Information Storage and Retrieval

CSCE 670
Texas A&M University
Department of Computer Science & Engineering
Instructor: Prof. James Caverlee

Link Analysis: Getting Started
2 February 2017

Empirical investigation of the effect of ranking

- How can we measure how important ranking is?
- Observe what searchers do when they are searching in a controlled setting
  - Videotape them
  - Ask them to "think aloud"
  - Interview them
  - Eye-track them
  - Time them
  - Record and count their clicks
- The following slides are from Dan Russell's JCDL talk
- Dan Russell is the "Über Tech Lead for Search Quality & User Happiness" at Google.

Interview video

So.. Did you notice the FTD official site?

To be honest, I didn't even look at that.

At first I saw "from $20" and $20 is what I was looking for.

To be honest, 1800-flowers is what I'm familiar with and why I went there next even though I kind of assumed they wouldn't have $20 flowers

And you knew they were expensive?

I knew they were expensive but I thought "hey, maybe they've got some flowers for under $20 here…"

But you didn't notice the FTD?

No I didn't, actually… that's really funny.

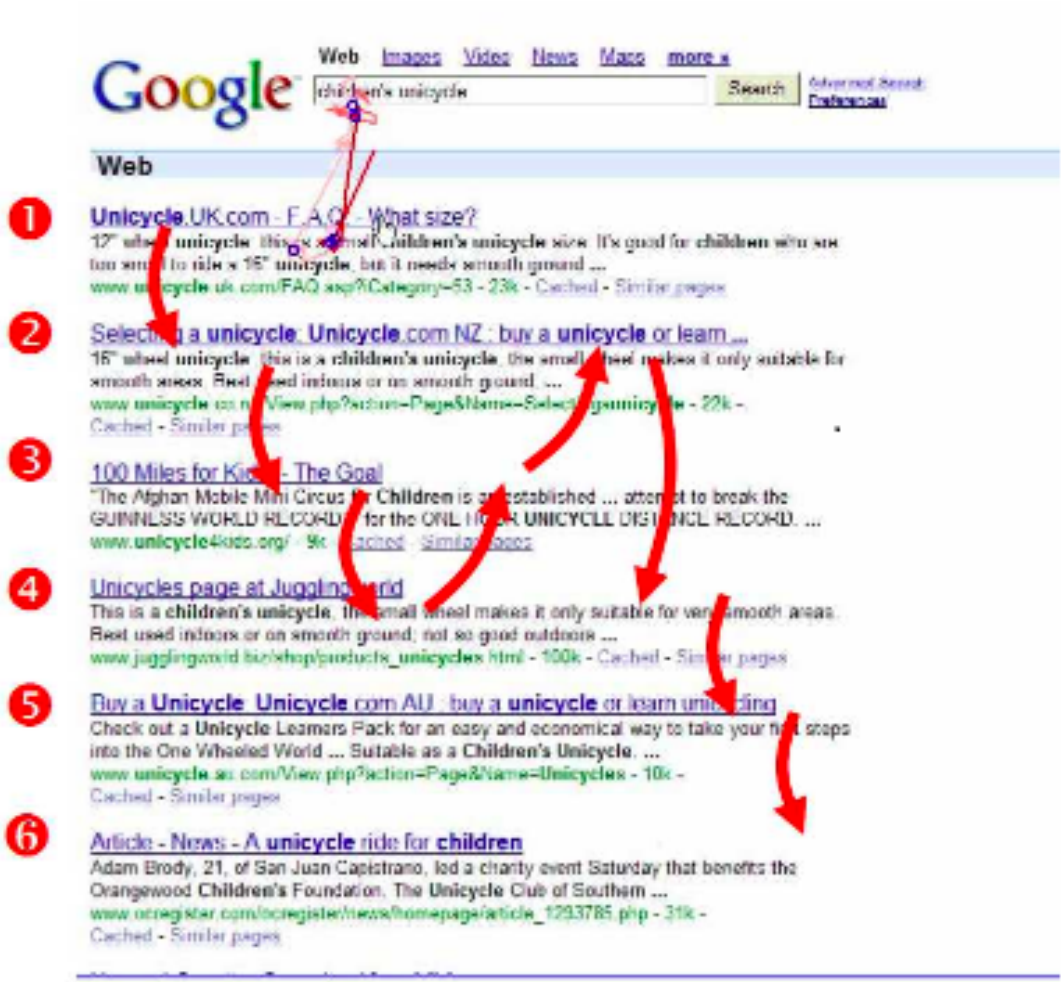
Rapidly scanning the results

Note scan pattern:

Page 3:
Result 1
Result 2
Result 3
Result 4
Result 3
Result 2
Result 4
Result 5
Result 6 <click>

Q: Why do this?

A: What's learned later influences judgment of earlier content.



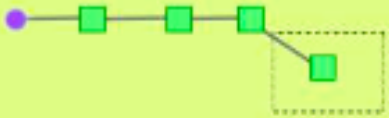
Google

Kinds of behaviors we see in the data

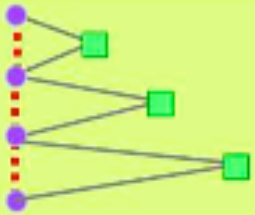
Short / Nav

Topic exploration

Topic switch

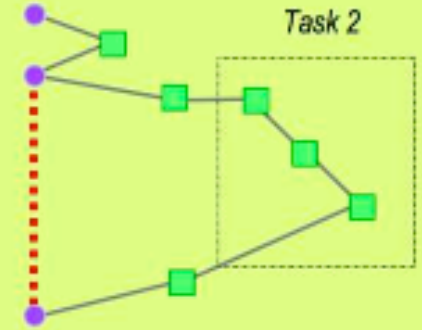


Methodical results exploration

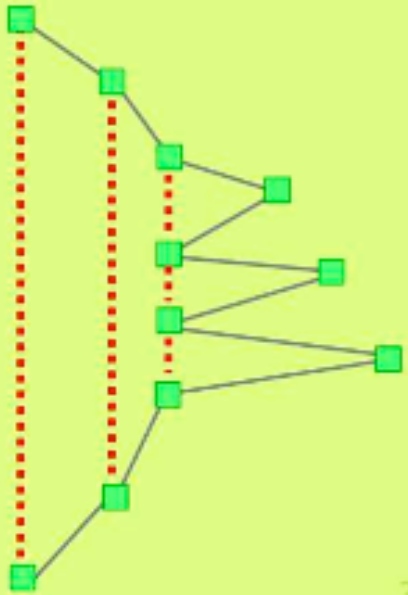


Query reform

Multitasking

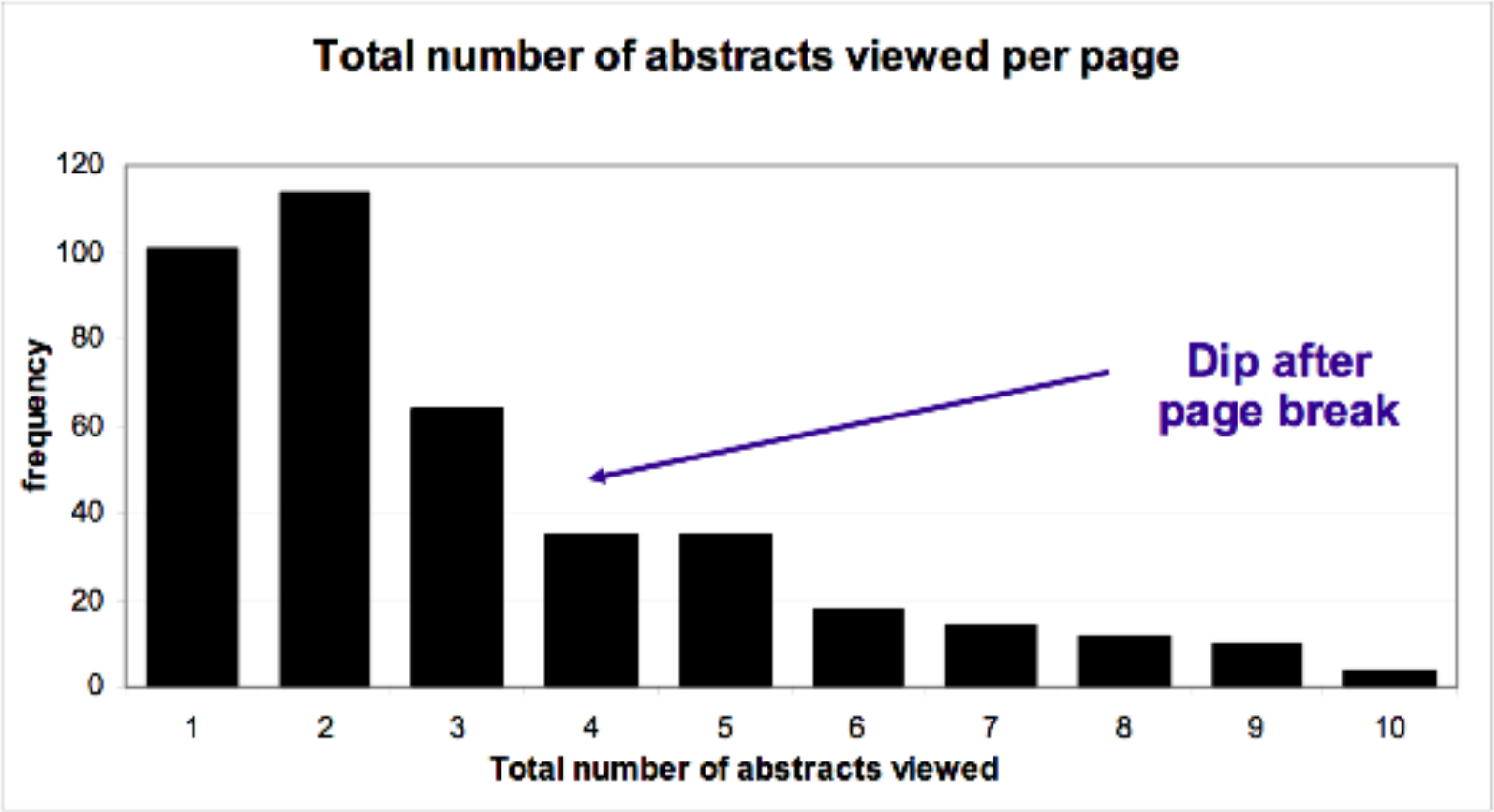


Stacking behavior



Google

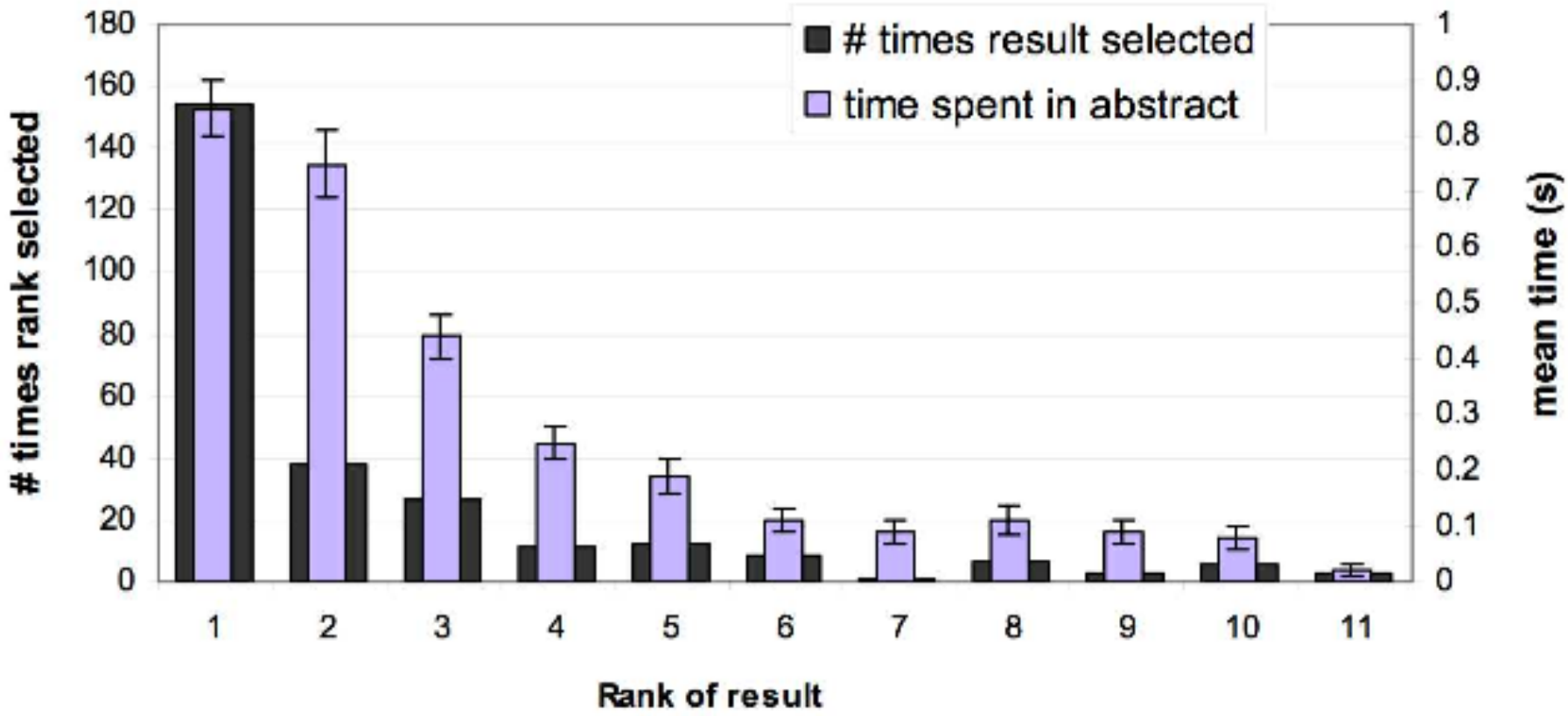
How many links do users view?



Mean: 3.07 Median/Mode: 2.00

Google

Looking vs. Clicking

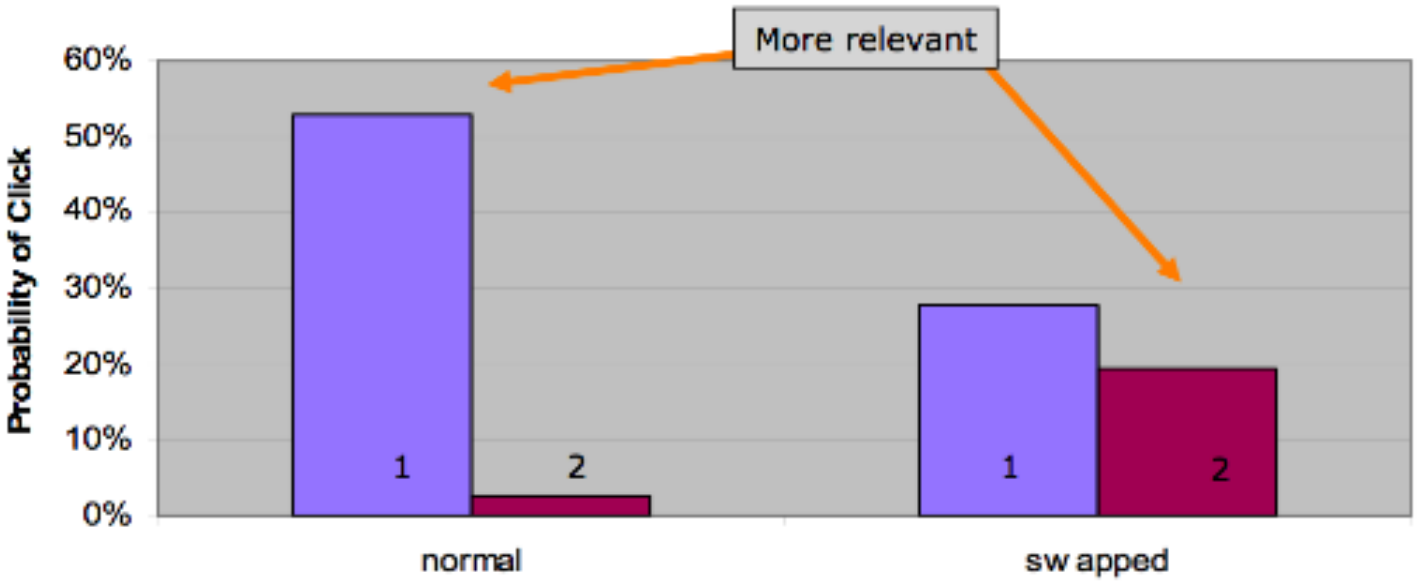


- Users view results one and two more often / thoroughly
- Users click most frequently on result one

Google

Presentation bias – reversed results

- Order of presentation influences where users look **AND** where they click



Google

IR History (Again)

Hans Peter Luhn

TF (1957)

The weight of a term that occurs in a document is simply proportional to the term frequency.

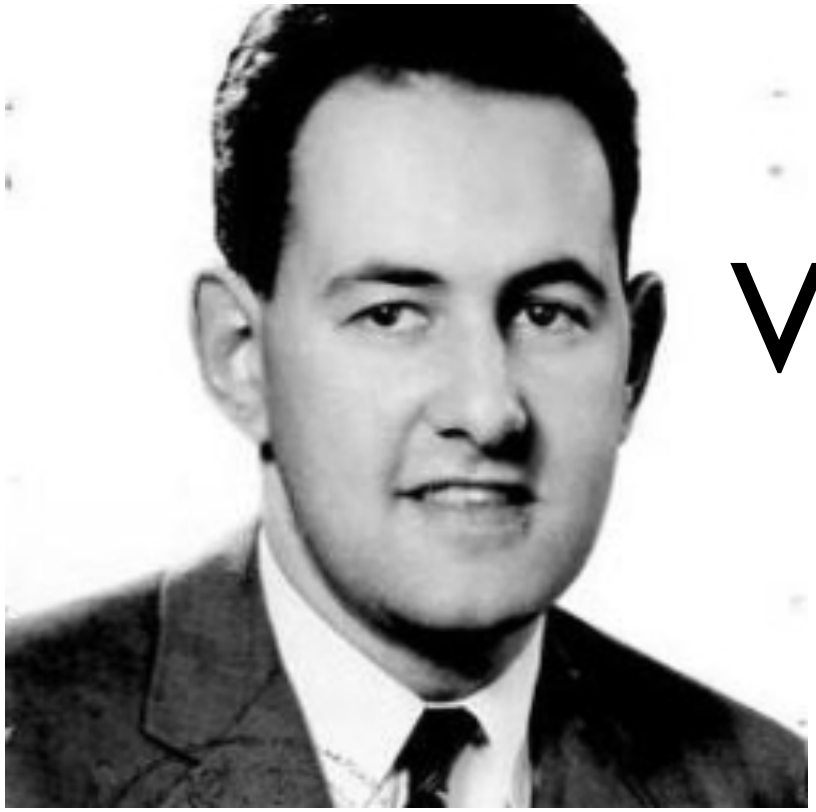
IBM Researcher
Foundational work in the 1950s
(invented an algorithm to checksum your credit card numbers!)

Karen Spärck Jones

IDF (1972)

Gerard Salton

Father of Information Retrieval



TF-IDF (1975)
Vector Space Model

SMART system (at Cornell)

Amit Singhal (PhD 1996 with Salton)

Joins Google 2000 : Later, Head of Search

Maron and Kuhns

Probabilistic IR (1960)

Stephen Robertson

with Karen Spark Jones
Binary Independence Model (1976)

Larry Page and Sergey Brin

PageRank (1997-8)

Jon Kleinberg

HITS (1998)

Professor at Cornell

Web Search: Pre-History

Brief (non-technical) history of Web search

- Early keyword-based engines ca. 1995-1997
  - Altavista, Excite, Infoseek, Inktomi, Lycos
- Paid search ranking: Goto (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords: ***casino*** was expensive!

ALTAVISTA
Technology, Inc.

View Multimedia From Our Vantage Point

Search the Web and Display the Results in Standard Form

Submit

Search with Digital's Alta Vista [Advanced Search]

Add your URL to the most popular Search Engines

Free Software
Download Now...

Contests
Make Me Laugh...

Creative Web
Create a Site...

ALTAVISTA
Technology, Inc.

[Creative][Search][Humor]

Search for information about:

in the World Wide Web Search Now

Infoseek Guide is best viewed with:

Want personalized news? Get Personal now!

Basic Search Tips:

- Click in the box above and type a few words that describe what you want to find. For example, typing **growing orchids indoors** will find sites about caring for orchids.
- If you are looking for a person or place, type the name, starting with capital letters. For example, typing **Florence Italy** will find sites about this famous city.
- These detailed search tips describe how to use the features of Infoseek Guide to find what you are looking for.
- For the broadest results, you can search the entire **World Wide Web**.
- To restrict your search to hand-picked and categorized sites, choose **Infoseek Select Sites**.
- Or just search for a category within Infoseek Select by choosing **Categories of Sites**.
- To search through Internet discussion forums (similar to bulletin boards), choose **Usenet Newsgroups**.
- To search for someone's e-mail address, choose **E-mail Addresses**.
- To search through news stories within the past month, choose **Reuters News**.
- To search through answers to Frequently Asked Questions, choose **Web FAQs**.

Explore these popular Infoseek Select topics:

- Arts & Entertainment
- Business & Finance
- Computers & Internet
- Education
- Government & Politics
- Health & Medicine
- Living
- News
- Reference
- Science & Technology
- Sports
- Travel

Try Infoseek Personal, your personalized news service

Customize | Add Site | Help | Feedback
Download iSeek | About Infoseek

Click here to try Microsoft Money 97 FREE

It's amazing where
Go Get It will get you.

Find: Go Get It

Enhance your search.

New Search • TopNews • Sites by Subject • Top 5% Sites • City Guide • Pictures & Sounds

PeopleFind • Point Review • Road Maps • Software • About Lycos • Club Lycos • Help

Add Your Site to Lycos

Questions & Comments

"Turbo Search!"
Download Excite Direct

Take an ExciteSeeing Tour

Excite on TV

Make your website searchable, FREE!

Excite Search: twice the power of the competition.

What: search

Where: World Wide Web

[Help]
[Advanced Search]

Excite Reviews: site reviews by the web's best editorial team.

- Arts
- Business
- Computing
- Education
- Entertainment
- Health
- Hobbies
- Life & Style
- Money
- News & Reference
- Personal Pages
- Politics & Law
- Regional
- Science
- Shopping
- Sports

Excite City.Net

Plan your weekend, your travels.

Find-A-Destination

Take me there!

Maps · Top Cities · Concierge

ExciteSeeing Tours

Choose from hundreds.

- X-Files: The truth is out there!
- Dr. Ruth's guide to safer sex
- Windows 95 shareware and freeware
- Celebrating Thanksgiving
- Investing in high-tech stocks
- New to the Net?

Excite Live!

Your news, your way.

- Latest news
- Sports scores
- Local weather
- Movie reviews
- Stock quotes
- TV listings
- Horoscopes
- Site reviews

Excite Reference

Just the facts, ma'am.

- Yellow Pages
- People Finder
- Email Lookup
- Maps
- Shareware
- Dictionary

Search the web using Google!

10 results | Google Search | I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

About Google!

Stanford Search Linux Search

Get Google! updates monthly!

your e-mail | Subscribe | Archive

Search the web using Google!

Google Search I'm feeling lucky

Special Searches
Stanford Search
Linux Search

Help!
About Google!
Company Info
Google! Logos

Get Google!
updates monthly:
your e-mail
Subscribe Archive

Jobs@Google About Google

Search the web using Google

Google Search I'm feeling lucky

Google Launches! Read the press release.

Brief (non-technical) history

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines save Inktomi
  - Great user experience in search of a business model
  - Meanwhile Goto/Overture's annual revenues were nearing $1 billion
- Result: Google added paid search "ads" to the side, independent of search results
  - Yahoo followed suit, acquiring Overture (for paid placement) and Inktomi (for search)
- 2005+: Google gains search share, dominating in Europe and very strong in North America
  - 2009: Yahoo! and Microsoft propose combined paid search offering

Google iphone

Web News Images Shopping Videos More Search tools

iPhone 5c
Ad www.apple.com/
For the colorful. Learn more.
iPhone 5s
iPhone 5c
Back to School
Buy now

iPhone 5s at T-Mobile®
Ad www.t-mobile.com/iPhone5s
3.5 ★★★½☆ rating for t-mobile.com
Unlimited talk, text & web plans on an advanced nationwide 4G network.
Ratings: Price 10/10 - Sign-up 10/10 - Website 9/10 - Customer service 9/10
T-Mobile has 184,607 followers on Google+
Simple Choice Plan - iPhone 5c - Find a store - Family plans

iPhone 5s on Sprint® - sprint.com
Ad www.sprint.com/iPhone5s
Qualify w/ $0 down iPhone 5s & get iPad Mini for $49.99. Learn more.
Ratings: Sign-up 10/10 - Website 8.5/10 - Plan selection 7.5/10 - Price 7/10
Sprint has 20,780 followers on Google+
iPhone 5c for free - iPhone 4s - Shop iPhone - iPhone 5s

News for iphone

iPhone 6's RAM still up in the air
CNET - by Lance Whitney - 2 hours ago
A blog site's report that the next iPhone will stick with 1GB of RAM turns out to be based on a misreading of a technical diagram, but the ...

To get cash for an iPhone 6, lock in your old iPhone trade-in ...
CNET - by Eric Mack - 4 hours ago

Apple iPhone 6 rumor roundup: Specs, price, release date
ZDNet - by Charlie Osborne - 3 hours ago

More news for iphone

Apple - iPhone
https://www.apple.com/iphone/ Apple Inc.
Discover everything iPhone, including the most advanced mobile OS in its most advanced form and great apps that let you be creative and productive.
iPhone 4s Tips and Tricks - iPhone 5s - Tips and Tricks - Apple Support - iPhone 5s

Shop for iphone on Google Sponsored


Apple iPhone 5c 16GB (wit...
$0.00
Sprint
★★★★½ (2k+)


iPhone 5c - 16GB in Yello...
$59.99
Verizon Wirel...
★★★★½ (2k+)


Apple iPhone 5c - 16GB - Bl...
$49.99
AT&T
★★★★½ (2k+)


Apple iPhone 4s - 8GB - Bla...
$0.00
AT&T
★★★★½ (4k+)


Apple iPhone 4s - 8GB - Wh...
$0.00
AT&T
★★★★½ (4k+)

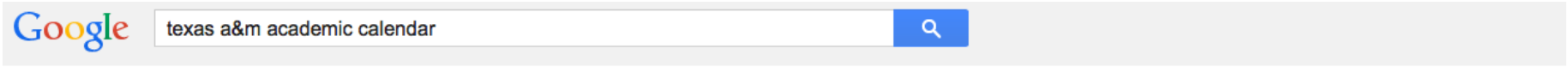

iPhone 4s - 8GB in White...
$0.00
Verizon Wirel...
★★★★½ (4k+)


Apple iPhone 5s - 16GB - G...
$199.99
AT&T
★★★★½ (5k+)


Apple iPhone 4s - 32GB - W...
$49.99
AT&T
★★★★½ (4k+)

Ad
iPhone At Best Buy®
www.bestbuy.com/iPhone
4.5 ★★★★½ rating for bestbuy.com
Free Shipping On The iPhone.
Shop Best Buy® For Great Deals.
805 Texas Ave S, College Station, TX
(979) 693-2745

See your ad here »



Web Shopping News Maps Images More Search tools

About 1,700,000 results (0.26 seconds)

Academic Calendar - Office of the Registrar - Texas A&M ...
registrar.tamu.edu/general/calendar.aspx ▾ Texas A&M University ▾
Academic Calendar. Print Print Calendar. University **Academic Calendar**. Spring 2014 | Summer 2014 | Fall 2014 | Spring 2015 | Summer 2015. Download iCal.

Final Examination Schedules
Final Examination Schedules. Spring 2014 | 1st Summer ...

Student Business Services
Tuition Estimator - Important Dates - Payment/Refunds - About SBS

More results from tamu.edu »

Academic Calendar - Texas A&M University Calendar
calendar.tamu.edu/academic/?&upcoming... ▾ Texas A&M University ▾
Events from **Academic Calendar** at **Texas A&M** University in College Station, TX, and from the university's colleges, departments, offices, and other organizations ...

Texas A&M University Calendar
calendar.tamu.edu/ ▾ Texas A&M University ▾
Events from **Texas A&M** University **Calendar** at **Texas A&M** University in College Station, TX, and from the university's colleges, departments, offices, and other ...

Academic Calendar - Texas A&M University School of Law
law.tamu.edu/.../AcademicCal... ▾ Texas Wesleyan University School of Law ▾
2014-2015 **Academic** Year. Printer-friendly version. Fall 2014. August 20, Graduation application open in Howdy for all students planning to graduate in ...

[PDF] **Calendar** 2014-2015 - **Texas A&M** University at Galveston
www.tamug.edu/.../TAMUG2014-2... ▾ Texas A&M University at Galveston ▾
Texas A&M University at Galveston **Academic Calendar** 2014-2015. 2014 Fall Semester*. August 20. Graduation application opens for all students planning to ...

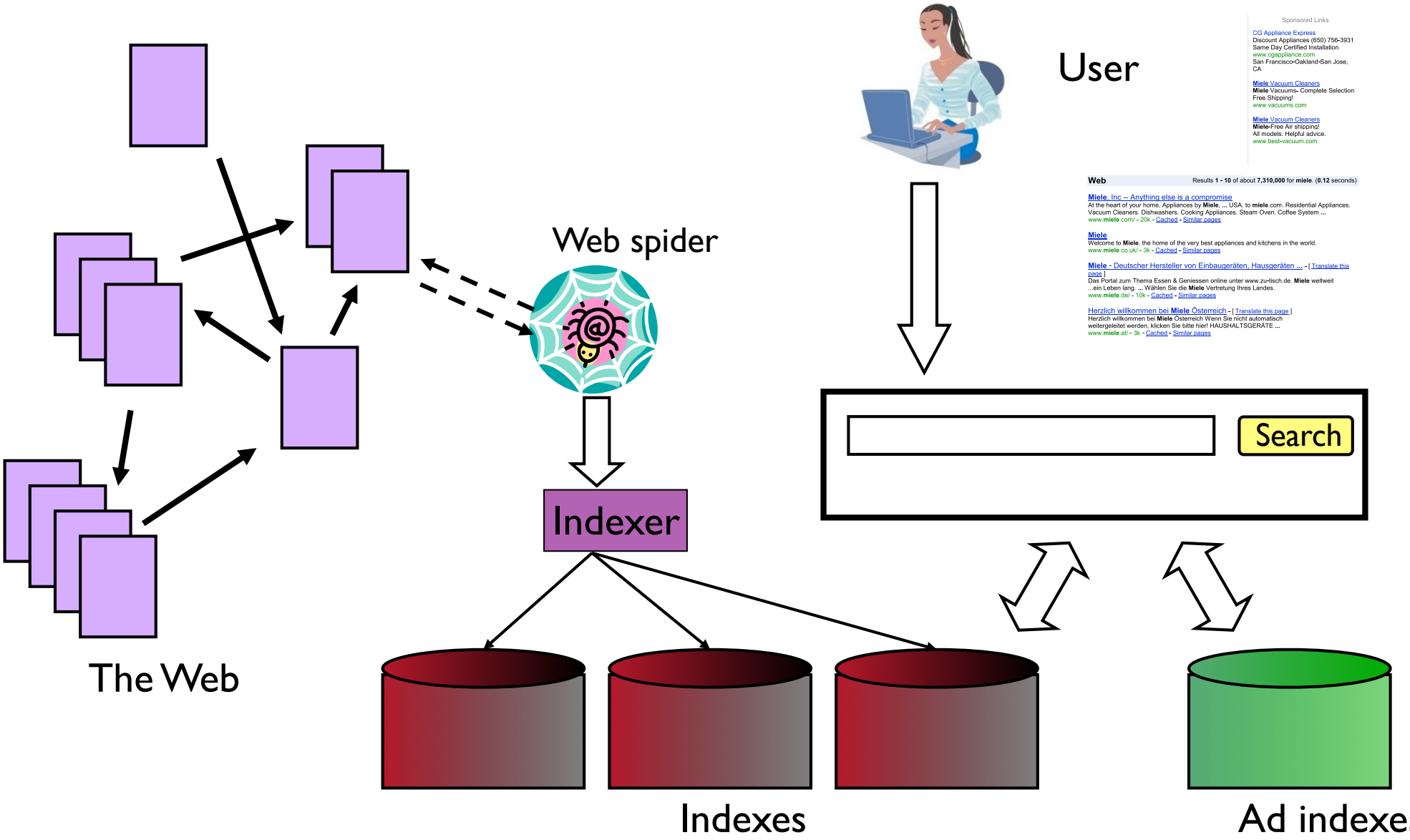
[PDF] Complete 2013-2014 **Calendar** PDF Version - **Texas A&**...
www.tamug.edu/.../TAMUG2013-2... ▾ Texas A&M University at Galveston ▾
Texas A&M University at Galveston **Academic Calendar** 2013-2014. 2013 Summer Term I*. May 15. Graduation application opens for all students planning to ...

Academic Calendar - Texas A&M University at Galveston
www.tamug.edu/.../calendar.html ▾ Texas A&M University at Galveston ▾
The **Texas A&M** University at Galveston **Academic Calendar** mirrors the **Texas A&M** University (College Station) Calendar, with minor changes. The Calendar is ...

Searches related to **texas a&m academic calendar**

texas a&m **football schedule**
texas a&m academic calendar **2012-13**
southwest airlines
texas a&m academic calendar **2013-14**

Web search basics

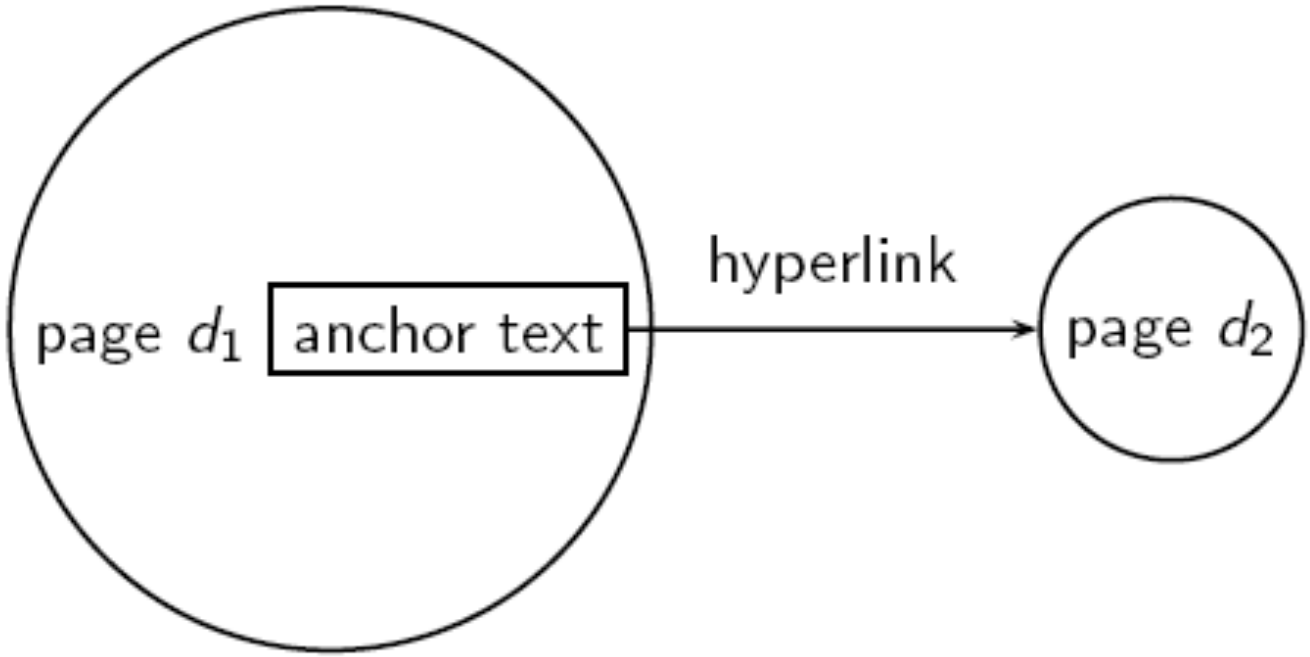


Today: Link Analysis

- Anchor text
- PageRank

Anchor text

The Web as a Directed Graph

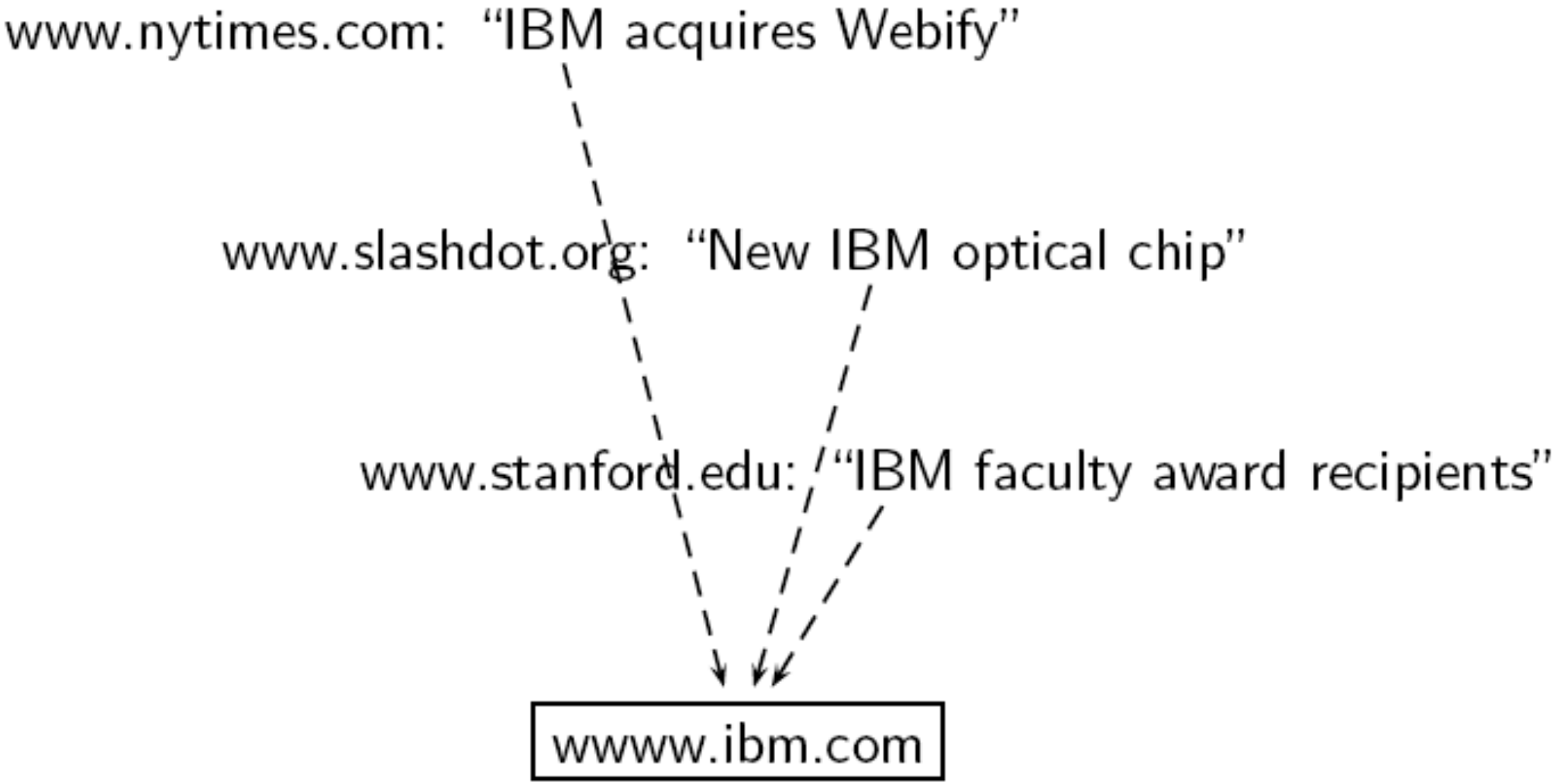


- **Assumption 1: A hyperlink is a quality signal**
  - A hyperlink between pages denotes that the author perceived relevance
- **Assumption 2: The anchor text describes the target page**
  - We use anchor text somewhat loosely here: the text surrounding the hyperlink. Example: "You can find cheap cars here."

[document text only] vs. [document text + anchor text]

- Searching on [document text + anchor text] is often more effective than searching on [document text 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 graphical)
- Searching on anchor text is better for the query IBM.
- **Represent each page by all the anchor text pointing to it.**
- In this representation, the page with the most occurrences of IBM is www.ibm.com.

Anchor text containing ***IBM*** pointing to www.ibm.com



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 Assumptions 1&2)
- Indexing anchor text can have unexpected side effects – Google bombs.
- A Google bomb is a search with "bad" results due to maliciously manipulated anchor text.
- Google introduced a new weighting function in January 2007 that fixed many google bombs.
- Any "live" Google bombs?

PageRank

Link-based ranking

- Query processing with link-based ranking:
  - First retrieve all pages meeting the query (say venture capital)
  - Order these by their link popularity (= citation frequency, first generation)
- . . . or by Pagerank (second generation)

- Simple link popularity (= number of inlinks of a page) is easy to spam.
- Why?