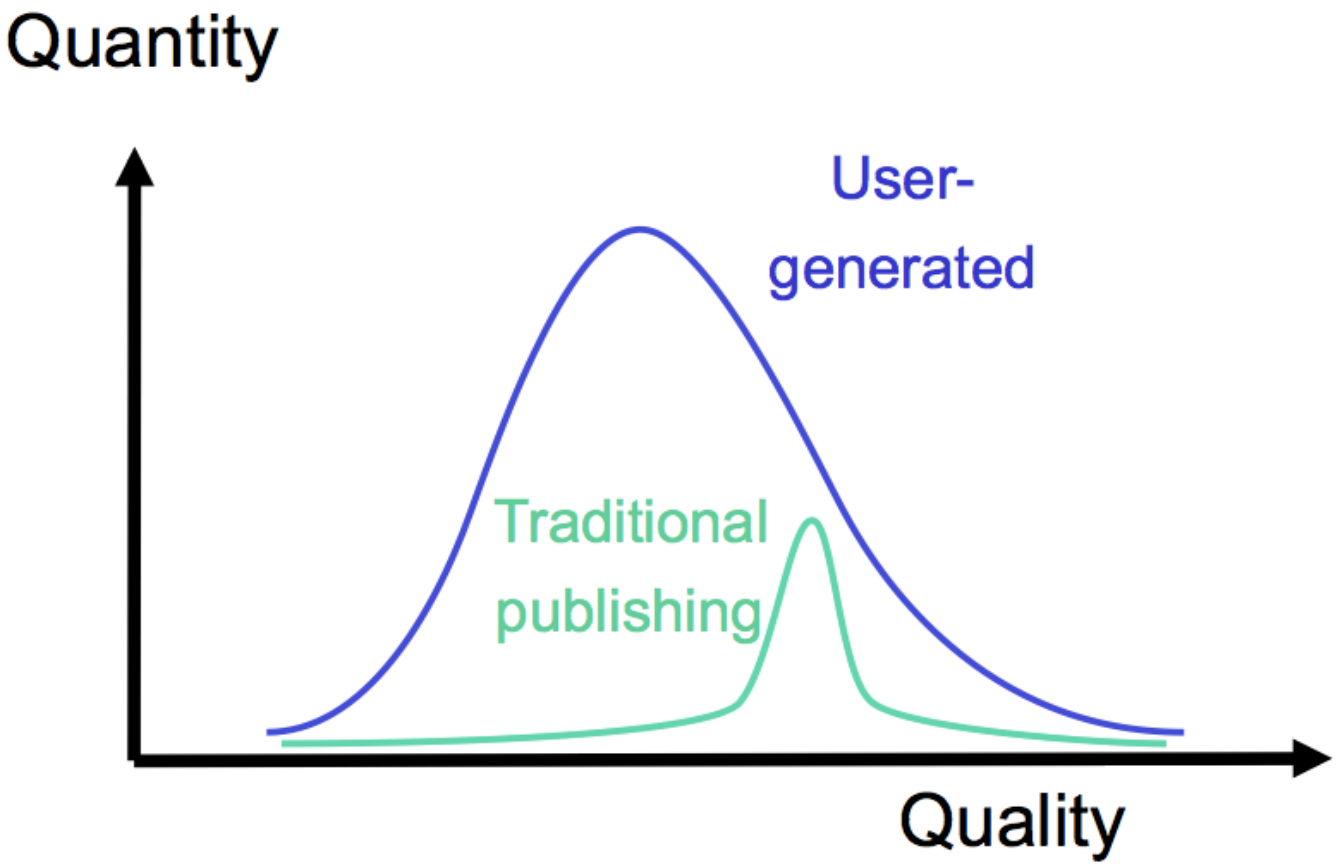
WORLD INTERNET USAGE AND POPULATION STATISTICS 2019 Mid-Year Estimates						
World Regions	Population (2019 Est.)	Population % of World	Internet Users 30 June 2019	Penetration Rate (% Pop.)	Growth 2000-2019	Internet World %
Africa	1,320,038,716	17.1 %	522,809,480	39.6 %	11,481 %	11.5 %
Asia	4,241,972,790	55.0 %	2,300,469,859	54.2 %	1,913 %	50.7 %
Europe	829,173,007	10.7 %	727,559,682	87.7 %	592 %	16.0 %
Latin America / Caribbean	658,345,826	8.5 %	453,702,292	68.9 %	2,411 %	10.0 %
Middle East	258,356,867	3.3 %	175,502,589	67.9 %	5,243 %	3.9 %
North America	366,496,802	4.7 %	327,568,628	89.4 %	203 %	7.2 %
Oceania / Australia	41,839,201	0.5 %	28,636,278	68.4 %	276 %	0.6 %
WORLD TOTAL	7,716,223,209	100.0 %	4,536,248,808	58.8 %	1,157 %	100.0 %

The changing web

- The Web is in practice unbounded
  - Dynamic pages are unbounded
  - Static pages over 20 billion?

- Each year 20% of sites die, but …
- … 110% increase so you don't see it

Not Managed or Curated



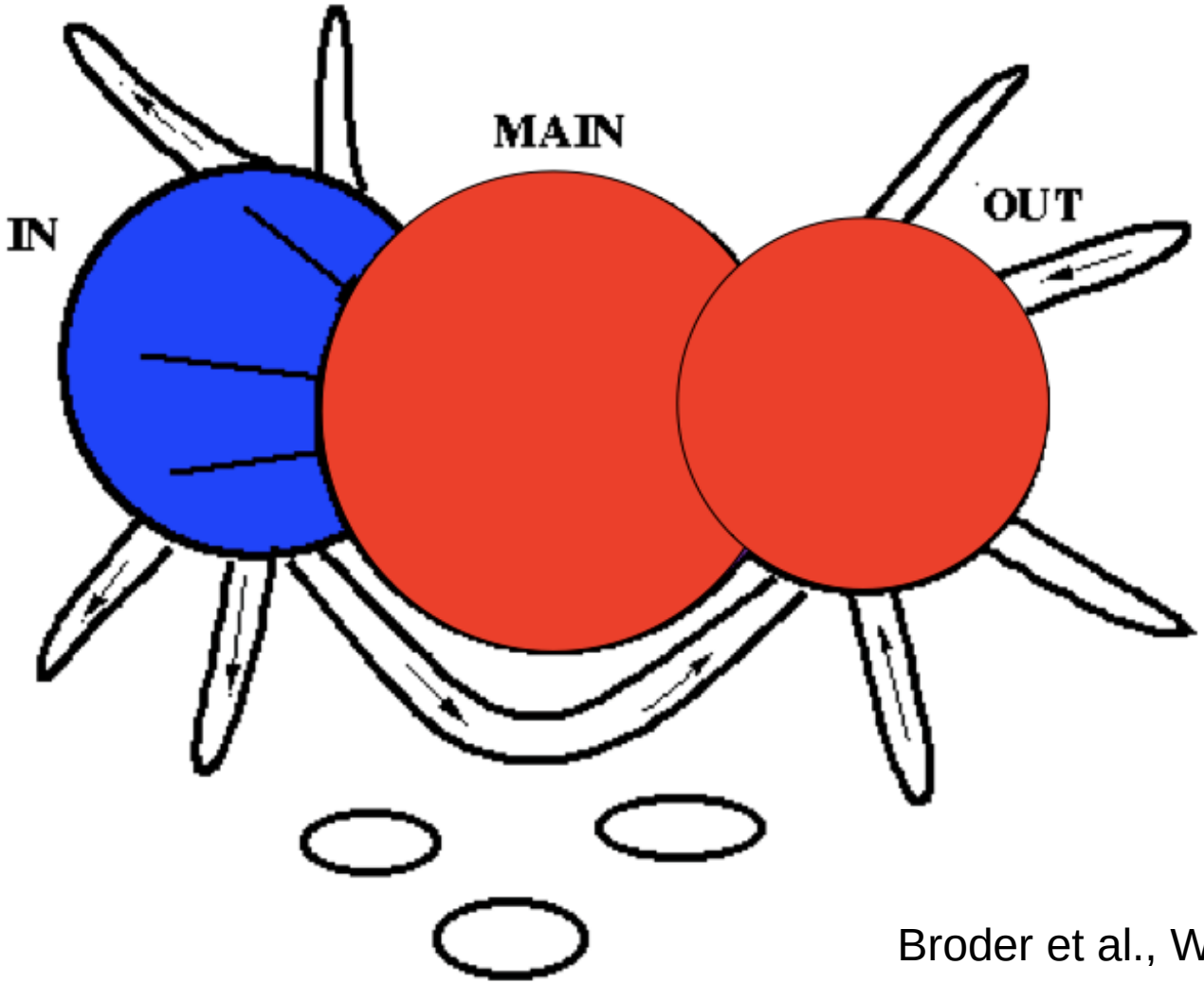
Trends in Content and Metadata (2007)

Content type	Amount of content produced per day
Published content	3-4 GB
Professional web content	~ 2 GB
User generated content	8-10 GB
Private text content	~ 3 TB
Upper bound on typed content	~700 TB

Metadata type	Amount of metadata produced per day
Anchortext	100 MB
Tags	40 MB
Pageviews	180 GB
Reviews	Around 10 MB

Raghu Ramakrishnan and Andrew Tomkins, Toward a PeopleWeb, IEEE Computer, 2007

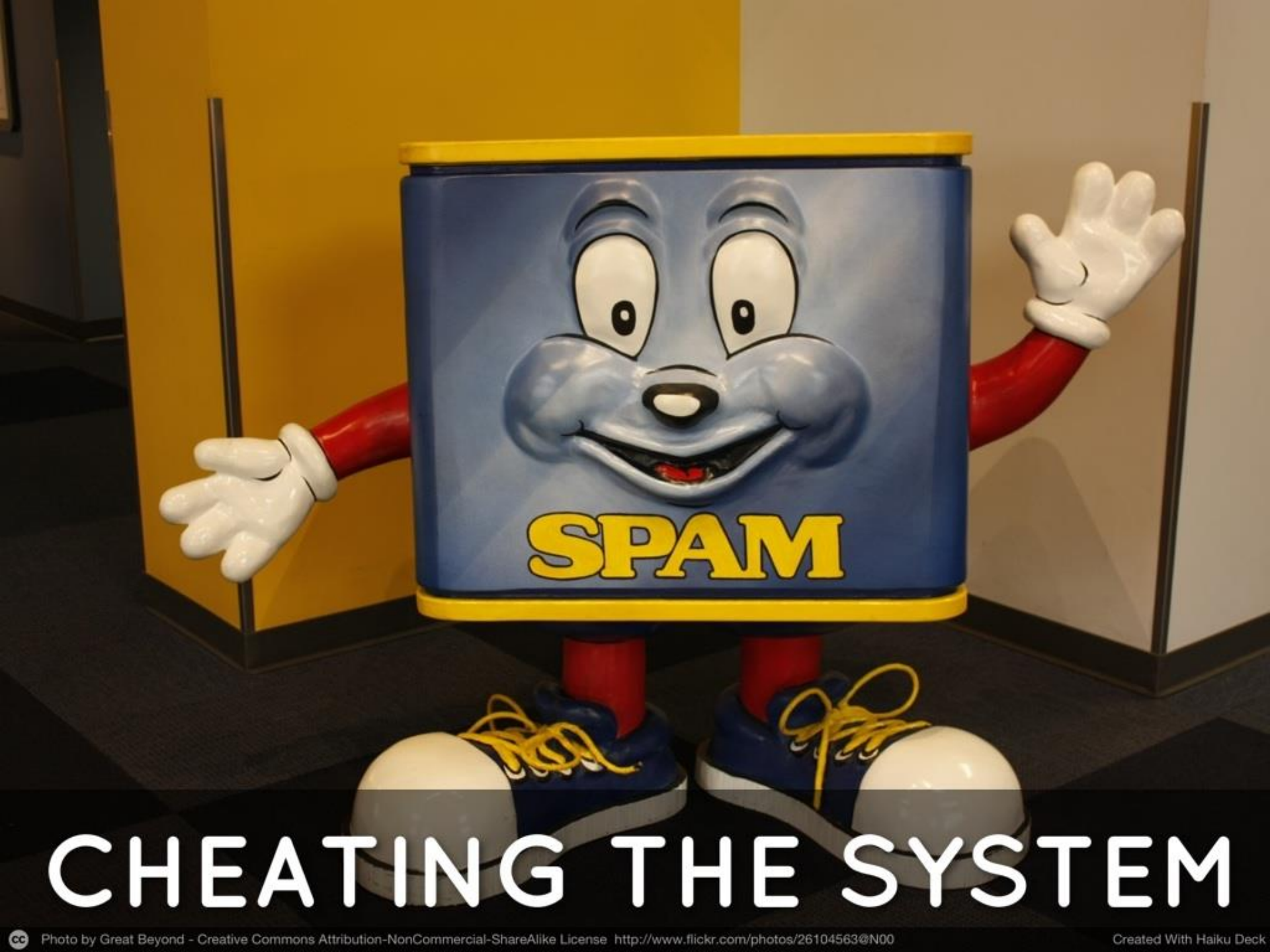
Structure of the Web: Bow Tie Model



Broder et al., WWW2000

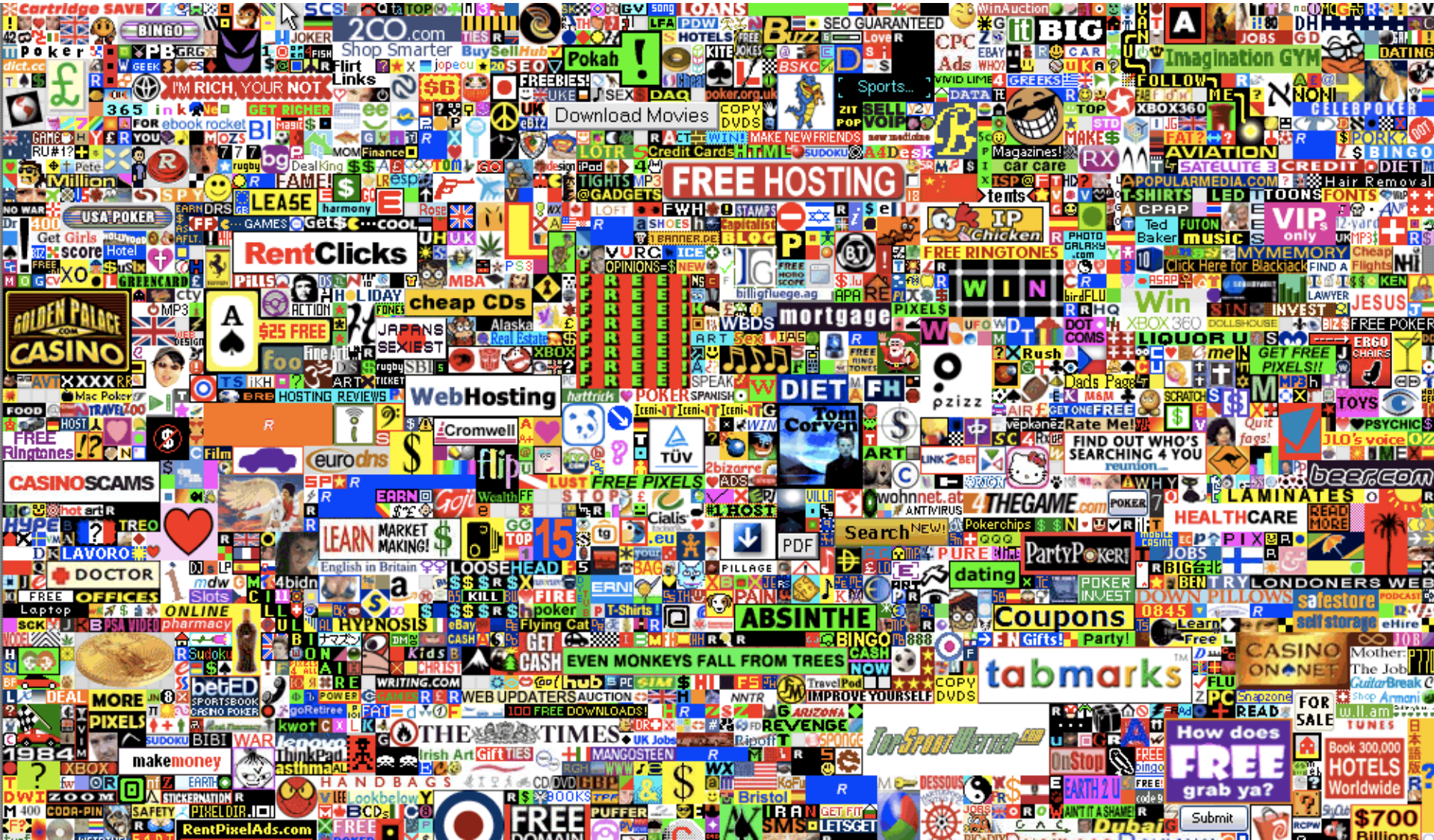
Why do people search the web?

- Very different to standard information retrieval tasks
- Broder's (2002) taxonomy of needs
  - Informational (want to learn something)
  - Navigational (want to go to that page)
  - Transactional (want to do something)
    - Access a service (e.g. Regensburg weather)
    - Download something
    - Buy something (e.g. Digital Camera)
- Grey areas - find a good source e.g. car rental Paris
- Exploratory search (what's out there?)



CHEATING THE SYSTEM

1 Million Dollar Website



Adversarial IR

- Adversarial IR is related to strategies for working with a data source where some portion of it has been manipulated maliciously
- It aims to combat Spamdexing, which is the deliberate manipulation of search engine indexes
- Search Engine Optimisation (SEO) is getting your web page to rank highly in a web search engine result list
  - SEO can be done in both a “good” and a “bad” way
- Constant battle between Search Engines and web page providers

Some Spamdexing/SEO Techniques

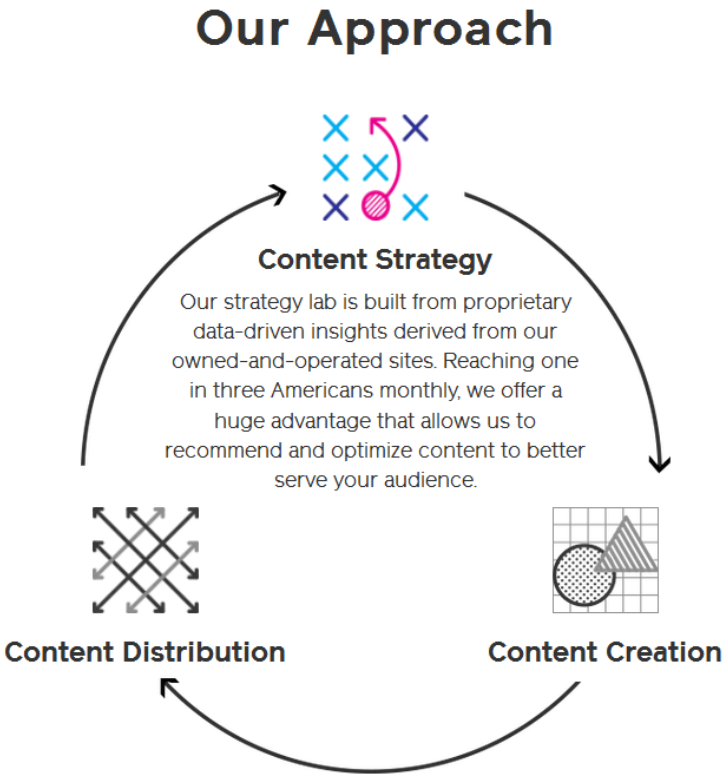
- Hidden or invisible text (very early approach):
  - Early web search engines relied heavily on tf.idf to rank results
  - Repeating words gave a pages a higher ranking (e.g. Repeating “maui resort” a few 100 times in white on a white background)
  - This usually no longer works!
- Link spam – take advantage of search engines doing link analysis
  - Link farms – communities of pages referencing each other
  - Spam blogs – link to websites from low-quality blogs
- Cloaking – serve up a different page to web crawlers than to a human user
  - Can also be used in a positive way (e.g. provide humans with content that web crawlers can’t parse)

Content generation

- Basic plan:
  - Analyse trending keywords and topics
  - Place articles or videos on these topics online (with lots of ads) so that they are found by search engines
- How are the articles generated?
  - Pay people to write articles or create videos from scratch
  - Article spinning
    - Pay people to rewrite existing articles on the trending topics
    - Use automated techniques to replace words in existing articles with synonyms

Example: DemandMedia

http://www.demandmedia.com/



No longer active

http://contentsolutions.demandstudios.com

Select a Case for Your Cell

By Jeff Grundy, eHow Contributor

Cell phones are so essential to some people that a lost or damaged mobile phone could mean loss of income or the freedom to travel. Protecting your cell phone from drops, scratches and other damage is therefore a prudent choice and may save you money in more ways than one. However, just as there are thousands of cell phone brands and models to choose from, an equally daunting selection of cell phone cases are available. Knowing what to look for in a cell phone case, and taking the environment in which you use your phone into consideration, can help you make the best selection for protecting your mobile phone investment. Have a question? Get an answer from a Nerd now!

Other People Are Reading


How to Make a Phone Cover


How to Fix a Cell Phone Dropped in the Toilet

Related Ads

Cell Phone Repeater

Phone Case

Touch and Type Cell Phone

Cellular Phone

Waterproof Plastic Case

A Summary of the Differences: Web IR vs. IR

- MUCH larger
- Nobody curating (issues of quality)
- Structured (links, anchor text, layered)
- Dynamic and constantly changing
- People are trying to cheat
- Lots of people contributing
  - We can take advantage of this (long tail)

Outline

① The Web: What makes it Unique?

② Crawling and Indexing the Web

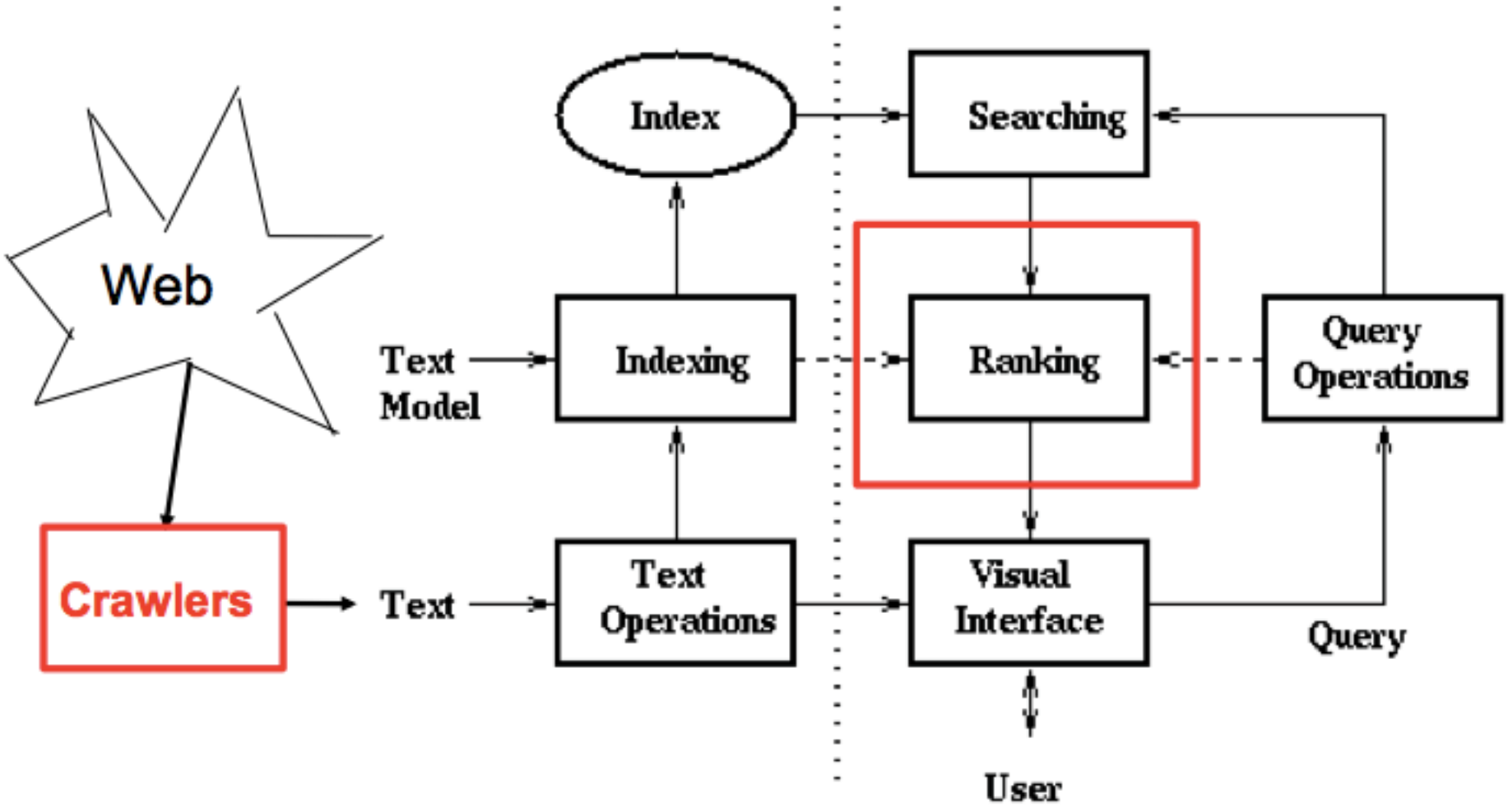
③ Ranking on the Web

④ Query Log Analysis

The Web as an Engineering Challenge

- **This is one of the most complex data engineering challenges today:**
  - Distributed in nature
  - Large volume of data
  - Highly concurrent service
  - Users expect very good & fast answers
- **Solution: Replicated centralized system**

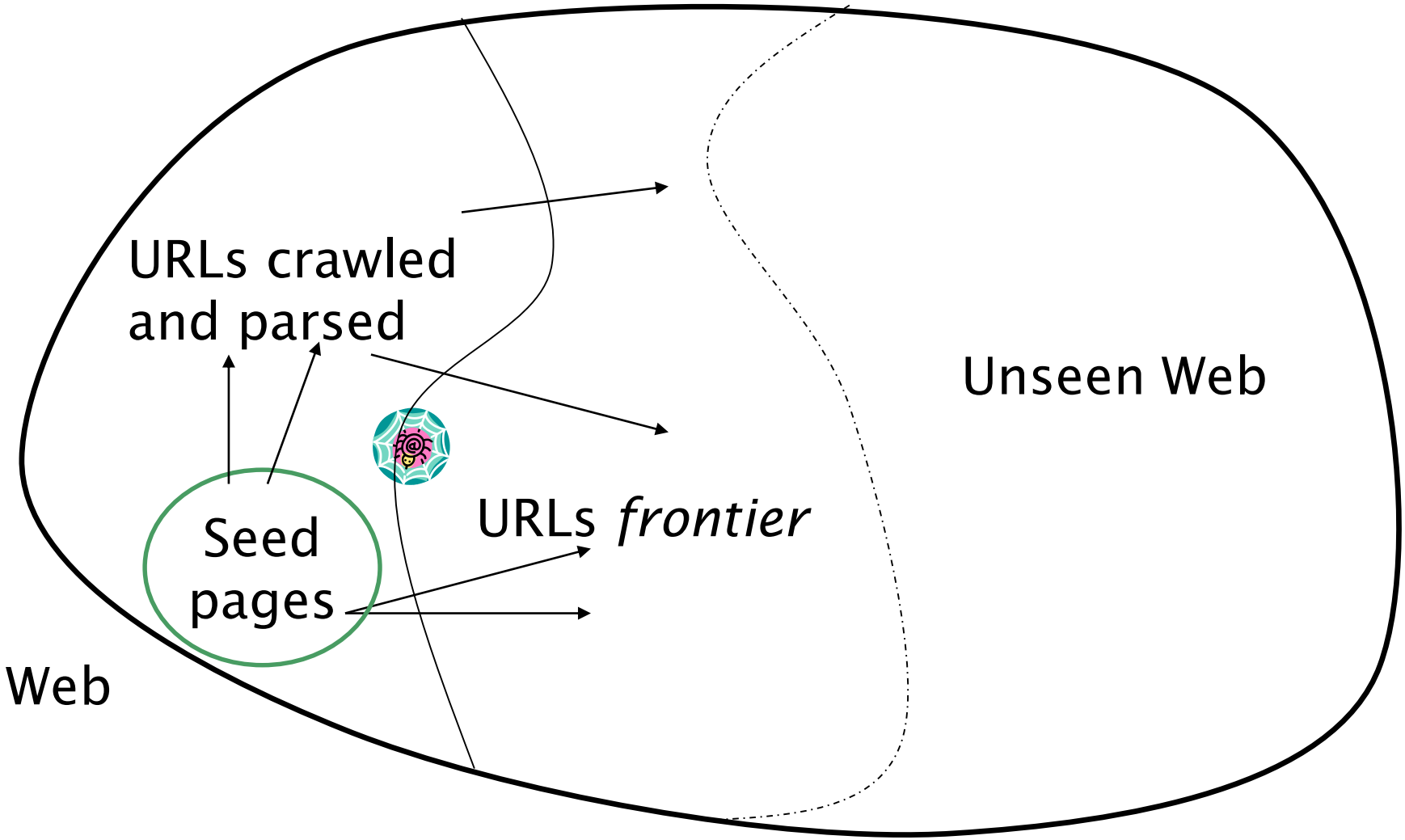
Overview


Web
Crawlers
Index
Searching
Text
Model
Indexing
Ranking
Query
Operations
Text
Text
Operations
Visual
Interface
Query
User

Basic crawler operation

- Begin with known “seed” URLs
- Fetch and parse them
  - Extract URLs they point to
  - Place the extracted URLs on a queue (the Frontier)
- Fetch each URL on the queue and repeat

Crawling picture



Simple picture – complications

- Web crawling isn't feasible with one machine
  - All of the above steps distributed
- Malicious pages
  - Spam pages
  - Spider traps – including dynamically generated
- Even non-malicious pages pose challenges
  - Latency/bandwidth to remote servers vary
  - Webmasters' stipulations
    - How "deep" should you crawl a site's URL hierarchy?
  - Site mirrors and duplicate pages
- Politeness – don't hit a server too often

What any crawler *must* do

- Be Polite: Respect implicit and explicit politeness considerations:
  - Explicit politeness: specifications from webmasters on what portions of site can be crawled
    - robots.txt
  - Implicit politeness: even with no specification, avoid hitting any site too often
- Be Robust: Be immune to spider traps and other malicious behavior from web servers

Robots.txt

- Protocol for giving spiders ("robots") limited access to a website, originally from 1994
  - http://www.robotstxt.org
- Website announces its request on what can(not) be crawled
  - For a server, create a file /robots.txt
  - This file specifies access restrictions

Robots.txt example

- No robot should visit any URL starting with "/yoursite/temp/", except the robot called "searchengine":

```
User-agent: *
Disallow: /yoursite/temp/

User-agent: searchengine
Disallow:
```

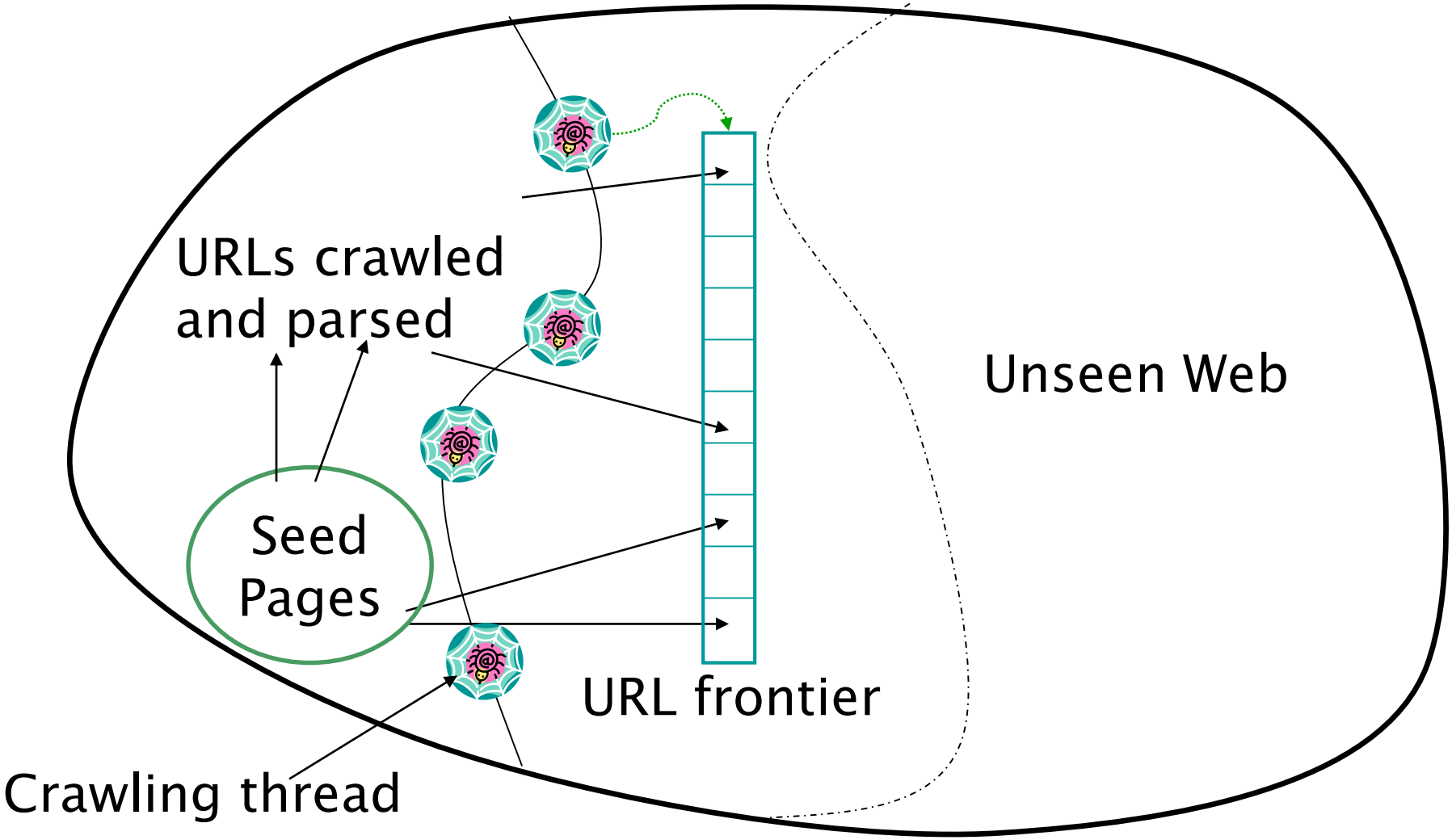
What any crawler *should* do

- Be capable of distributed operation: designed to run on multiple distributed machines
- Be scalable: designed to increase the crawl rate by adding more machines
- Performance/efficiency: permit full use of available processing and network resources

What any crawler *should* do

- Fetch pages of "higher quality" first
- Continuous operation: Continue fetching fresh copies of a previously fetched page
- Extensible: Adapt to new data formats, protocols

Updated crawling picture



URL frontier

- Contains URLs yet to be fetched in the current crawl (or re-fetched for continuous crawling)
- Can include multiple pages from the same host
- Must avoid trying to fetch them all at the same time
- Must try to keep all crawling threads busy

URL frontier: two main considerations

- Politeness: do not hit a web server too frequently
- Freshness: crawl some pages more often than others
  - E.g., pages (such as News sites) whose content changes often

Conflict Goals

(E.g., simple priority queue fails – many links out of a page go to its own site, creating a burst of accesses to that site.)

Processing steps in crawling

- Pick a URL from the frontier
- Fetch the document at the URL
- Parse the document
  - Extract links from it to other docs (URLs)
- Check if URL has content already seen
  - If not, add to indexes
- For each extracted URL
  - Ensure it passes certain URL filter tests
  - Check if it is already in the frontier (duplicate URL elimination)

Which one?

E.g., only crawl .edu, obey robots.txt, etc.

Implementation

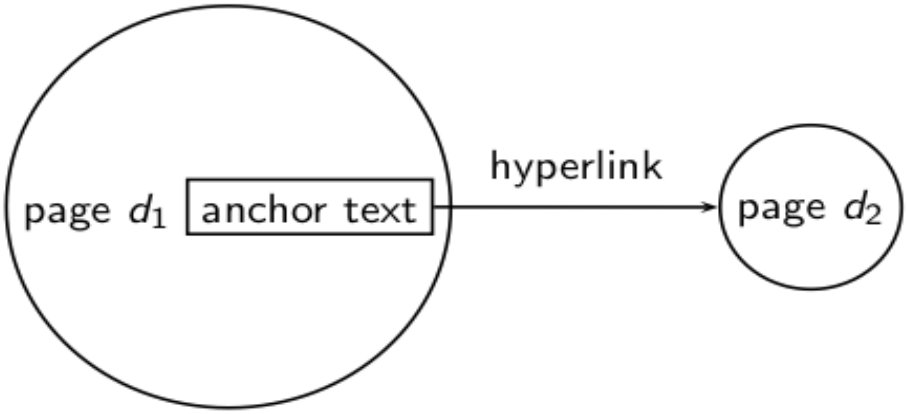
- Run multiple crawl threads, under different processes – potentially at different nodes
  - Geographically distributed nodes
- The Mercator Crawler approach is described in Manning, Chapter 20.2
- Apache Nutch is a widely-used open source web crawler

- But there are many more available: http://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining/

Outline

1. The Web: What makes it Unique?
2. Indexing the Web
3. Ranking on the Web
4. Query Log Analysis

The web as a directed graph

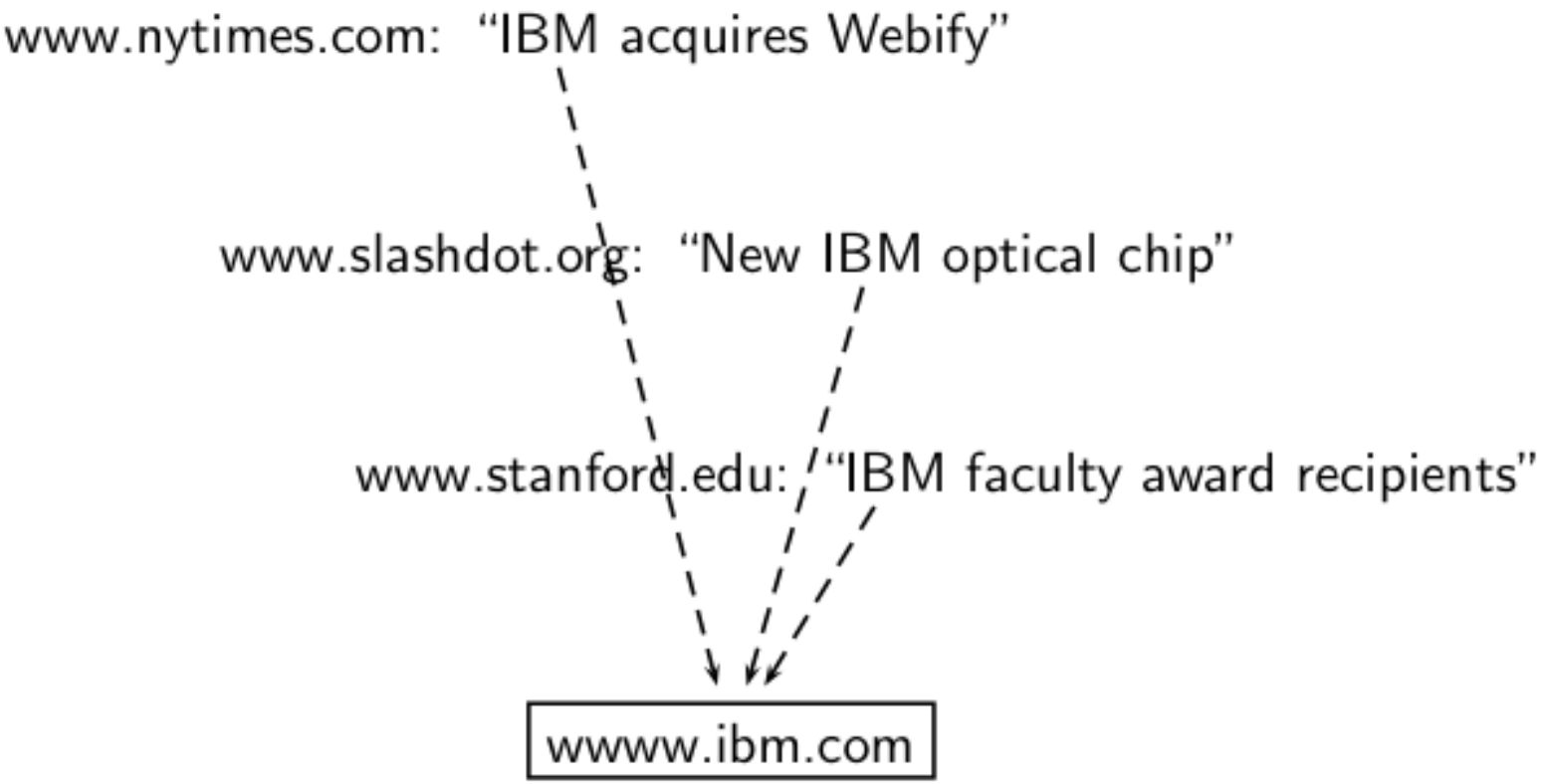


- Assumption 1: A hyperlink is a quality signal.
  - The hyperlink $d_1 \rightarrow d_2$ indicates that d_1's author deems d_2 high-quality and relevant.
- Assumption 2: The anchor text describes the content of d_2.
  - We use anchor text somewhat loosely here for: the text surrounding the hyperlink
  - Example: "You can find cheap cars <a href =http://…>here . "
  - Anchor text: "You can find cheap cars here"

[text of d_2] only vs. [text of d_2] + [anchor text → d_2]

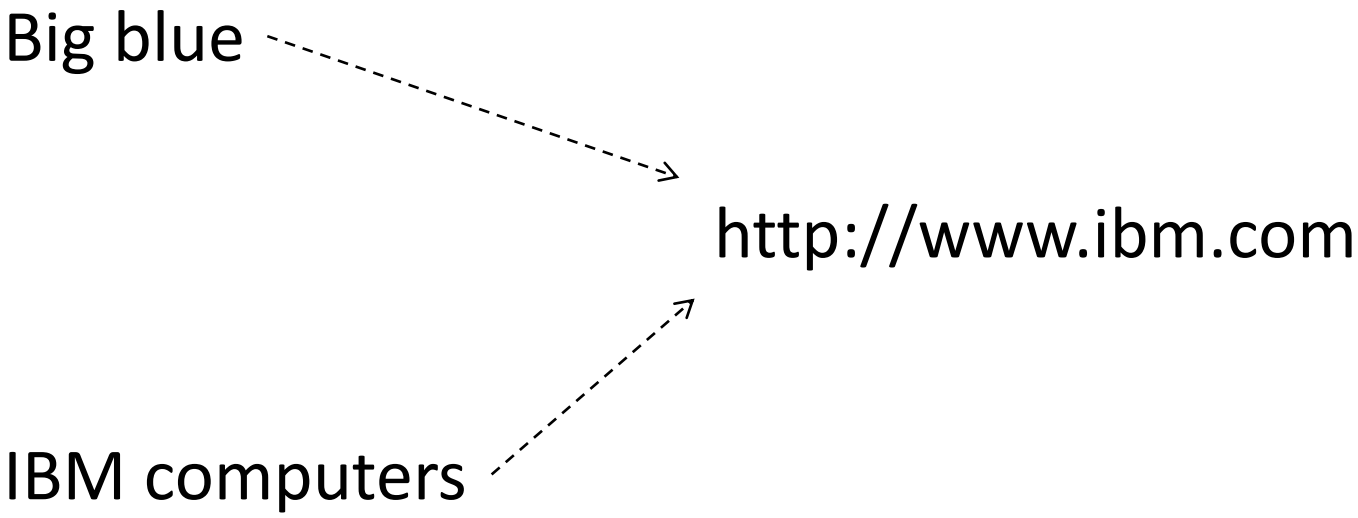
- Searching on [text of d_2] + [anchor text → d_2] is often more effective than searching on [text of d_2] only.
- Example: Query *IBM*
  - Matches IBM's copyright page
  - Matches many spam pages
  - Matches IBM wikipedia article
  - May not match IBM home page!
  - … if IBM home page is mostly graphics
- Searching on [anchor text → d_2] is better for the query *IBM*.
  - In this representation, the page with most occurrences of *IBM* is www.ibm.com

Anchor text "*IBM*" pointing to ibm.com



Other useful anchor text

Very common nickname for IBM



Indexing anchor text

- Thus: Anchor text is often a better description of a page's content than the page itself.
- Anchor text can be weighted more highly than document text.
- (based on Assumption 1&2)

Exercise: Testing our Assumptions

- Assumption 1: A link on the web is a quality signal – the author of the link thinks that the linked-to page is high-quality.
- Assumption 2: The anchor text describes the content of the linked-to page.
- Is assumption 1 true in general?
- Is assumption 2 true in general?

Origins of PageRank: Citation analysis

- Citation frequency can be used to measure the impact of an article
  - Simplest measure: Each article gets one vote
  - On the web: citation frequency = inlink count
  - A high inlink count does not necessarily mean high quality → Link spam!
- Better measure: weighted citation frequency or citation rank
  - An article's vote is weighted according to its citation impact

REFERENCES

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proc. of HLT-NAACL*.

[2] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *JAIR* (2014).

[3] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *American society for inf. science* (1990).

[4] Mostafa Dehghani, Hamed Zamani, A. Severyn, J. Kamps, and W Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proc. of SIGIR*.

[5] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query expansion with locally-trained word embeddings. *Proc. of ACL* (2016).

[3] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *American society for inf. science* (1990).

[HTML] Indexing by latent semantic analysis

S **Deerwester**, ST **Dumais**, GW Furnas... - Journal of the ..., 1990 - search.proquest.com

Indexing by Latent Semantic Analysis Scott Deerwester Center for Information and Language Studies, University of Chicago, Chicago, IL 60637 Susan T. Dumais*, George W. Furnas, and Thomas K. Landauer Bell Communications Research, 445 South St.,

☆ 99 Cited by 12803 Related articles All 85 versions Web of Science: 3906

About 12,803 results (**0.15** sec)

Indexing by latent semantic analysis

☐ Search within citing articles

Latent dirichlet allocation

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying

☆ 99 Cited by 20731 Related articles All 127 versions Web of Science: 6257

[BOOK] Data Mining: Practical machine learning tools and techniques

IH Witten, E Frank, MA Hall, CJ Pal - 2016 - books.google.com

Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition, offers a thorough grounding in machine learning concepts, along with practical advice on applying these tools and techniques in real-world data mining situations. This highly anticipated

☆ 99 Cited by 32051 Related articles All 31 versions »

[BOOK] Usability engineering

J Nielsen - 1994 - books.google.com

Written by the author of the best-selling HyperText & HyperMedia, this book is an excellent guide to the methods of usability engineering. The book provides the tools needed to avoid usability surprises and improve product quality. Step-by-step information on which method to

☆ 99 Cited by 17922 Related articles All 12 versions »

Introduction

DAC Manning - Introduction to Industrial Minerals, 1995 - Springer

Abstract Human exploitation of minerals extends back for many thousands of years and, contrary to popular belief, mining may in fact be the 'oldest profession'. Early people used minerals, initially for pigments, and stone tools for grinding and cutting. We still use some of

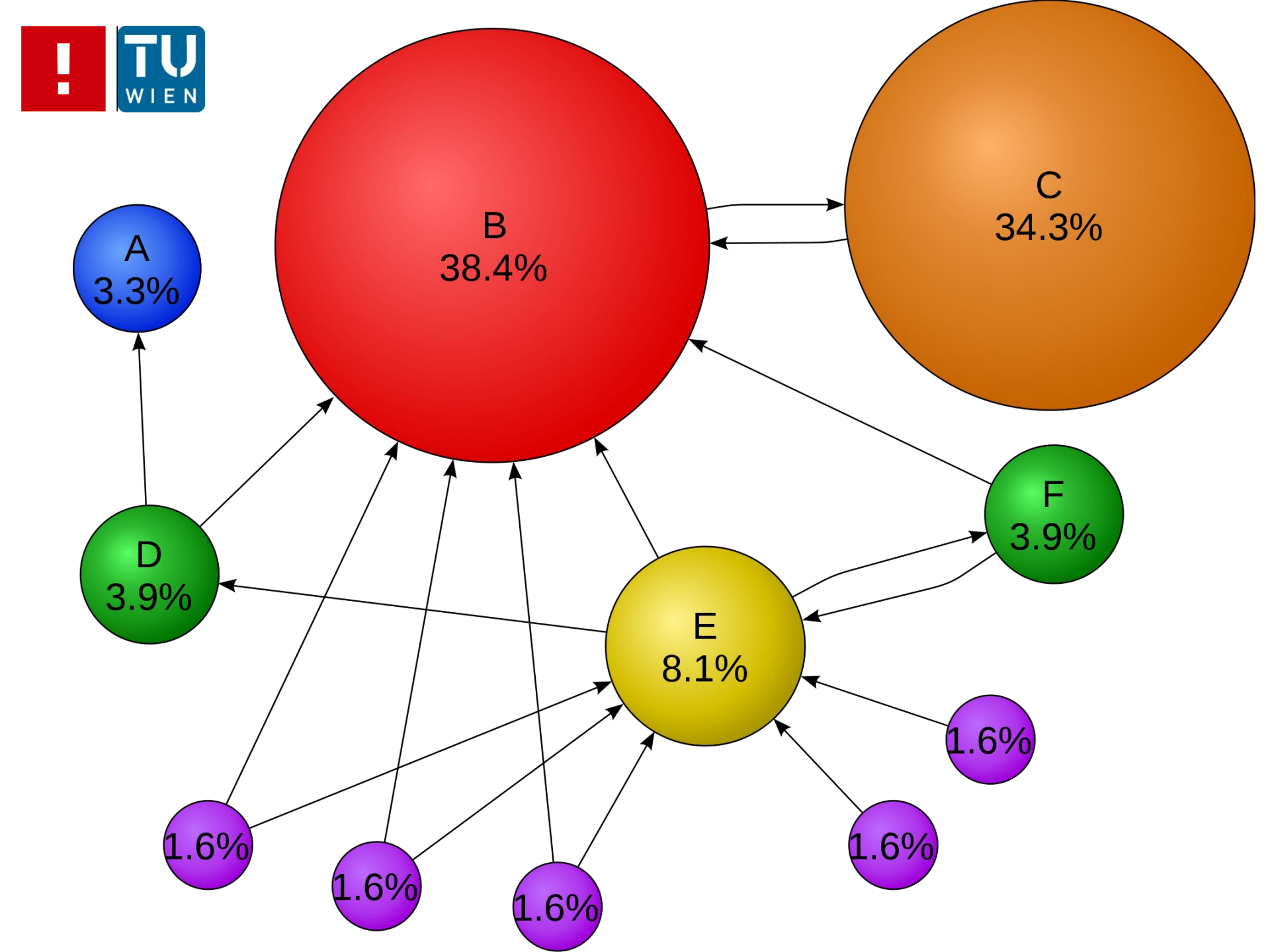
☆ 99 Cited by 13322 Related articles All 16 versions »

Origins of PageRank: Citation analysis

- Better measure: weighted citation frequency or citation rank.
- This is the idea behind PageRank.
- PageRank was invented in the context of citation analysis by Pinsker and Narin in the 1960s.
- Citation analysis is a big deal: The hiring decisions and funding of university staff are often highly influenced by the impact of their publications!

Origins of PageRank: Summary

- We can use the same formal representation for
  - Citing scientific literature
  - Citing web links
- Appropriately weighted citation frequency is a measure of quality ...
  - ... both for web pages and for scientific publications.
- Next: PageRank algorithm for computing weighted citation frequency on the web

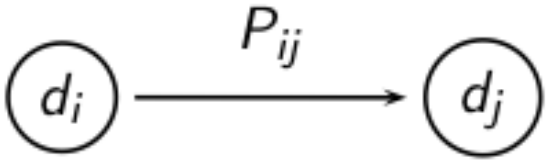

A
3.3%
B
38.4%
C
34.3%
D
3.9%
E
8.1%
F
3.9%
1.6%
1.6%
1.6%
1.6%
1.6%

PageRank Model: Random walk

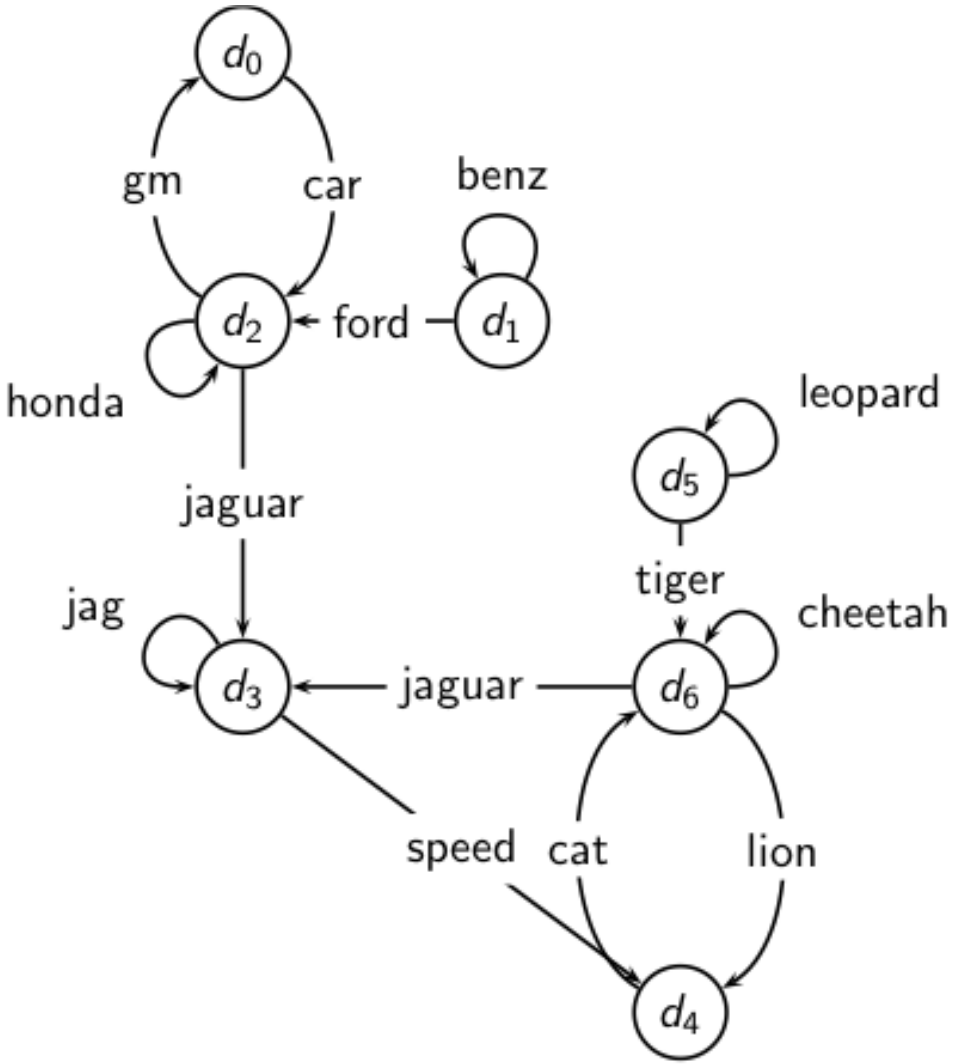
- Imagine a web surfer doing a random walk on the web
  - Start at a random page
  - At each step, go out of the current page along one of the links on that page, with equal probability
- In the steady state, each page has a long-term visit rate.
- This long-term visit rate is the page's PageRank.
- PageRank = long-term visit rate = steady state probability

Formalizing random walk: Markov chains

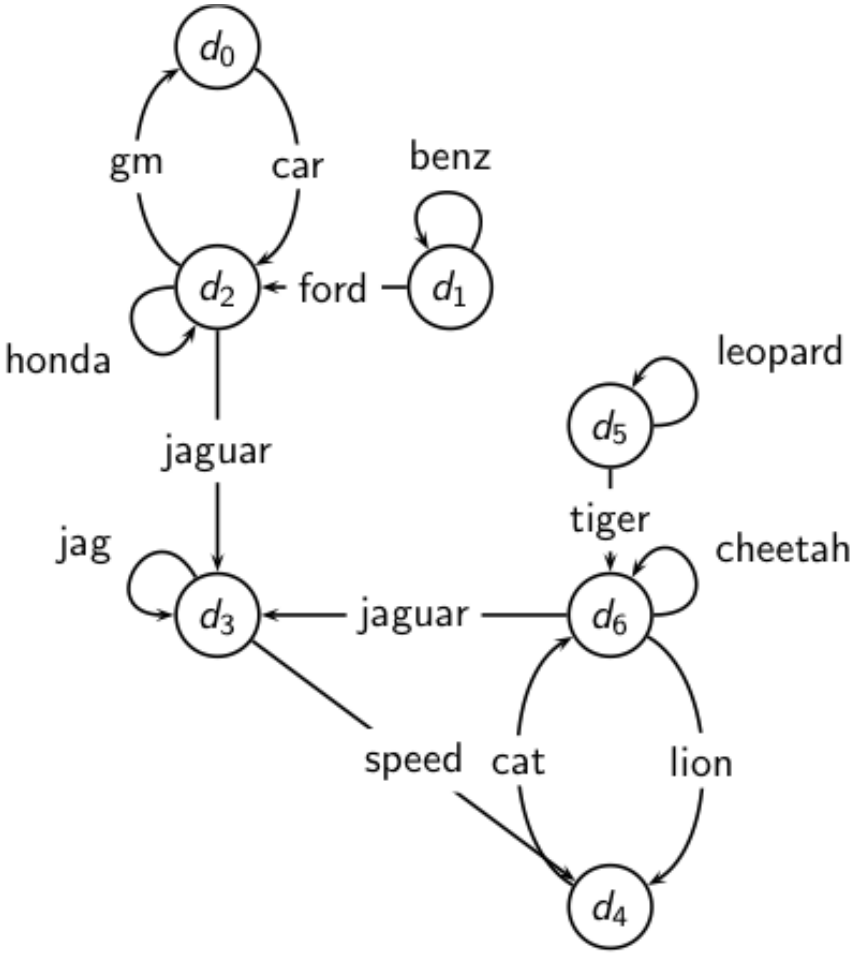
- A Markov chain consists of N states, plus an $N \times N$ transition probability matrix P
- state = page
- At each step, we are on exactly one of the pages.
- For $1 \leq i, j \leq N$, the matrix entry P_{ij} tells us the probability of j being the next page, given we are currently on page i
- Clearly, for all i, $\sum_{j=1}^{N} P_{ij} = 1$ = Row sum is 1



Example web graph

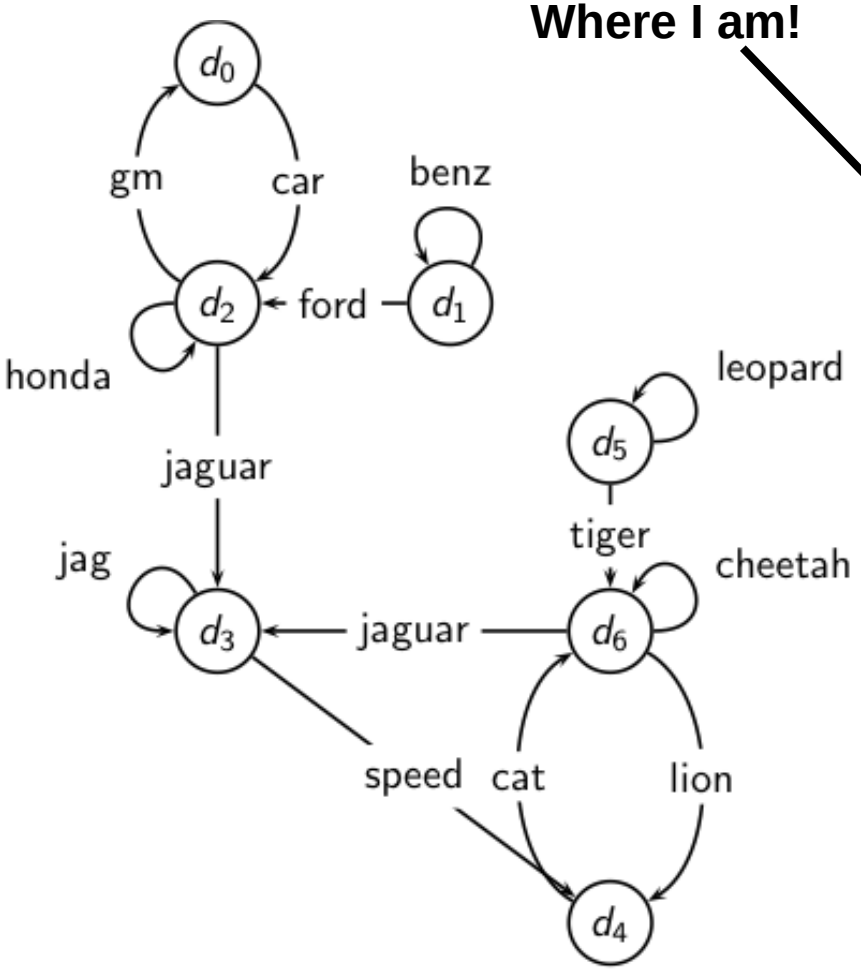


Example: Link (Adjacency) Matrix



	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Example: Link (Adjacency) Matrix



Where I am!

Where I am going!

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

Example: Transition probability matrix *P*

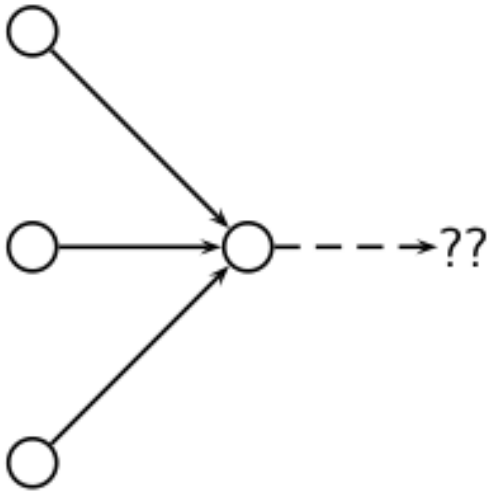
Use: $\sum_{j=1}^{N} P_{ij} = 1$

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

Long-term visit rate

- PageRank = long-term visit rate.
- Long-term visit rate of page *d* is the probability that a web surfer is at page *d* at a given point in time.
- Next: what properties must hold of the web graph for the long-term visit rate to be well defined?
- First a special case: The web graph must not contain dead ends

Dead ends



- The web is full of dead ends
- Random walk can get stuck in dead ends
- If there are dead ends, long-term visit rates are not well-defined

Teleporting – Escaping dead ends

- At a dead end, jump to random web page with prob. 1/ *N*
- At a non-dead end, with probability 10%, jump to a random web page (to each with a probability of 0.1/*N*).
- With remaining probability (90%), go out on a random hyperlink.
  - For example, if the page has 4 outgoing links: randomly choose one with probability (1-0.10)/4=0.225
- 10% is a parameter, the teleportation rate α.
- Note: "jumping" from dead end is independent of teleportation rate

Example: Transition probability matrix *P*

Add teleportation to this matrix… α=0.1

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

Adding Teleportation

- If a row has no 1's, then replace each element by $1/N$
- For all other rows: Divide each 1 by the number of 1's in its row. Thus, if there is a row with three 1's, then each of them is replaced by 1/3
- Multiply the resulting matrix by $1 - \alpha$
- Add α/N to every entry of the resulting matrix

Example: Transition probability matrix *P*

Matrix with teleportation (α=0.1) (and rounding errors)

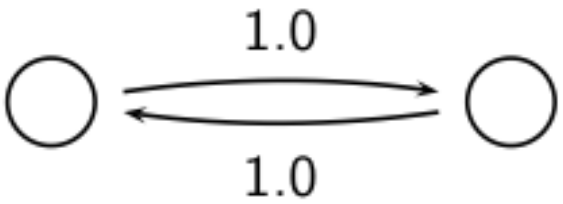
	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.017	0.017	0.917	0.017	0.017	0.017	0.017
d_1	0.017	0.467	0.467	0.017	0.017	0.017	0.017
d_2	0.317	0.017	0.317	0.317	0.017	0.017	0.017
d_3	0.017	0.017	0.017	0.467	0.467	0.017	0.017
d_4	0.017	0.017	0.017	0.017	0.017	0.017	0.917
d_5	0.017	0.017	0.017	0.017	0.017	0.467	0.467
d_6	0.017	0.017	0.017	0.317	0.317	0.017	0.317

Result of teleporting

- Avoids getting stuck in a dead end
- But even without dead ends, a graph may not have well-defined long-term visit rates
- More generally, we require that the Markov chain be ergodic

Ergodic Markov chains

- A Markov chain is ergodic if it is irreducible and aperiodic
- Irreducibility. Roughly: there is a path from any other page
- Aperiodicity. Roughly: The pages cannot be partitioned such that the random walker visits the partitions sequentially
- A non-ergodic Markov chain:



Ergodic Markov chains

- Theorem: For any ergodic Markov chain, there is a unique long-term visit rate for each state
- This is the steady-state probability distribution
- Over a long time period, we visit each state in proportion to this rate
- Does not matter where we start
- Teleporting makes the web graph ergodic

$\Longrightarrow$ Web-graph + teleporting has steady-state probability distribution

$\Longrightarrow$ Each page in the web-graph + teleporting has PageRank

Formalization of “visit”: Prob. vector

- A probability (row) vector $\vec{x} = (x_1, \ldots, x_N)$ tells us where the random walk is at any point
- Example:

(	0	0	0	…	1	…	0	0	0	)
	1	2	3	…	*i*	…	N-2	N-1	N	

- More generally: the random walk is on the page *i* with probability x_i
- Example:

(	0.05	0.01	0.0	…	0.2	…	0.01	0.05	0.03	)
	1	2	3	…	*i*	…	N-2	N-1	N	

Formalization of “visit”: Prob. vector

- A probability (row) vector $\vec{x} = (x_1, \ldots, x_N)$ tells us where the random walk is at any point
- Example:

(	0	0	0	…	1	…	0	0	0	)
	1	2	3	…	*i*	…	N-2	N-1	N	

- More generally: the random walk is on the page *i* with probability x_i
- Example:

(	0.05	0.01	0.0	…	0.2	…	0.01	0.05	0.03	)
	1	2	3	…	*i*	…	N-2	N-1	N	

- $\Sigma\, x_i = 1$

Change in probability vector

- If the probability vector is $\vec{x} = (x_1, \ldots, x_N)$, at this step, what is it at the next step?
- Recall that row i of the transition probability matrix P tells us where we go next from state i
- So from $\vec{x}$, our next state is distributed as $\vec{x}P$
- *… and then* $(\vec{x}P)P = \vec{x}P^2$
- *… and then* $((\vec{x}P)P)P) = \vec{x}P^3$
- *… and so on …* $= \vec{x}P^\infty$

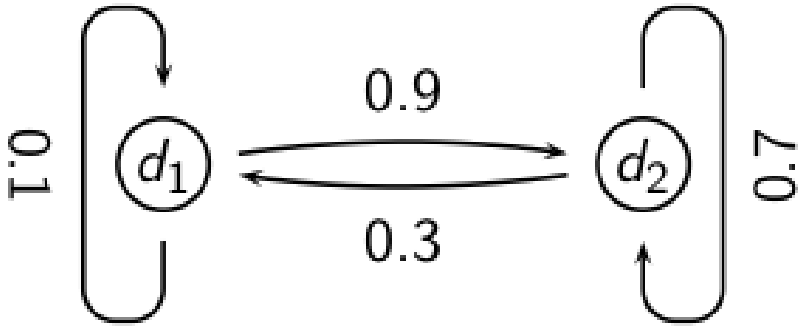
- Known as *Power Iteration*

Steady state in vector notation

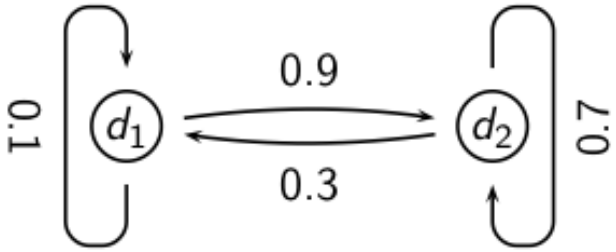
- The steady state in vector notation is simply a vector
- $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ of probabilities
- (We use $\vec{\pi}$ to distinguish it from the notation for the probability vector $\vec{x}$)
- π is the long-term visit rate (or PageRank) of page *i*
- So we can think of PageRank as a very long vector – one entry per page

Power method: Example

- What is the PageRank / steady state in this example?



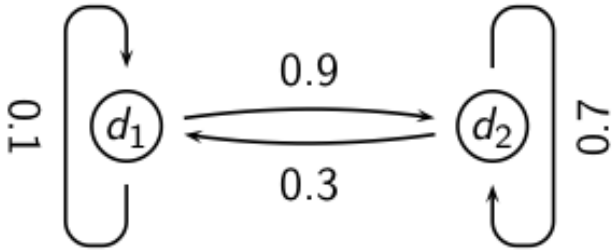
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1			$= \vec{x}P$
t_1					$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

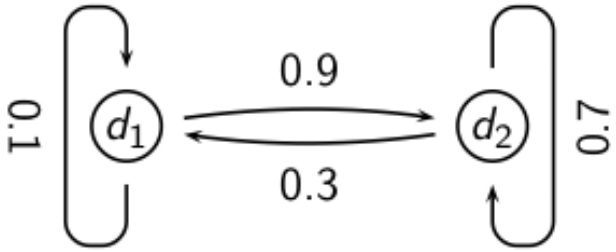
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1					$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

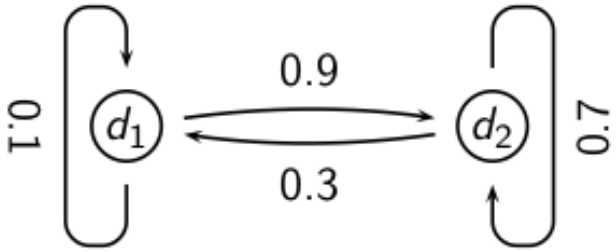
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7			$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

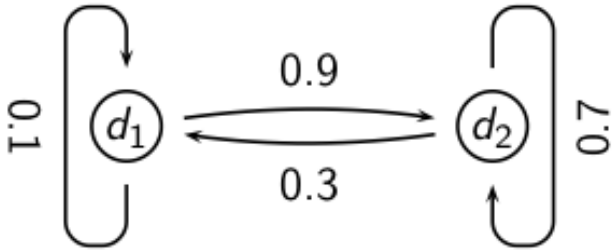
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2					$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

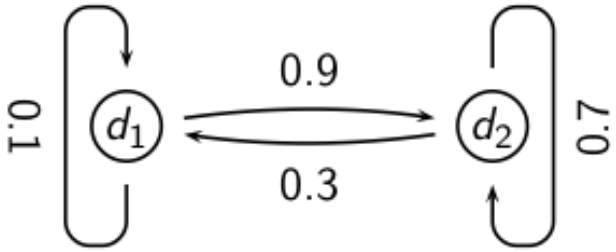
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76			$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

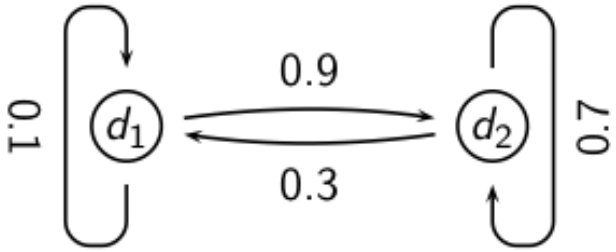
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3					$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

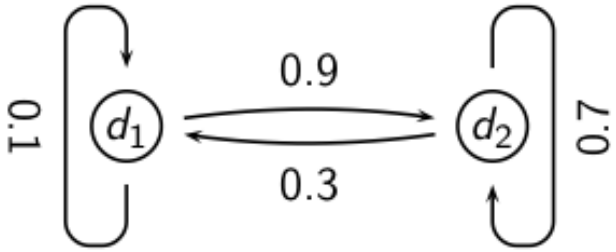
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748			$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

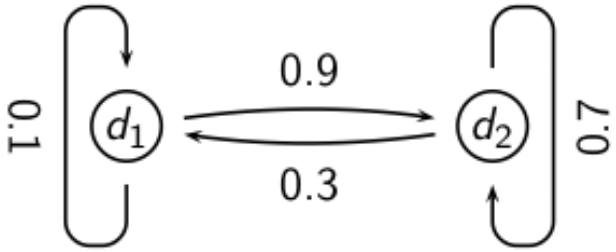
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
					. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

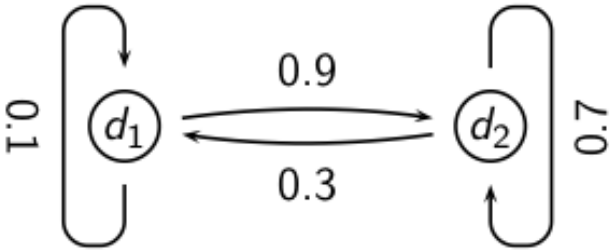
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			. . .		. . .
t_∞					$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

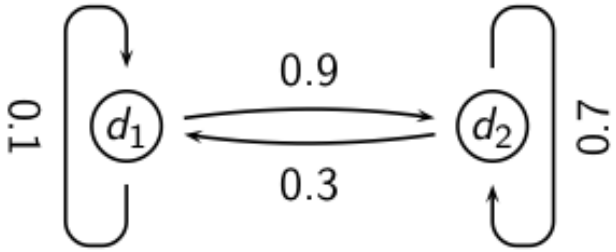
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			. . .		. . .
t_∞	0.25	0.75			$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

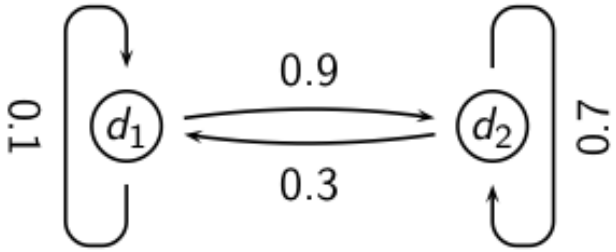
Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			. . .		. . .
t_∞	0.25	0.75	0.25	0.75	$= \vec{x}P^\infty$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$
$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

Computing PageRank: Power Example



	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
			. . .		. . .
t_∞	0.25	0.75	0.25	0.75	$= \vec{x}P^\infty$

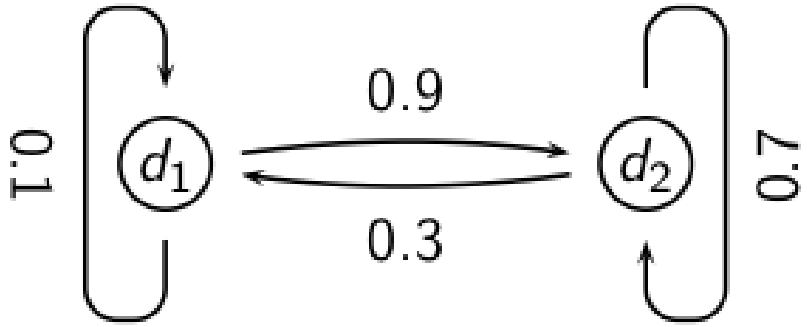
PageRank vector = $\vec{\pi}$ = (π_1, π_2) = (0.25, 0.75)

$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$

$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$

Power method: Example

- What is the PageRank / steady state in this example?



- The steady state distribution (= the PageRanks) in this example are 0.25 for d_1 and 0.75 for d_2

PageRank summary

- Preprocessing
  - Given graph of links, build matrix *P*
  - Apply teleportation
  - From modified matrix, compute π
  - $\vec{\pi}_i$ is the PageRank of page *i*.

- Query processing
  - Retrieve pages satisfying the query
  - Rank them by their PageRank
  - Return reranked list to the user

PageRank issues

- Real surfers are not random surfers
  - Examples of nonrandom surfing: back button, short vs. long paths, bookmarks, directories – and search!
  - → Markov model is not a good model of surfing
  - But it's good enough as a model for our purposes

PageRank issues

- Simple PageRank ranking produces bad results for many pages
  - Consider the query [video service]
  - The Yahoo home page (i) has a very high PageRank and (ii) contains both *video* and *service*
  - If we rank all Boolean hits according to PageRank, then the Yahoo home page would be top-ranked
  - Clearly not desirable

How important is PageRank?

- Frequent claim: PageRank is the most important component of web ranking
- The reality:
  - There are several components that are at least as important: e.g., anchor text, phrases, proximity
  - PageRank in his original form (as presented here) now has a negligible impact on ranking!
  - However, variants of a page's PageRank are still an essential part of ranking
  - Addressing link spam is difficult and crucial

Learning to Rank

- Ranking as a machine learning problem
- Supervised learning
- Training and Testing sets created
- Lots of features, page-ranking, anchor text terms, etc.
- Ranking algorithm's job is to find weights for those features that maximise performance
- Evaluated on the test set using methods similar to those we have seen already

Outline

1. The Web: What makes it Unique?
2. Indexing the Web
3. Ranking on the Web
4. Query Log Analysis

Query Log Analysis

"Query log mining[/analysis] is concerned with all those techniques aimed at discovering interesting patterns from query logs of web search engines with the purpose of enhancing either effectiveness or efficiency of an online service provided through the web"

Fabrizio Silvestri
FNTIR Vol. 4, Nos. 1-2, 2010

- Also for enhancing users' search experience, search-based advertisement and web marketing in general

Query Log Example

- AOL query log, released in 2006

AnonID	Query	QueryTime	ItemRank	ClickURL
142	rentdirect.com	2006-03-01 07:17:12		
142	www.prescriptionfortime.com	2006-03-12 12:31:06		
142	staple.com	2006-03-17 21:19:29		
142	staple.com	2006-03-17 21:19:45		
142	www.newyorklawyersite.com	2006-03-18 08:02:58		
142	www.newyorklawyersite.com	2006-03-18 08:03:09		
142	westchester.gov	2006-03-20 03:55:57	1	http://www.westchestergov.com
142	space.comhttp	2006-03-24 20:51:24		
142	dfdf	2006-03-24 22:23:07		
142	dfdf	2006-03-24 22:23:14		
142	vaniqa.comh	2006-03-25 23:27:12		
142	www.collegeucla.edu	2006-04-03 21:12:14		
142	www.elaorg	2006-04-03 21:25:20		
142	207 ad2d 530	2006-04-08 01:31:04		
142	207 ad2d 530	2006-04-08 01:31:14	1	http://www.courts.state.ny.us
142	broadway.vera.org	2006-04-08 08:38:23		
142	broadway.vera.org	2006-04-08 08:38:31		
142	vera.org	2006-04-08 08:38:42	1	http://www.vera.org
142	broadway.vera.org	2006-04-08 08:39:30		
142	frankmellace.com	2006-04-09 02:19:24		
142	ucs.ljx.com	2006-04-09 02:20:44		
142	attornyleslie.com	2006-04-13 00:25:27		
142	merit release appearance	2006-04-22 23:51:18		
142	www.bonsai.wbff.org	2006-05-06 08:49:34		
142	loislaw.com	2006-05-12 22:43:36		
142	rapny.com	2006-05-18 09:21:57		
142	whitepages.com	2006-05-19 19:36:31		
217	lottery	2006-03-01 11:58:51	1	http://www.calottery.com
217	lottery	2006-03-01 11:58:51	1	http://www.calottery.com
217	ameriprise.com	2006-03-01 14:06:23	1	http://www.ameriprise.com
217	susheme	2006-03-02 12:31:08		
217	united.com	2006-03-03 14:54:13		
217	mizuno.com	2006-03-07 22:41:17	1	http://www.mizuno.com

Twenty million search keywords for over 650,000 users over a 3-month period

AOL query log word cloud

From: Silvestri, FNTIR

AOL User
#427326

www.akidimicshoes.com		
www.yahoo.com		
www.snagajob.com	→	www.snagajob.com
women dress shoes		
www.music.yahoo.com	→	music.yahoo.com
www.chrisbrownworld.com	→	www.chrisbrownworld.com
www.tandemfcu.com	→	www.tandemfcu.com
where is my computer		
www.tandemfcu.com		
www.mich.guv		
secretary of statewww.mich.guv		
state of michigan		
state of michigan secretaryof state	→	www.michigan.gov
www.tandemfcu.com	→	www.tandemfcu.com
tandemfcu.com	→	www.tandemfcu.com
buildersquarecom		
buildersquare		
builder square	→	www.thehomeshow.com
jetted tubs		
lowes	→	www.lowes.com
builder square		
lowes		
america west		
www.americawestairlines		
american west		
aol toolbar context search.com		

AOL User
#4417749

Thelma Arnold

rescue of older dogs	→	www.srdogs.com
school supplies for the iraq children	→	www.operationiraqichildren.org
pine straw lilburn delivery		
pine straw delivery in gwinnett county		
landscapers in lilburn ga.		
pne straw in lilburn ga.		
pine straw in lilburn ga.	→	gwinnett-online.com
gwinnett county yellow pages	→	directory.respond.com
seffects of nicotine		
effects of nicotine x 4	→	www.nida.nih.gov
	→	www.bic.uci.edu
	→	*www.ash.org.uk*
safest place to live x 2	→	ezinearticles.com
	→	homebuying.about.com
new zealand	→	www.immigration.govt.nz
new zealand real estate	→	www.realenz.co.nz
retirement in new ealand		
retirement in austrailia	→	www.globalaging.org
best retirement in the world	→	www.escapeartist.com
best retirement place in usa x 3	→	www.amazon.com
	→	www.clubmarena.com
	→	www.committment.com
jarrett arnold x 2	→	www.oregoncountryfair.org
jarrett t. arnold		
jarrett t. arnold eugene oregon x 2	→	www2.eugeneweekly.com
jack t. arnold		
juice plus x 2	→	www.juiceplus.com
	→	www.mlmwatch.org

Privacy

- Unless you work for a search company, you're unlikely to get access to a good query log
- Therefore, published results on search log analysis tend to be on:
  - Old query logs (university researchers)
  - Query logs that are not available (search company researchers)

Nature of web queries

- Short (unlikely to contain more than 3 terms)
- Search operators rarely used (e.g. “”, +, -)
- Some queries more popular – the distribution of query popularity follows a power law – the most popular queries account for a very small fraction of the total number of unique queries

Likely due to "More Like This" function

query	freq.
Empty Query	2,586
sex	229
chat	58
lucky number generator	56
p****	55
porno	55
b****y	55
nude beaches	52
playboy	46
bondage	46
porn	45
rain forest restaurant	40
f****ing	40
crossdressing	39
crystal methamphetamine	36
consumer reports	35
xxx	34
nude tanya harding	33
music	33
sneaker stories	32

Excite (1997-2001)

query	freq.
christmas photos	31,554
lyrics	15,818
cracks	12,670
google	12,210
gay	10,945
harry potter	7,933
wallpapers	7,848
pornografia	6,893
"yahoo com"	6,753
juegos	6,559
lingerie	6,078
symbios logic 53c400a	5,701
letras de canciones	5,518
humor	5,400
pictures	5,293
preteen	5,137
hypnosis	4,556
cpc view registration key	4,553
sex stories	4,521
cd cover	4,267

Altavista (1998)

Fig. 2.3 The most popular queries out of the Excite and publicly available Altavista Logs. Potentially offending terms have been replaced by similar terms containing asterisks ('*'). Query have not previously filtered to remove stop-words and terms in queries have not been reordered.

From: Silvestri, FNTIR

Power Law

- $y = Kx^{-\alpha}$
  - K – constant corresponding to the query with the highest popularity
  - x – popularity rank
  - α – real parameter measuring how popularity decreases against the rank
- $\log(y) = -\alpha\log(x) + \log(K)$
  - Power law distributions have the form of a straight line when plotted on a log-log scale.

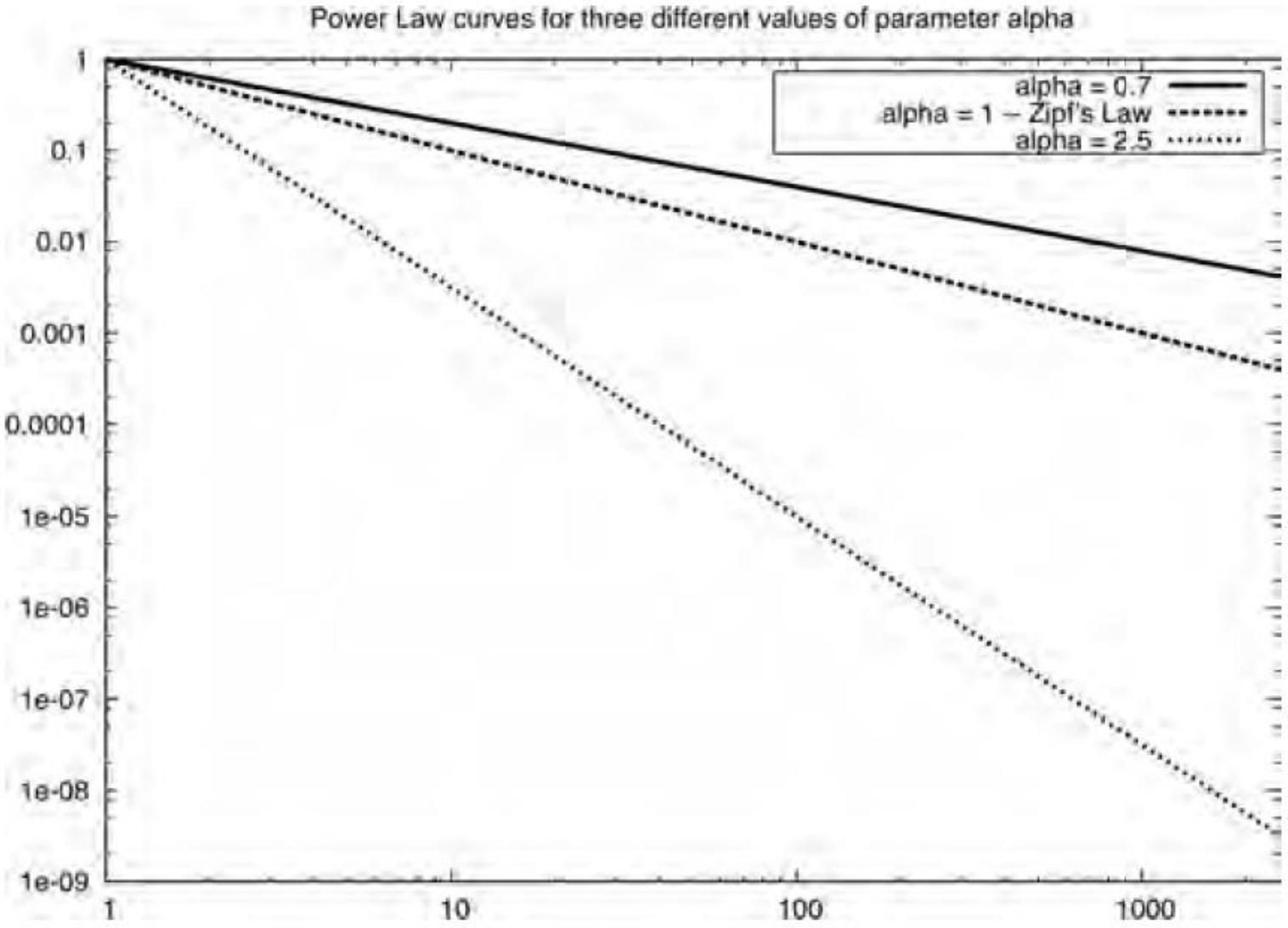
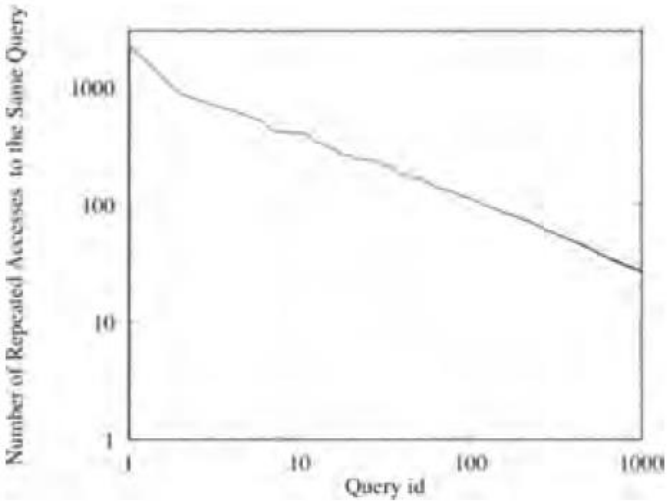
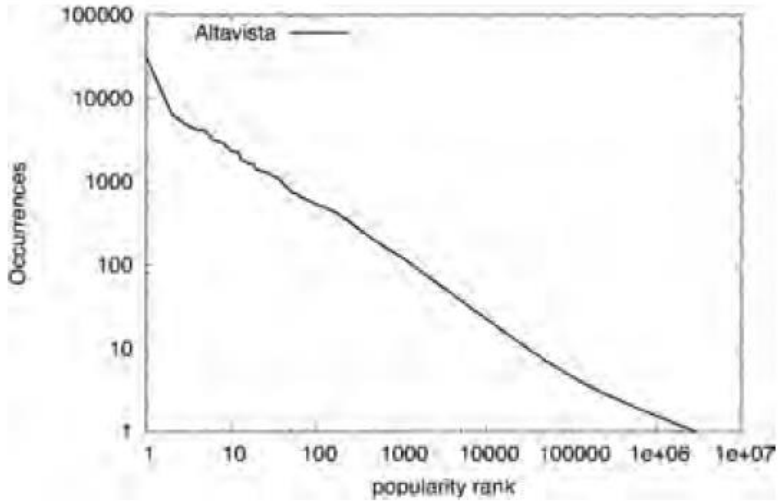


Fig. 2.1 Three examples of power-law curves for different values of the parameter α. The curve corresponding to $\alpha = 1$ is usually said to identify the Zip's law [245].

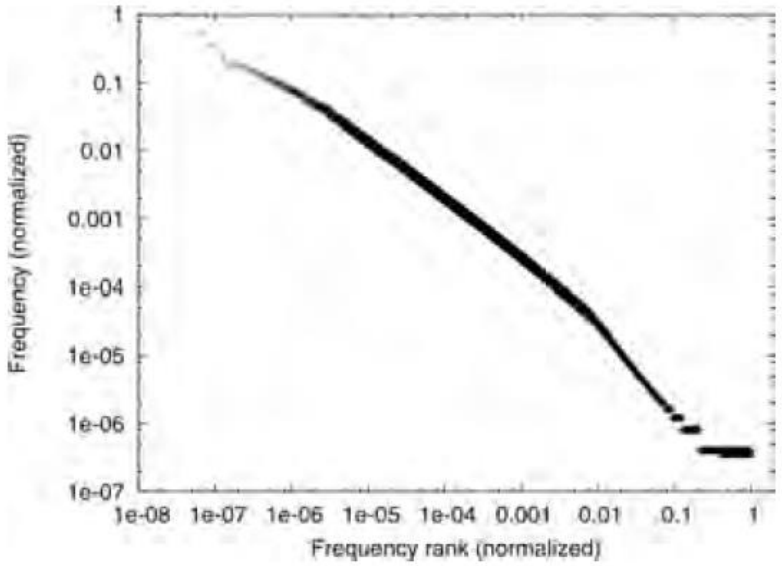
From: Silvestri, FNTIR



(a)



(b)



(c)

Fig. 2.2 Plots displaying query popularity for various query logs. (a) The 1,000 most popular queries in the Excite Log [144]; (b) Query popularity of Altavista queries [129] and (c) Query popularity of Yahoo! queries [15].

From: Silvestri, FNTIR

User Intent

- Queries can be classified as:
  - **Informational** – Users looking for information on a particular topic (e.g. *San Francisco* or *normocytic anemia*)
  - **Navigational** – Mostly users looking for the URL of a particular page (e.g. *Greyhound Bus*, *American airlines home*, *Don Knuth*)
  - **Transactional** – Users looking for websites that enable the buying of goods (e.g. *online music*, *online flower delivery service*)
- Almost even distribution across categories

Type	Surveyed	Estimated (from Query Log)
Navigational	24.5%	20%
Informational	~ 39%	48%
Transactional	~ 36%	30%

A. Z. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

- Survey of Altavista users on their intent through pop-up windows
- Manual estimation from 1000 randomly selected query log entries

Search Sessions

- A series of queries that are part of a single, information seeking activity
- Show how users interact with the search engine and how they modify queries depending on the results obtained

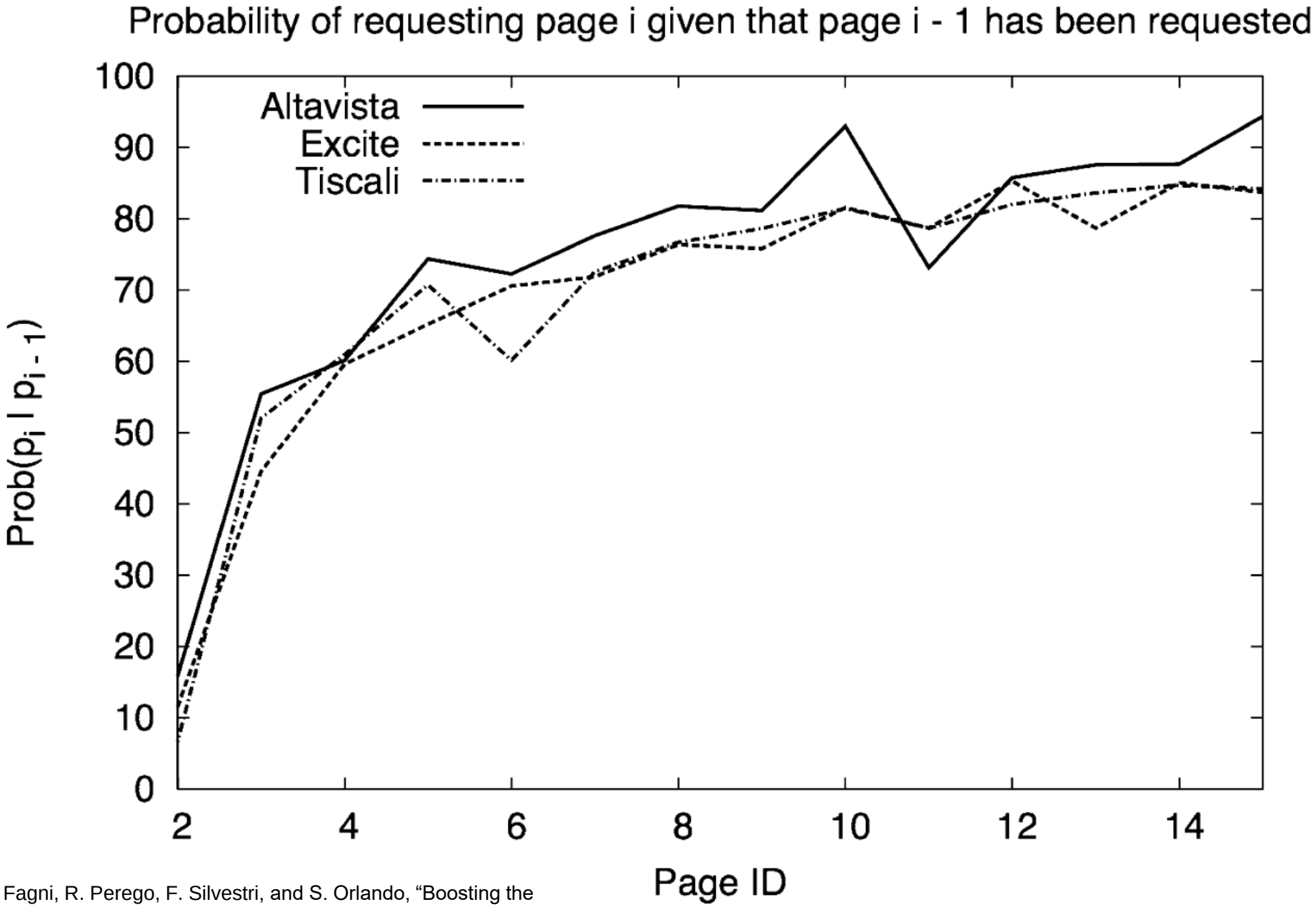
Observations

- Users, in the vast majority of cases, look at the first page of results only. If they do go to the second page, the likelihood of looking at further pages is high.

Table II. Percentage of Queries in the Logs as a Function of the Index of the Page Requested

Query Log	1	2	3	4	5	6	7	8	9	10
Excite	77.59	8.92	3.98	2.37	1.54	1.09	0.78	0.60	0.45	0.37
Tiscali	83.20	5.50	2.86	1.74	1.23	0.74	0.54	0.41	0.32	0.26
Alta Vista	64.76	10.25	5.68	3.41	2.54	1.83	1.42	1.16	0.94	0.88

T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data," *ACM Transactions on Information Systems*, vol. 24, no. 1, pp. 51–78, 2006.



T. Fagni, R. Perego, F. Silvestri, and S. Orlando, "Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data," *ACM Transactions on Information Systems*, vol. 24, no. 1, pp. 51–78, 2006.

Observations (2)

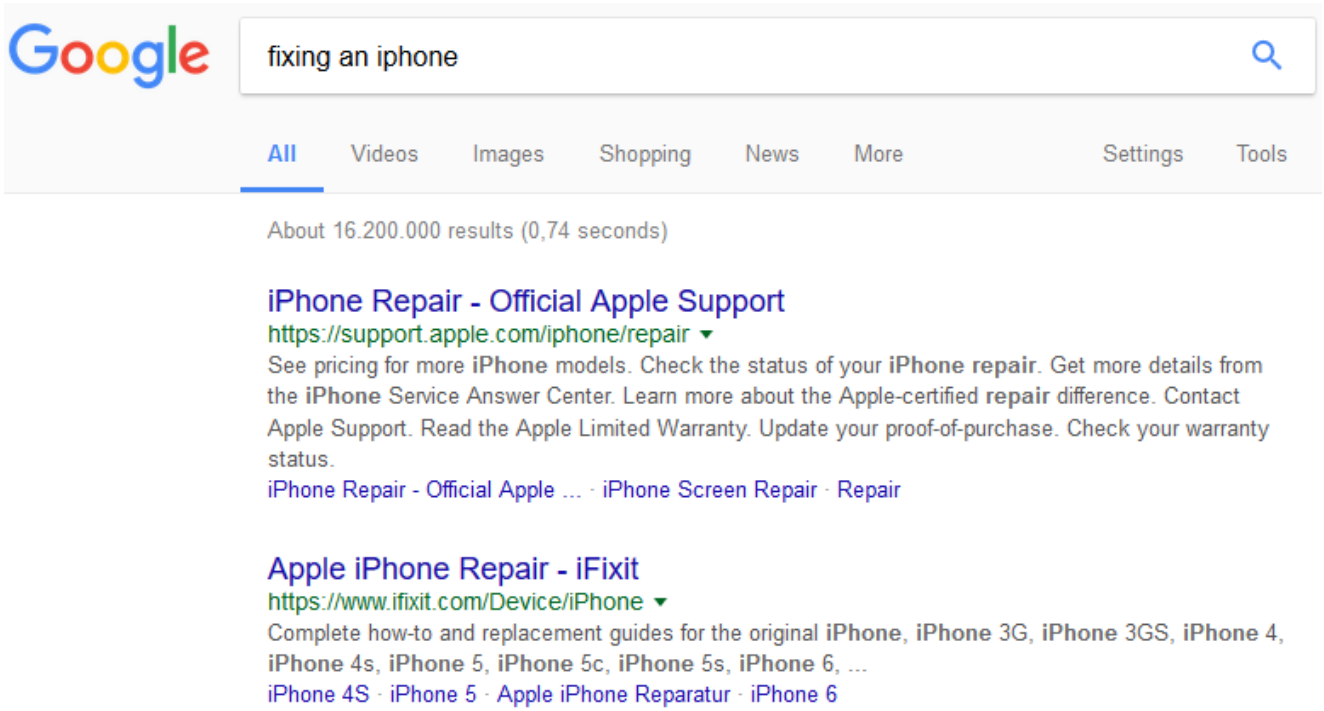
- Users re-submit the same queries over and over
- Studies have found that repeated queries by the same user are between 24% and 50% of the total number of queries

J. Teevan, E. Adar, R. Jones, and M. A. S. Potts, "Information re-retrieval: Repeat queries in yahoo's logs," in *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 151–158, New York, NY, USA: ACM, 2007.

M. Sanderson and S. T. Dumais, "Examining repetition in user search behavior," in *ECIR*, pp. 597–604, 2007.

Query Log Analysis for Enhancing Effectiveness of Search Systems

- **Query Expansion** – expand a query with terms that previous users have specified to improve the query

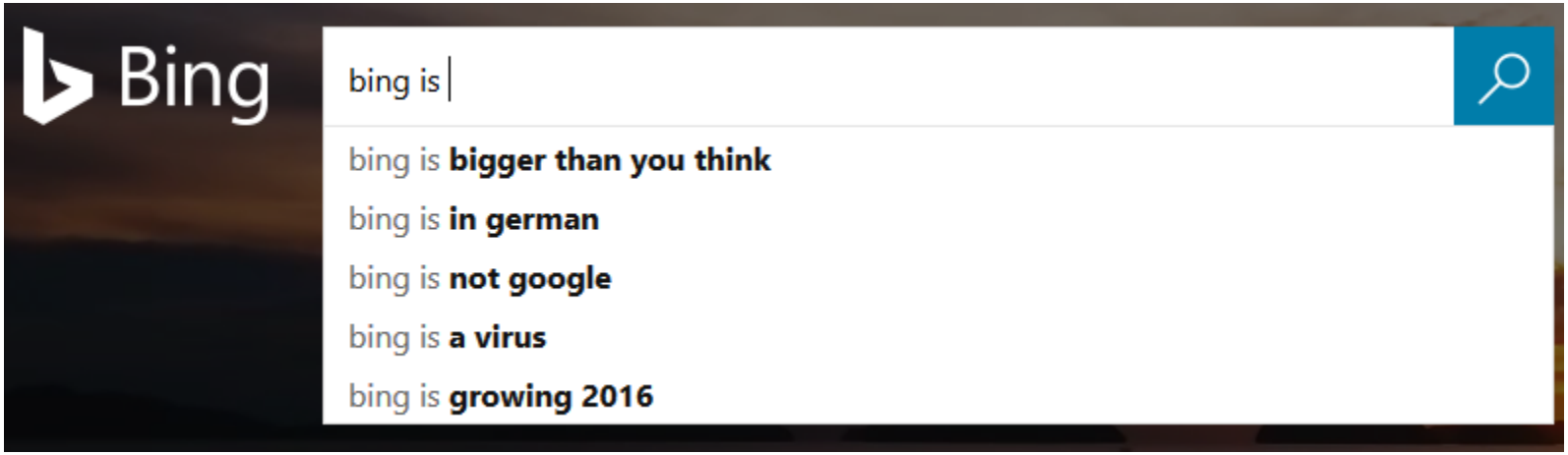


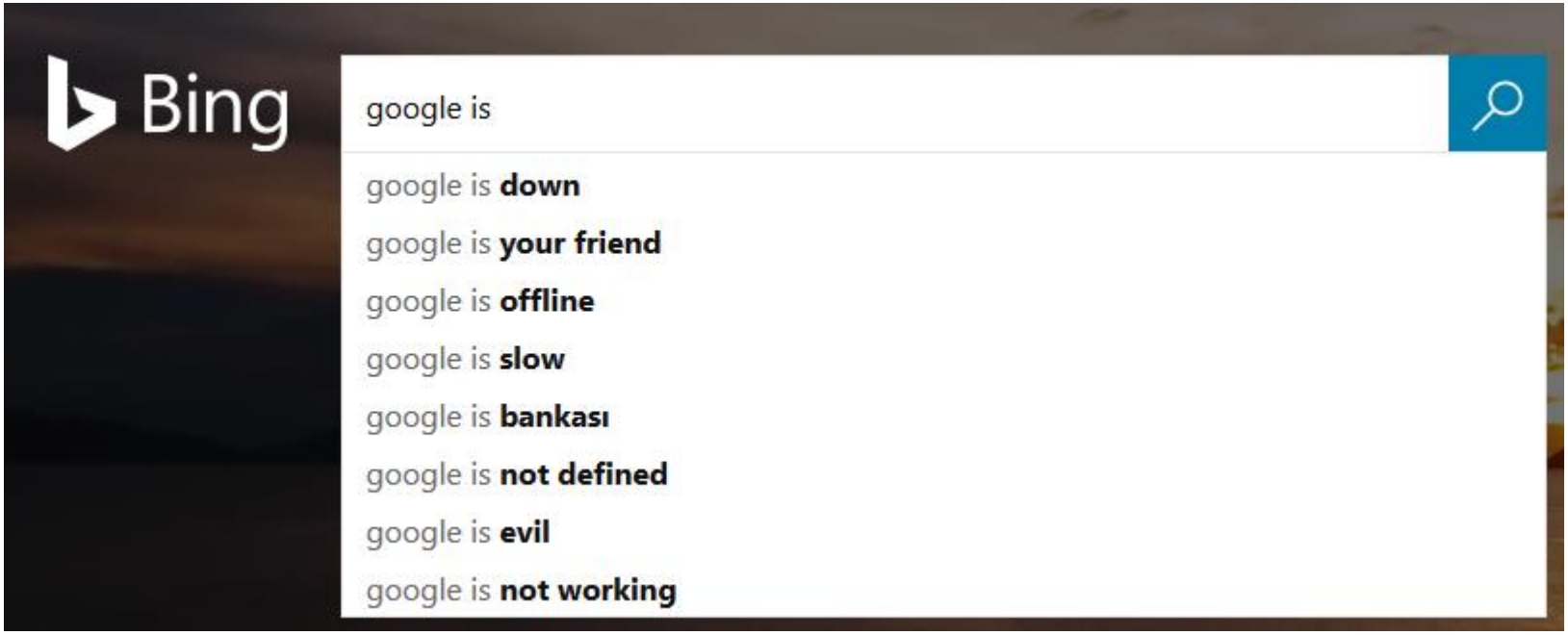
Query Log Analysis for Enhancing Effectiveness of Search Systems

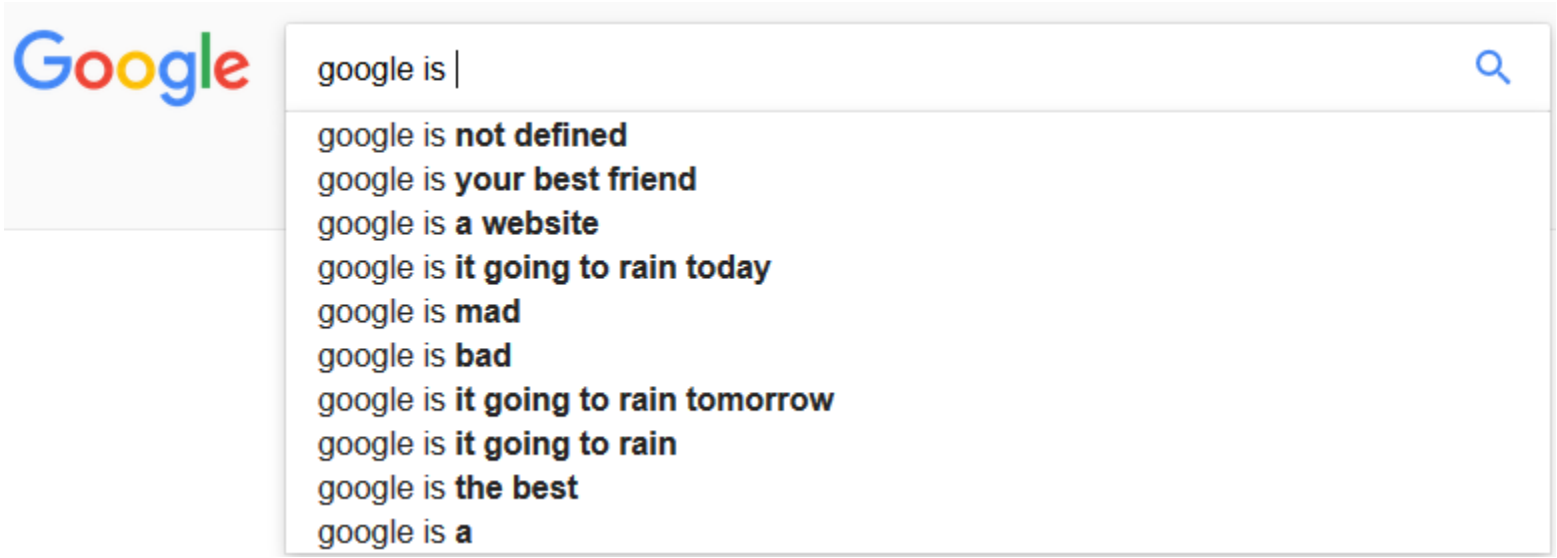
- **Query Suggestion** – use search log information to propose a list of queries related to the current query

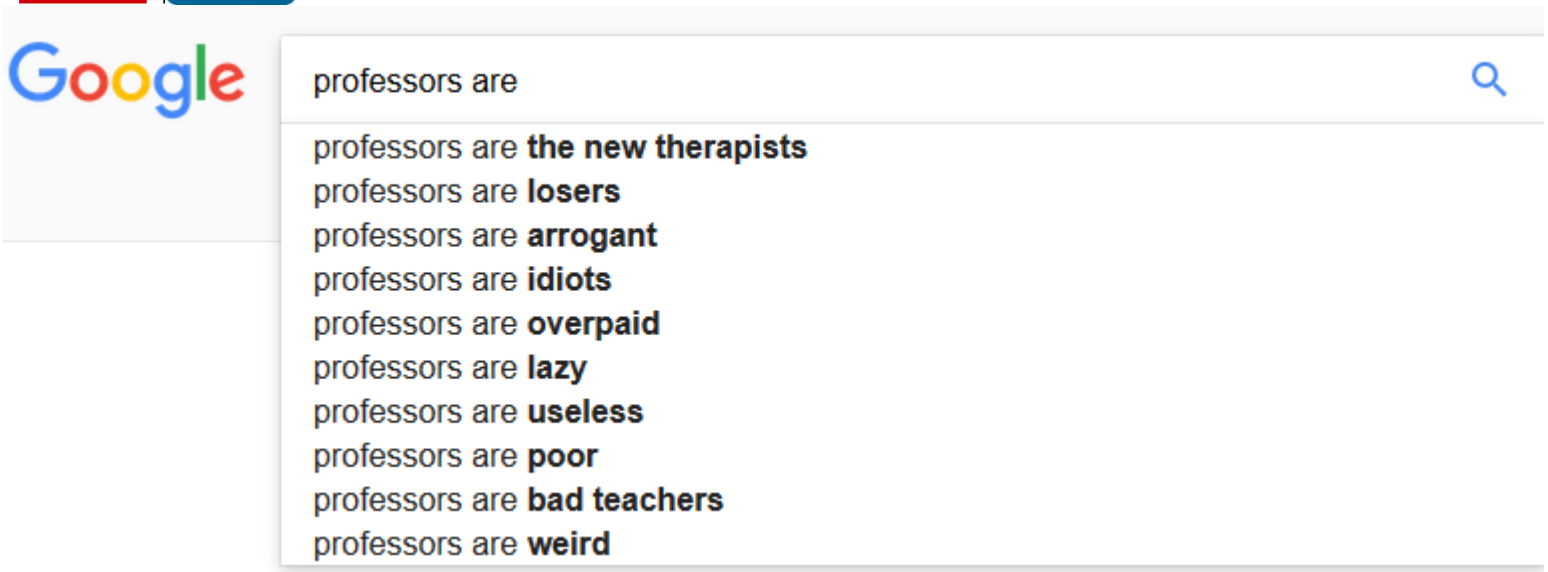
Bing

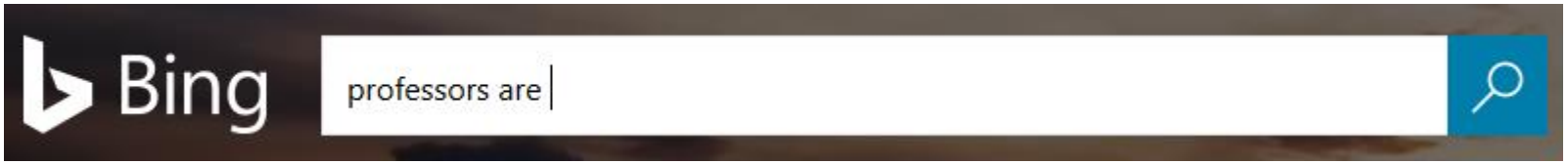
Google


Bing
bing is
bing is bigger than you think
bing is in german
bing is not google
bing is a virus
bing is growing 2016


Bing
google is
google is down
google is your friend
google is offline
google is slow
google is bankası
google is not defined
google is evil
google is not working


Google
google is
google is not defined
google is your best friend
google is a website
google is it going to rain today
google is mad
google is bad
google is it going to rain tomorrow
google is it going to rain
google is the best
google is a
Google
bing is
bing is not google
bing is trash
bing is bad
bing is owned by
bing is better
bing is dead
bing is a
bing is down
bing is not working
bing is the best search engine

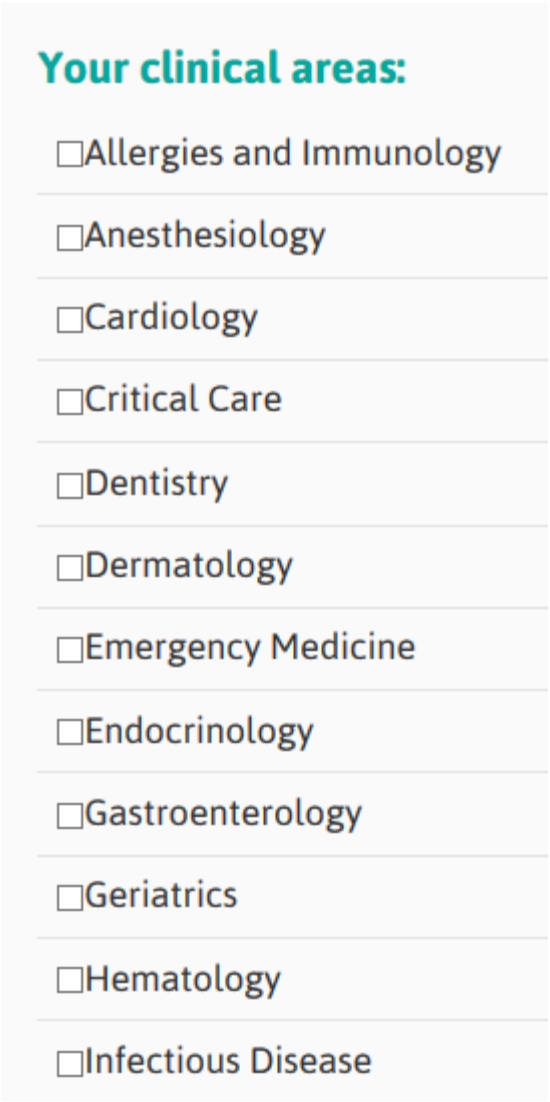

Google
professors are
professors are the new therapists
professors are losers
professors are arrogant
professors are idiots
professors are overpaid
professors are lazy
professors are useless
professors are poor
professors are bad teachers
professors are weird


Bing
professors are


Bing
professoren sind
professoren sind schlechter als ihr ruf

- **Personalised Query Results** – Delivering query results ranked according to the particular tastes of a precise user (or class of users).
  - Usually done by re-ranking search results according to a specific user profile built automatically
  - Alternatively, one can ask a user to give information about him/herself

TRIP Database



data

Sign in

Not logged in

Web Images Maps Videos Books More Search tools

About 1,400,000,000 results (0.29 seconds)

Ad related to **data**

IBM Big **Data** Lösungen - Big **Data** verstehen - ibm.com
www.ibm.com/at
IBM E-Book lesen!
IBM has 25,593 followers on Google+

Data - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Data**
Data is a set of values of qualitative or quantitative variables; restated, **data** are individual pieces of information. **Data** in computing (or **data** processing) are ...
Value - Data (disambiguation) - Data (computing) - Level of measurement

Data (Star Trek) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Data**_(Star_Trek)
Lieutenant Commander **Data** is a character in the fictional Star Trek universe portrayed by actor Brent Spiner. He appears in the television series Star Trek: The ...
Brent Spiner - List of Star Trek characters (N–S) - The Offspring

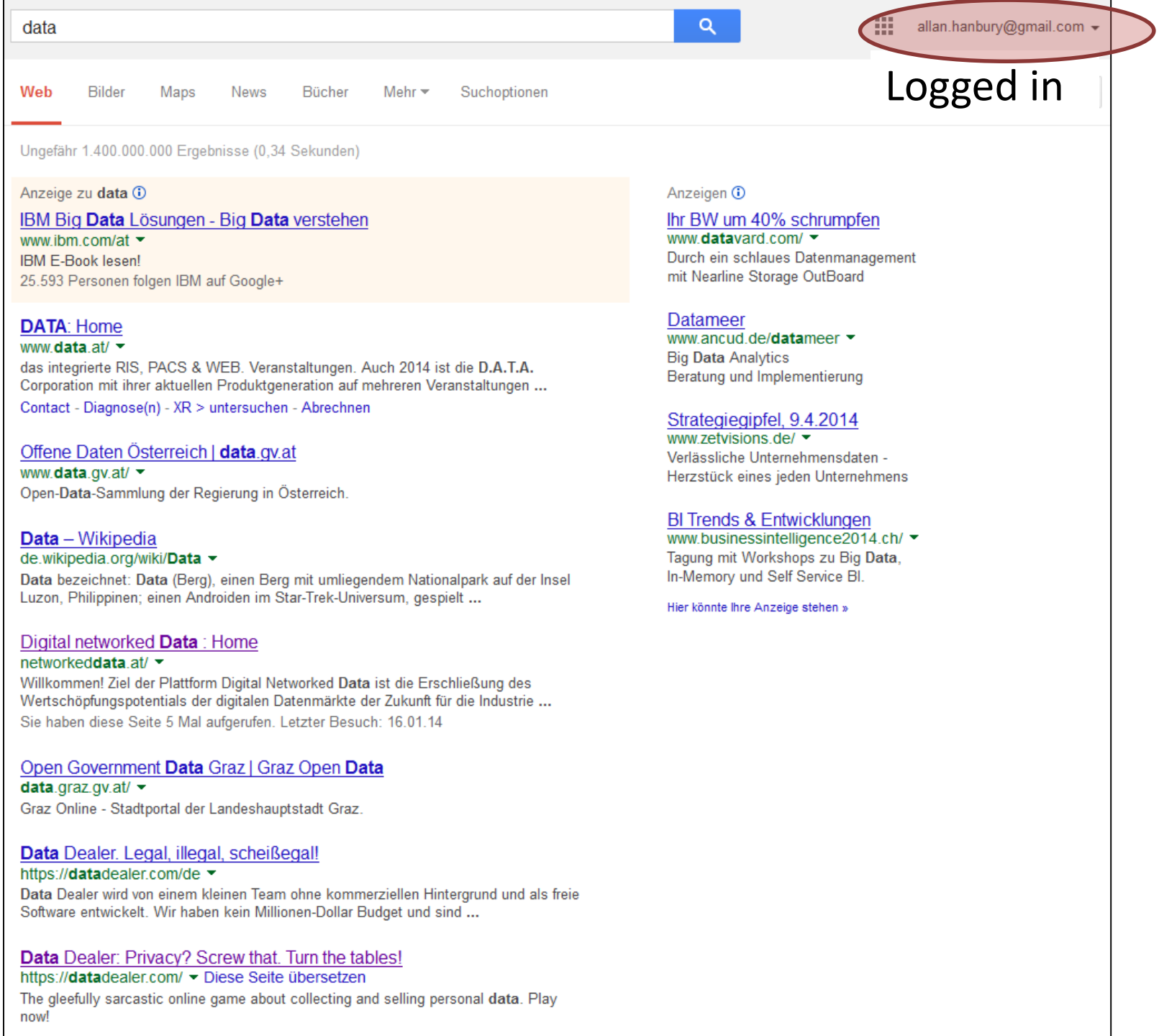
What is **data**? - A Word Definition From the Webopedia Computer ...
www.webopedia.com › TERM › D
This page describes the term **data** and lists other pages on the Web where you can find additional information.

Data - Definition and More from the Free Merriam-Webster Dictionary
www.merriam-webster.com/dictionary/**data**
da·ta. noun plural but singular or plural in construction, often attributive \ˈdā-tə, ˈda- also ˈdä-\. : facts or information used usually to calculate, analyze, or plan ...
Merriam-Webster Pronunciation - Data mining - Data bank - Data structure

The World Bank: **Data**
data.worldbank.org/
Indonesia: Spicing up research on sub-national development through open **data**. Indira Maulani Hapsari and Cut Dian | Posted on 5 Feb 2014. How open **data** ...

What is **data**? - Definition from WhatIs.com

searchdatamanagement.techtarget.com › ... › Data stewardship
by Margaret Rouse - in 1,480 Google+ circles
(1) In computing, **data** is information that has been translated into a form that is more convenient to move or process. Relative to today's computers and ...


data
allan.hanbury@gmail.com
Logged in
Web
Bilder
Maps
News
Bücher
Mehr
Suchoptionen
Ungefähr 1.400.000.000 Ergebnisse (0,34 Sekunden)
Anzeige zu data
IBM Big Data Lösungen - Big Data verstehen
www.ibm.com/at
IBM E-Book lesen!
25.593 Personen folgen IBM auf Google+
DATA: Home
www.data.at/
das integrierte RIS, PACS & WEB. Veranstaltungen. Auch 2014 ist die D.A.T.A. Corporation mit ihrer aktuellen Produktgeneration auf mehreren Veranstaltungen ...
Contact - Diagnose(n) - XR > untersuchen - Abrechnen
Offene Daten Österreich | data.gv.at
www.data.gv.at/
Open-Data-Sammlung der Regierung in Österreich.
Data – Wikipedia
de.wikipedia.org/wiki/Data
Data bezeichnet: Data (Berg), einen Berg mit umliegendem Nationalpark auf der Insel Luzon, Philippinen; einen Androiden im Star-Trek-Universum, gespielt ...
Digital networked Data : Home
networkeddata.at/
Willkommen! Ziel der Plattform Digital Networked Data ist die Erschließung des Wertschöpfungspotentials der digitalen Datenmärkte der Zukunft für die Industrie ...
Sie haben diese Seite 5 Mal aufgerufen. Letzter Besuch: 16.01.14
Open Government Data Graz | Graz Open Data
data.graz.gv.at/
Graz Online - Stadtportal der Landeshauptstadt Graz.
Data Dealer. Legal, illegal, scheißegal!
https://datadealer.com/de
Data Dealer wird von einem kleinen Team ohne kommerziellen Hintergrund und als freie Software entwickelt. Wir haben kein Millionen-Dollar Budget und sind ...
Data Dealer: Privacy? Screw that. Turn the tables!
https://datadealer.com/ Diese Seite übersetzen
The gleefully sarcastic online game about collecting and selling personal data. Play now!
Anzeigen
Ihr BW um 40% schrumpfen
www.datavard.com/
Durch ein schlaues Datenmanagement mit Nearline Storage OutBoard
Datameer
www.ancud.de/datameer
Big Data Analytics
Beratung und Implementierung
Strategiegipfel, 9.4.2014
www.zetvisions.de/
Verlässliche Unternehmensdaten - Herzstück eines jeden Unternehmens
BI Trends & Entwicklungen
www.businessintelligence2014.ch/
Tagung mit Workshops zu Big Data, In-Memory und Self Service BI.
Hier könnte Ihre Anzeige stehen »

- **Learning to Rank** – Use query logs as training data for learning to rank algorithms
- **Query Spelling Correction** – Use query logs to build spelling correction models that are based on actual usage of a language and not (only) on a pre-built vocabulary of terms

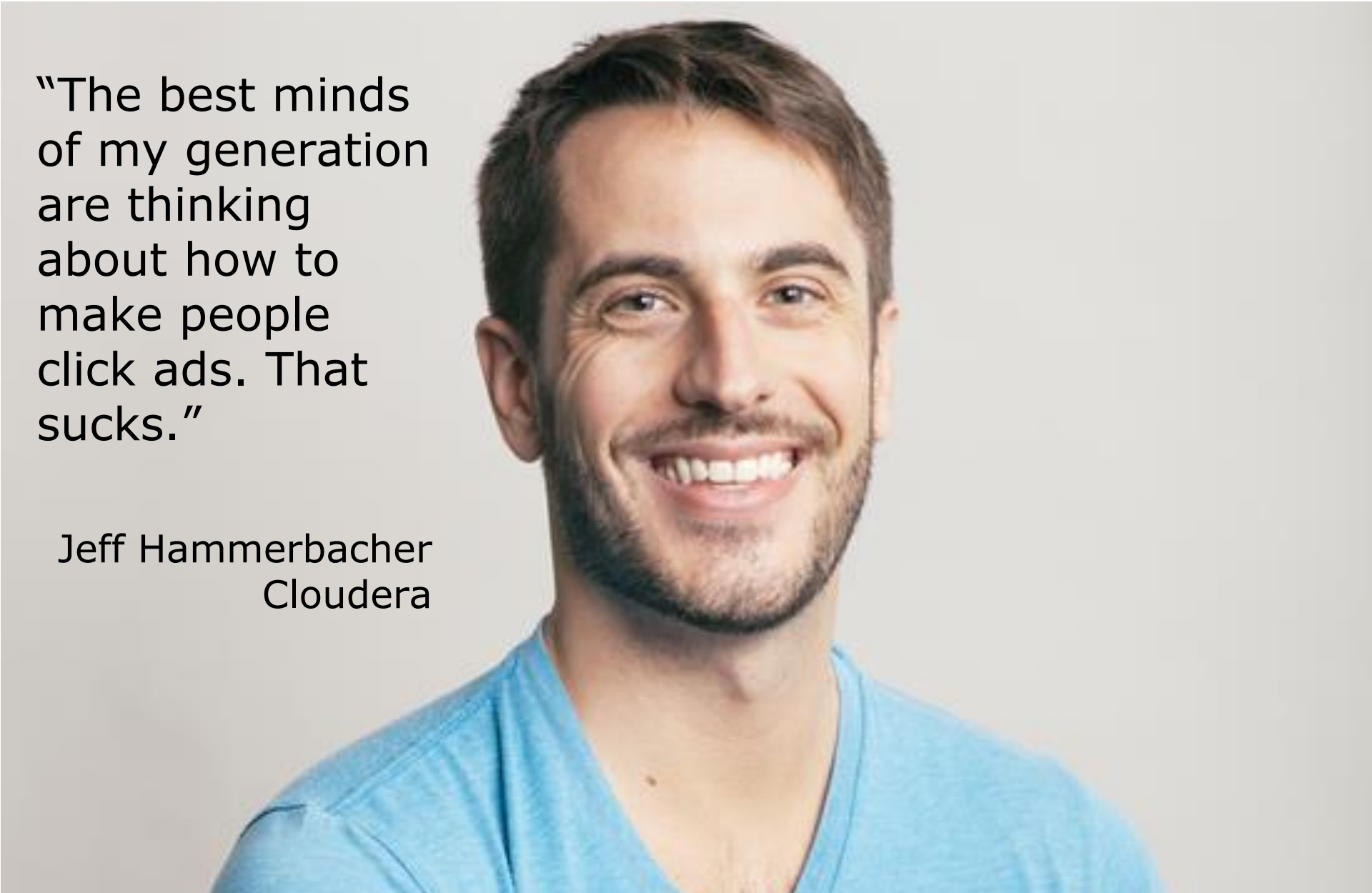
Drawbacks of Query Log Analysis

- Clicks…
  - Clicks can be taken as a proxy for the relevance of a document
  - However, users do not only click on documents that are relevant to the query
- People may not always want results that correspond to their profile…

Query Log Analysis for Enhancing Efficiency of Search Systems

- Usage patterns in web search engine logs can be exploited to design effective methods for enhancing efficiency
- **Caching** – exploit past usage information to build cache replacement policies suitable for search engine workloads
- **Data Partitioning** – design strategies to improve placement of data within a distributed web search engine.

Web Search Advertisement



Summary

- Crawling should be done efficiently but politely
- PageRank is important, but not a crucial component of web search
- Many insights can be gained from query log analysis