IS2140 Information Storage and Retrieval

Unit 9: Web Information Retrieval

Daqing He
School of Information Sciences
University of Pittsburgh

October 29, 2018

Midterm Exam on November 5

- The first 100 minutes of the class
  - Do not be late
- Covers the materials up to this week: unit 1 to unit 9
  - The topics covered by the lectures only
    - Using slides as the guidance for coverage
  - Reading/assignments help to consolidate the topics
- Normal exam questions
  - Concentrate on important concepts and some simple calculations
  - Will not require to programming
- Open book exam, complete on your computers
  - You can bring all the related books, lecture notes, reading, reference books etc.
- But complete the exam yourself, no collaboration during the exam
- Cannot leave once the exam starts until you submit the exam paper

Muddiest Points

- Relevance Feedback
  - When we implement the explicit relevance feedback and the implicit relevance feedback, are there some conflicts in their results?
  - we mentioned that nowadays search engines could be able to monitor user's behavior and then imply whether the user is satisfied with the search results. How can search engine be able to achieve that? Does this kind of monitoring also violate user's privacy?
  - I know there is a kind of feedback approach been used by the search engine, that is the engine will record the click amount of each hits, and the total time users stay on the page which linked to the hit. Then modify the weight of each query result. So, what type of relevance feedback should this kind of approach belong to?

Muddiest Points

- Blind Relevance Feedback
  - this kind of feedback is given by the system instead of the user. Why it can improve the precision? It may misunderstand the user's intention.
- Rocchio Method
  - For Rocchio Algorithm, why don't we let alpha + beta - gamma = 1?
  - How to set parameters in Rocchio Algorithm? Why gamma is typically smaller than beta?
  - If we use blind relevance feedback to produce query vector based on Rocchio Algorithm, how can we acquire the known non-relevant document part while we only get relevant part by using blind relevance feedback?

Muddiest Points

- Interface Design
  - How important is the user interface? If the search engine is a not so bad design with super powerful algorithm, will it be able to beat Google?
  - Is this design truly an aspect of Information Retrieval itself? It seems the same results were returned and the same information was displayed to the user, only the layout changed. This seems more like an issue of design or UX that doesn't interact with the information-retrieval aspects of the system.
  - Why most search engine uses left side of screen rather than center of screen?
  - We talked about that document surrogate influence user clicks. However, as a search engine/website, why do we care about the click for each document if all of them are presented the same way?

Muddiest Points

- Interface Design
  - Is a higher click-through rate always a desirable goal or can this measure be gamed? A system that generates document summaries in the form of click-bait article titles will have a higher click-through rate, but this doesn't necessarily reflect higher relevance or user satisfaction. How do tell if our system is falsely presenting relevance even for irrelevant results?
  - Since IR system encourage user to type long/complex queries, and Google allows users to type long query as they want, why do they make their search box a one-line input field instead of a multiple-line textarea? Wouldn't it confuse the users and make them to think that they can only type a very few characters as the queries?

Muddiest Points

- Query formulation
  - how do we select the documents if a user searches "Jaguar"? Do we return the most frequently searched topic or do we return them as equal?
  - In the PPT p.42, how to obtain contextual information from users such as previous queries, previous relevant documents? Whether IR systems have a database in the background to save histories of each user and relevant information or just save them in user's local.
  - I tried the type 1+1, 1 add 1, 1 add one in google, in these three cases, I can get correct results. However, when I typed in one plus one, 1 plus one, one add one, under these circumstance, I cannot get a calculated result, I got document search result instead.

Muddiest Points

- Document Surrogates
  - Why search engine like google or baidu do not use group-result strategy like cat-a-cone?

Outline

- Web and Web search
- Web Crawler
- Link Analysis

Class Goals

- After this class, you should be able to
  - know the basic characteristics of Web and Web searches
  - Know the basics of HITS and PageRank ideas
  - Be able to integrate HITS or PageRank method into future implementation

Interactive IR

Grouping Search Results

- Categorizing search results
  - Key idea: utilizing existing categories to group documents into categories
  - Example: Cat-a-Cone
- Clustering search results
  - Based on inter-document similarity
    - Computed using the cosine measure, for example
  - Text; Vivisimo's clusty, scatter/gather
  - 1D: scatter/gather
  - 2D: Kartoo
  - 3D: ThemeView

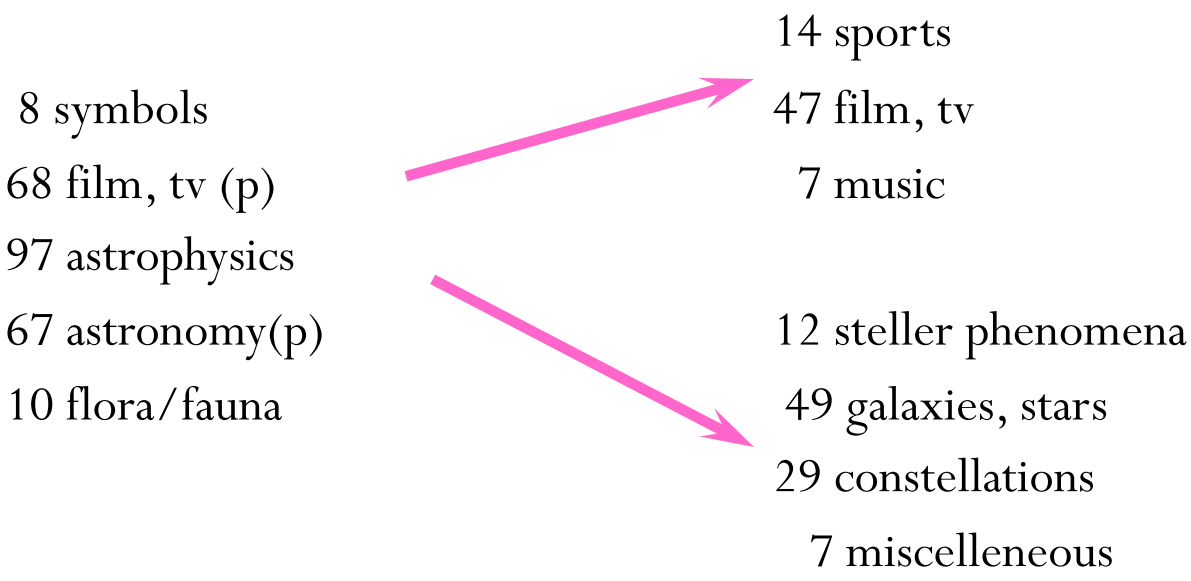
Scatter/Gather

Cutting, Pedersen, Tukey & Karger 92, 93, Hearst & Pedersen 95

- How it works
  - Cluster sets of documents into general "themes", just like a table of contents
  - Display the contents of the clusters by showing topical terms and typical titles
  - User chooses subsets of the clusters and re-clusters the documents within
  - Resulting new groups have different "themes"
- Originally used to give collection overview
- Evidence suggests more appropriate for displaying retrieval results in context

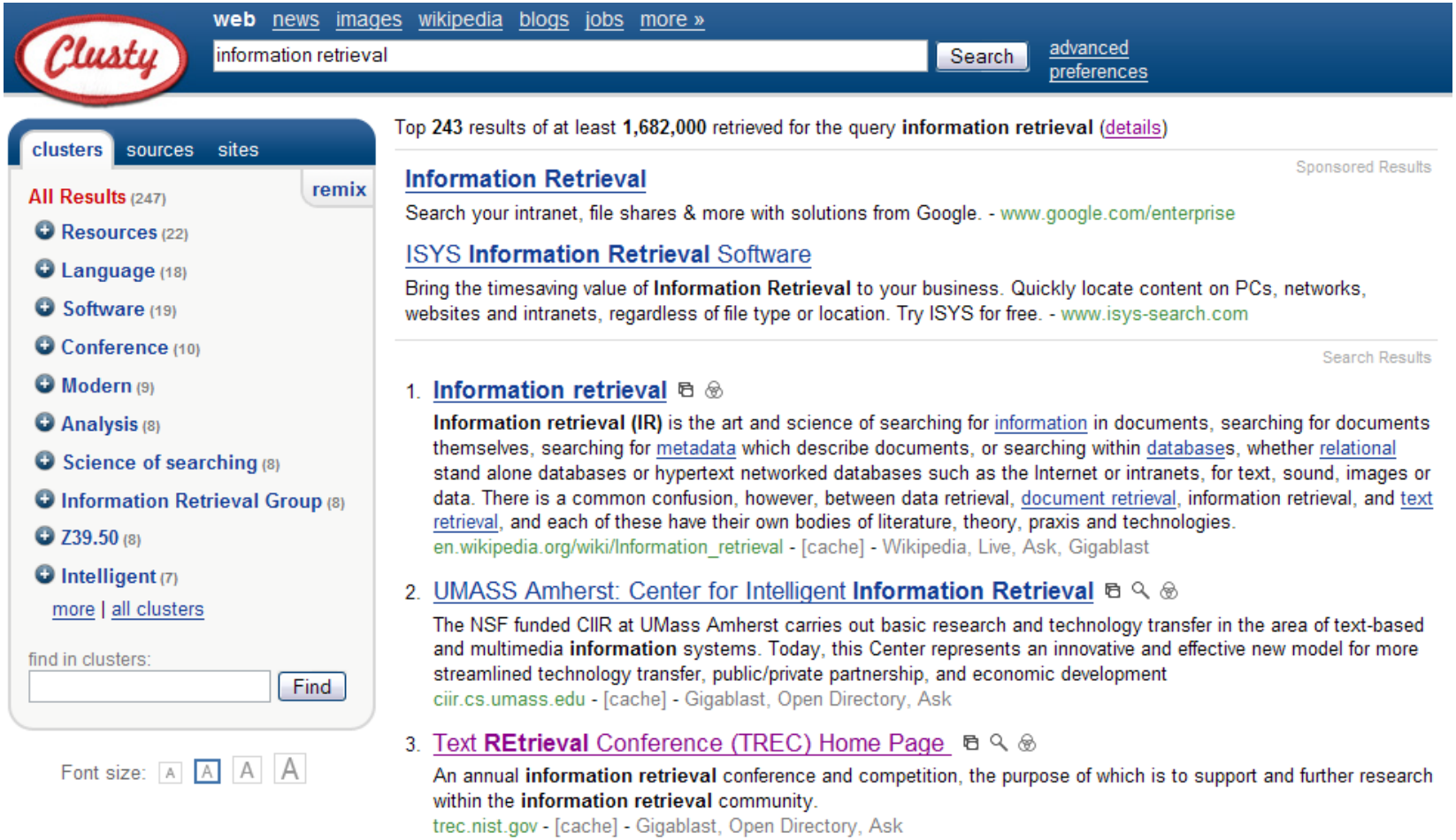
Example: query on “star”

Encyclopedia text

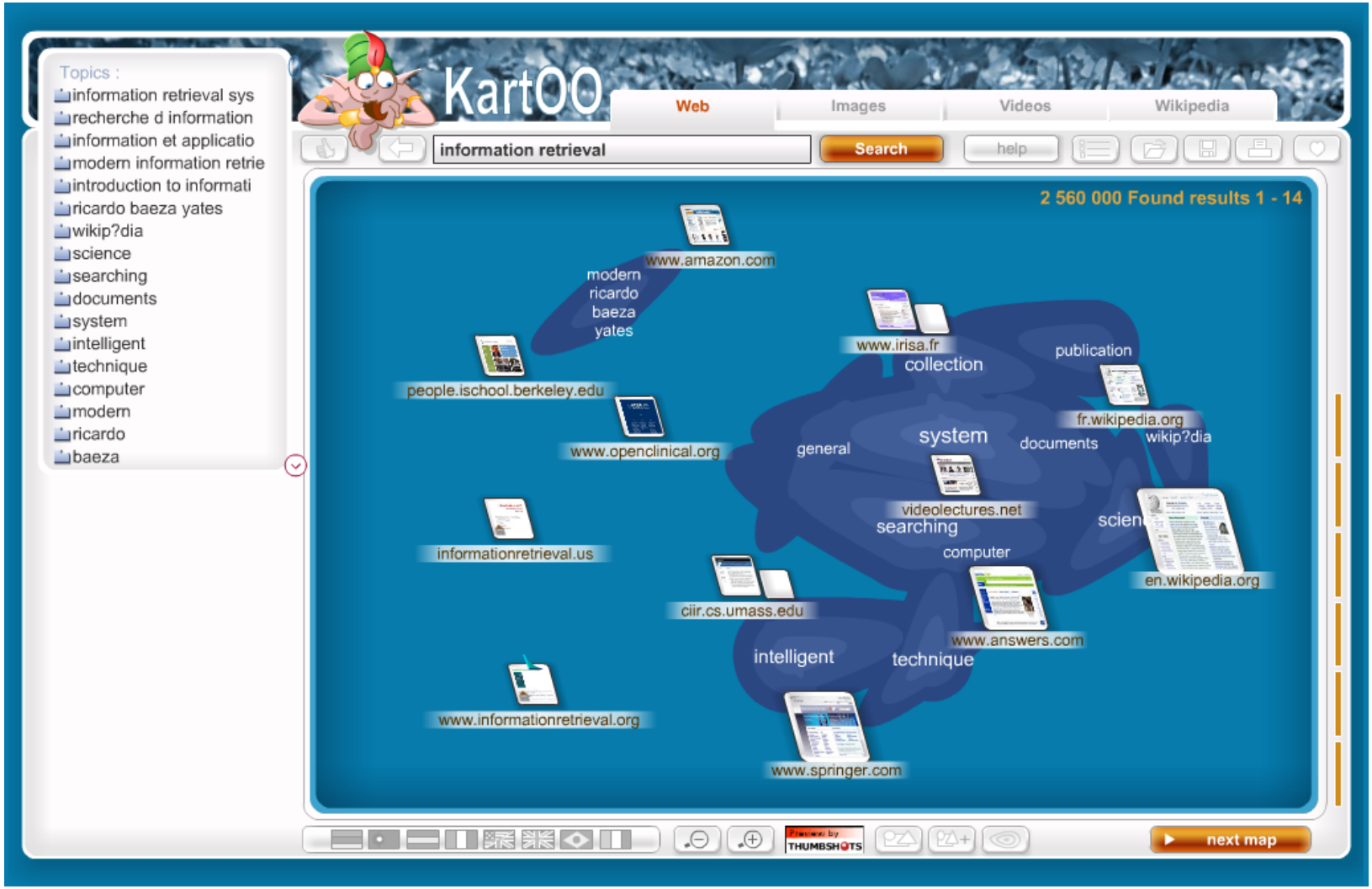


Clustering and re-clustering is entirely automated

Vivisimo's Clustering Results

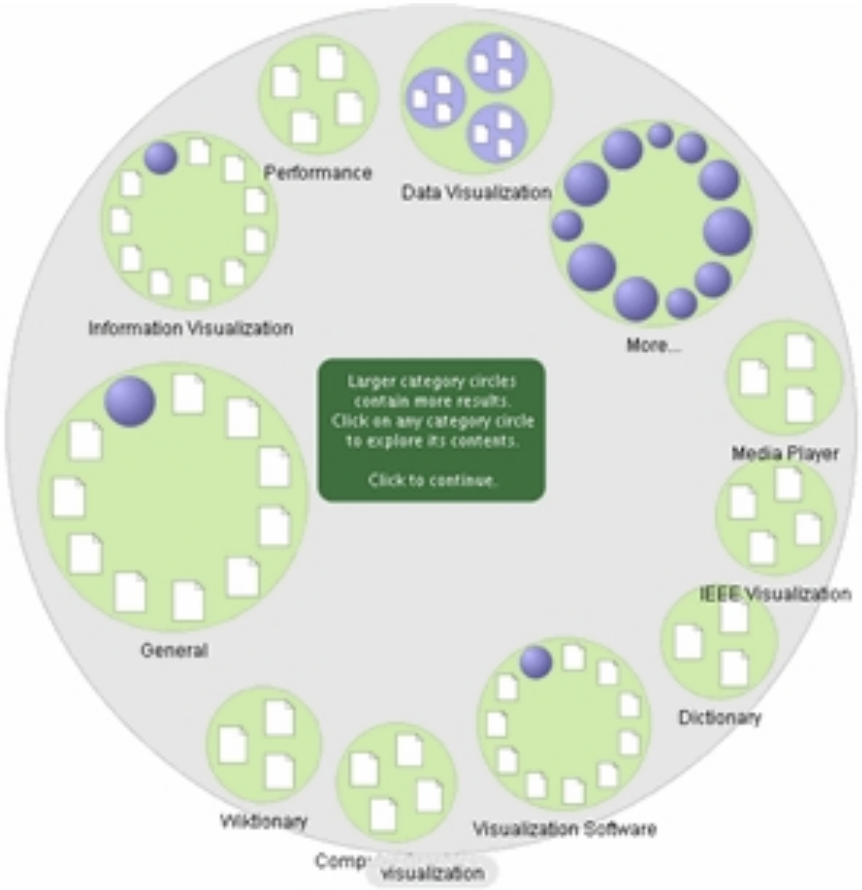


Kartoo's Cluster Visualization



Grokker's Clustering Visualization

- clusters its results and presents them in a unique circular map
- www.grokker.com no longer work now



Visualizing Named Entities in Results

NameSieve pittsburgh Search

Carnegie Museum of Art

Matching 1000 records | Pages 1 2 3 4 5 6 7 8 9 10 Next» | Note selection

1 **CMA-59085**

Inscription on Kodak Royal Pan film box reads "May - 56." Charlene Foggie Barnett identified location (visit 7/16/2008). Published in the Pittsburgh Courier Newspaper, March 31, 1956, page A1.
[Group portrait including, left to right, guest speaker and director of the Washington D.C. NAACP office, Clarence Mitchell, Pittsburgh Mayor David L. Lawrence, executive secretary of the Pittsburgh NAACP affiliate, Marion Bond Jordon, Pittsburgh NAACP president, Rev. Charles H. Foggie, Pittsburgh councilman and chairman of the dinner, Paul F. Jones, and Pittsburgh NAACP executive board member, John Golightly, gathered for the Human Rights Dinner at Gateway Plaza, Downtown]

2 **CMA-56466**

Published in Pittsburgh Courier Newspaper, September 1954, page 1. Inscription on Kodak Super Panchro Press film box reads "Courier Aug 54"
[Pittsburgh Mayor David L. Lawrence signing the proclamation declaring September 5 - 10 Urban League Week, seated with him are Pittsburgh Urban League board members, Jessie Vann, and Henry Pearson, standing over are, Pittsburgh Urban League secretary, Mrs. Nathaniel Dandridge, and Councilman and Pittsburgh Urban League president, Patrick T. Fagan, gathered in the Office of the Mayor at the City County Building, Downtown]

3 **CMA-60163**

Inscription on Kodak Royal Pan film box reads "May 57". Published in the Pittsburgh Courier Newspaper, May 23, 1959, page A15.
[Group portrait of four women from the Pittsburgh Club of Negro Business and Professional Women, including, from left to right, Marion Sappington Bryant, Wilnette B. Price, Florence Allen Holmes, and Blanche Russell Crayton, Pittsburgh Club of the Negro Business and Professional Women, gathered for the Founders Day Breakfast in the Pittsburgh Room at the Penn Sheraton Hotel]

4 **CMA-17297**

Published in the Pittsburgh Courier Newspaper, January 12, 1946, page 1.
[Pittsburgh Mayor David L. Lawrence swearing in Robert E. "Pappy" Williams in Pittsburgh City Council Chambers at the City County Building, Downtown]

5 **CMA-36776**

Note on Pittsburgh Courier paper attached to negatives reads " Pitt Gad. [University of Pittsburgh Graduates], Feb. '52"
[Group portrait of four men wearing dark graduation gowns and mortar boards, possibly University of Pittsburgh graduates]

6 **CMA-36779**

Note on Pittsburgh Courier paper attached to negatives reads " Pitt Gad. [University of Pittsburgh Graduates], Feb. '52"
[Portrait of woman wearing dark graduation gown and mortar board, standing between two plaques,possibly University of Pittsburgh graduate]

7 **CMA-36780**

Note on Pittsburgh Courier paper attached to negatives reads " Pitt Gad. [University of Pittsburgh Gr

Query Terms

pittsburgh (1000)

Named Entities

Who | Where | When | What

A. H. Burchfield » Ani Howell » Aug 54 » Blanche Russell Crayton » Booker » Chairman » Charles H. Foggie » Clarence Mitchell » Councilman » Courier » Courier Newspaper » Dapper Dan » David L. Lawrence » David W. Walker » Delilah Court » Eugene Geller » Executive Board Member » Executive Secretary » Florence Allen Holmes » Garland » Graduates » Guest Speaker » Harry Kenny » Henry Pearson » J. H. Campbell » James Hamlett » John Golightly » Leonard Wright » Man » Marion Sappington

Total 0 NE selected

Apply Filter | Clear Filters

Task Model Notes

Note something to put it here

http://ir.exp.sis.pitt.edu/namesieve_cma/

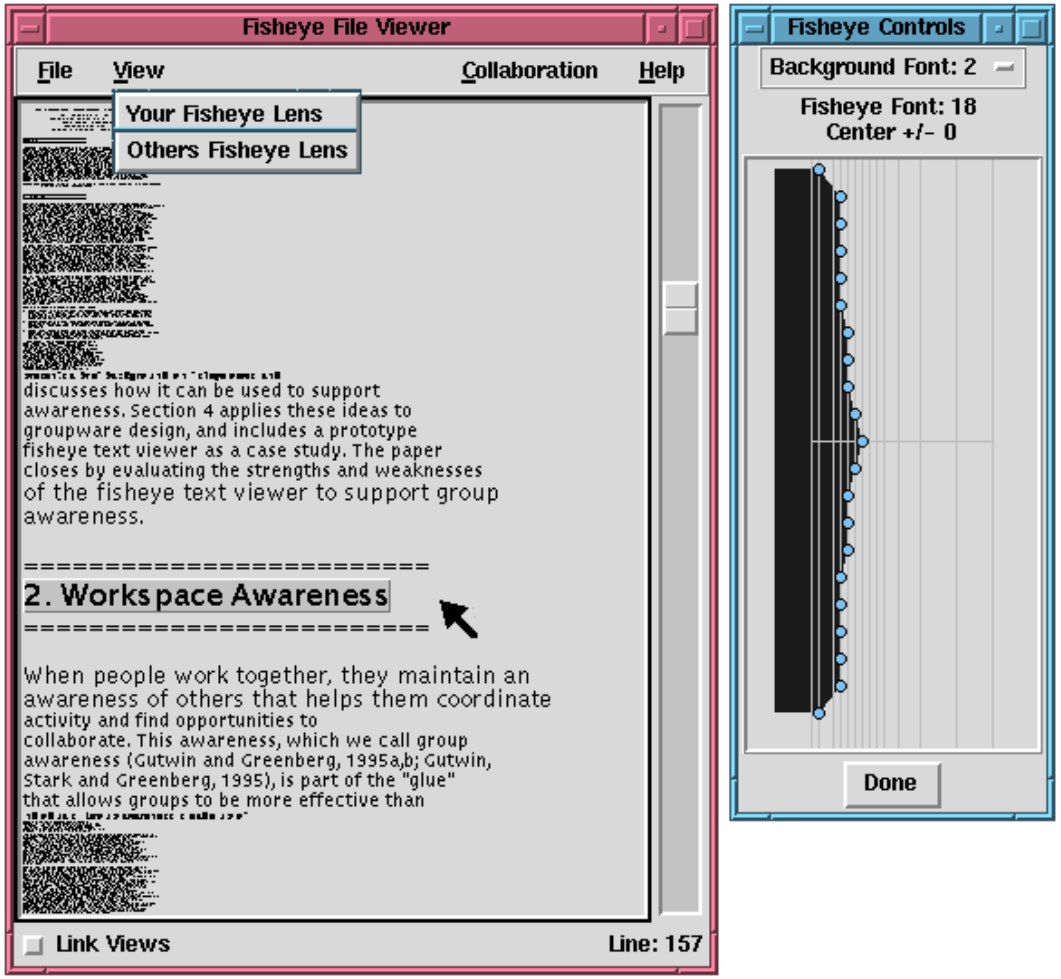
Result Examination Interfaces

- Present result documents in details
  - Text window to show the content
  - KWIC in documents
- To help presenting long documents
  - "Best passage" function helps users get started
    - Overlapping 300 word passages work well
  - Often visualization techniques are used

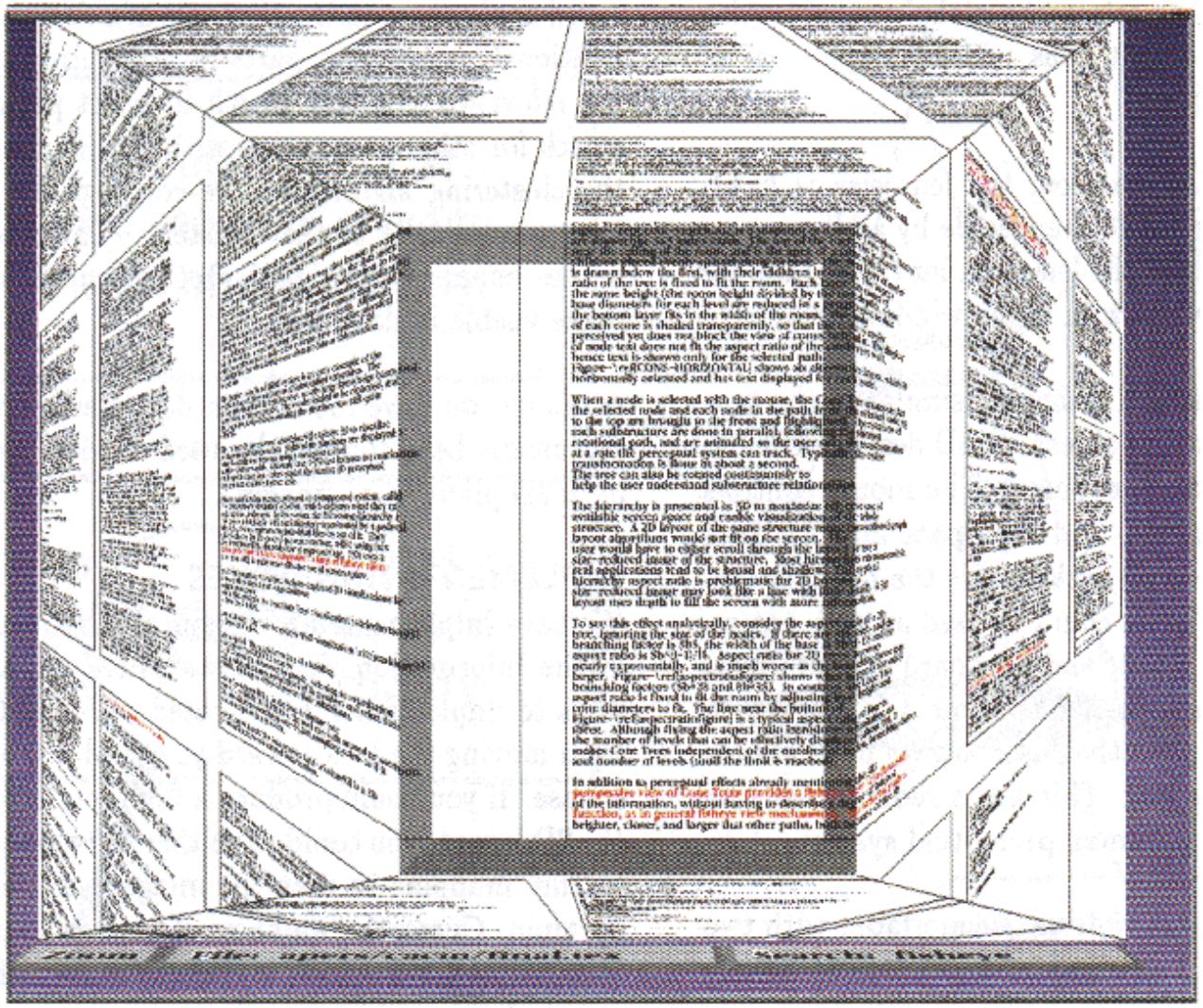
Overview + Details in ICDL

http://www.icdlbooks.org/

Fisheye Views of Documents



The Document Lens



http://doi.acm.org/10.1145/168642.168652

http://www.scils.rutgers.edu/~aspoerri/Teaching/InfoVisResources/images/visiontothink/infosphere/document_robertson/index.htm

Mobile Search Interface

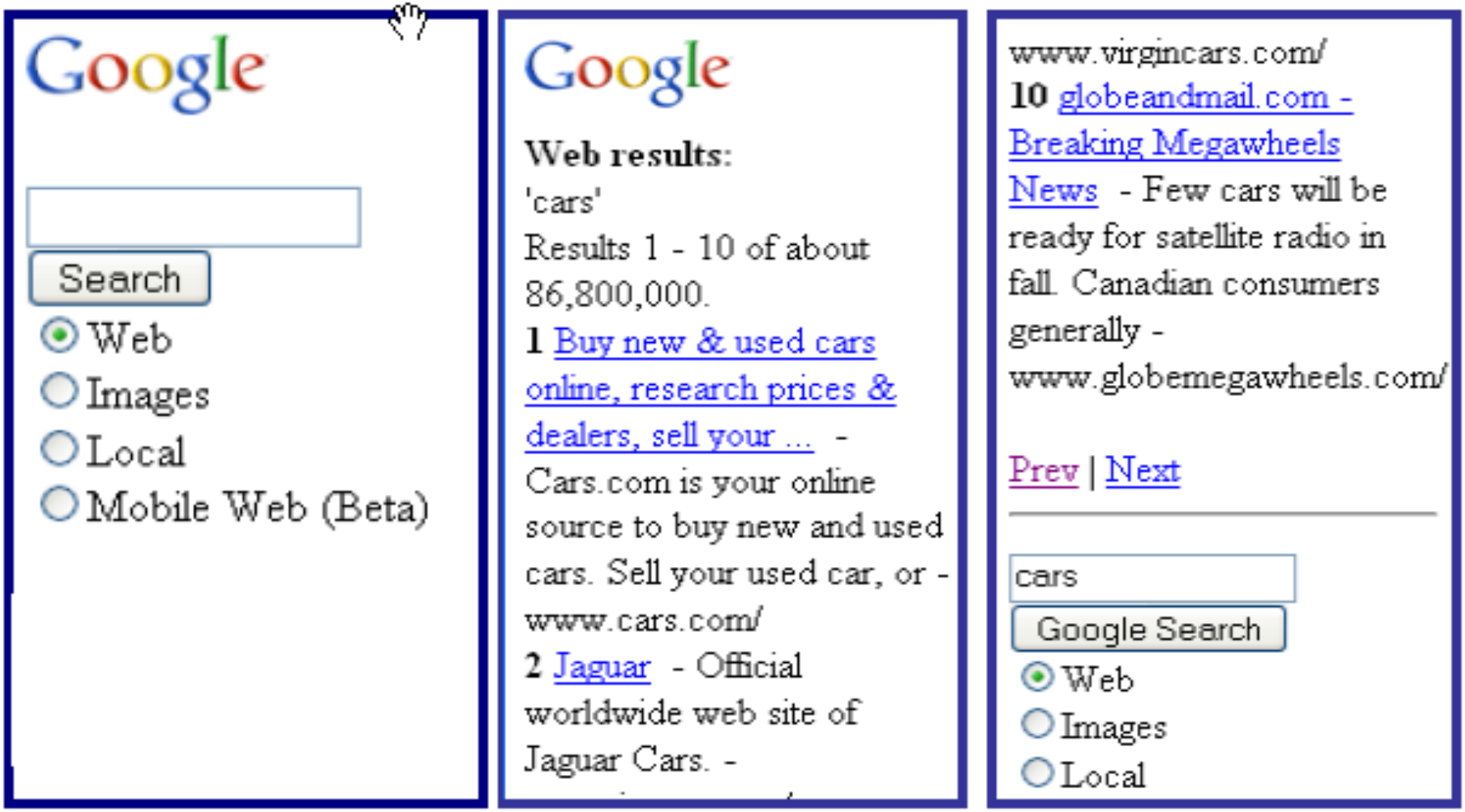
Mobile search interfaces

- Increasingly popular
  - Predict by 2020, most people will use mobile system to access internet
- Just emerge
  - Only in the past 4-5 years, relative large, high resolution screen become popular, such as iphone, ipad, etc.
- A testbed for new technologies
  - Android mobile operating system
  - Siri speech recognition system

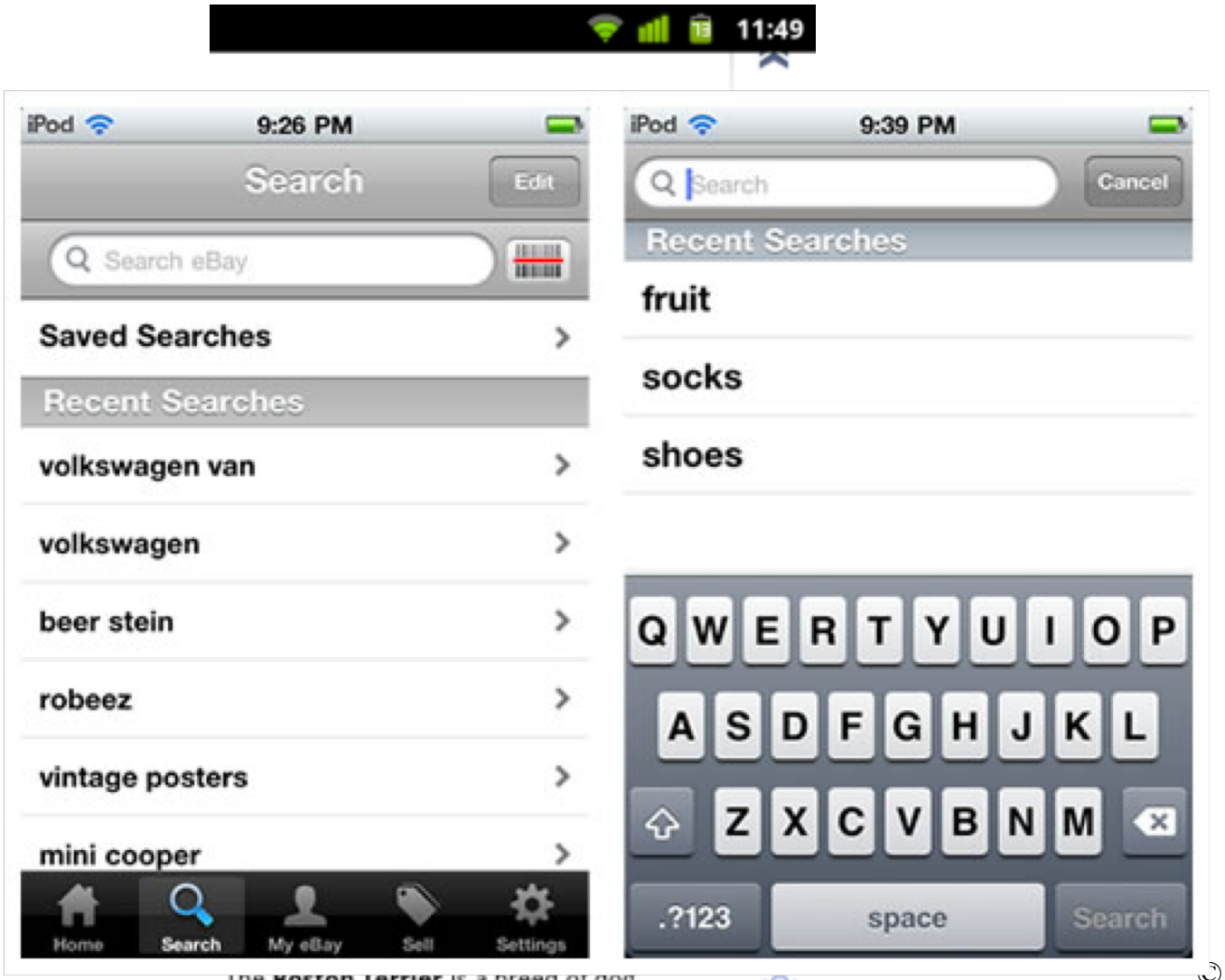
Mobile Search Usages

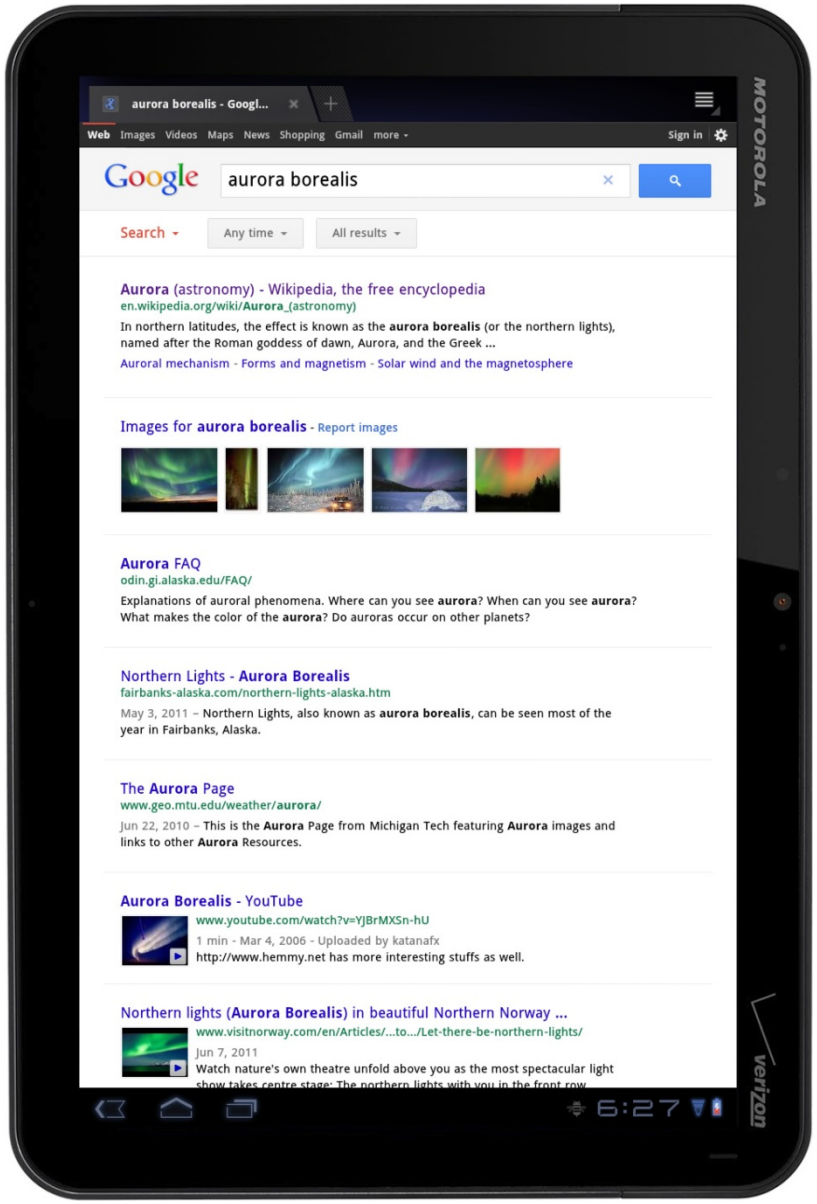
- Unique information needs
  - Information: Temporal and geolocation information can be the focus
  - Needs: route-finding, planning, related to current location
    - Answer questions that came up in current activity, location, time and conversation
- More focused, more repeated queries
  - Most frequent 1000 mobile queries account 22% of all mobile queries
    - In contrast, most frequent 1000 desktop queries account 6%
- Shallower interactions with results
  - Fewer page views per query
  - But more rely on current context

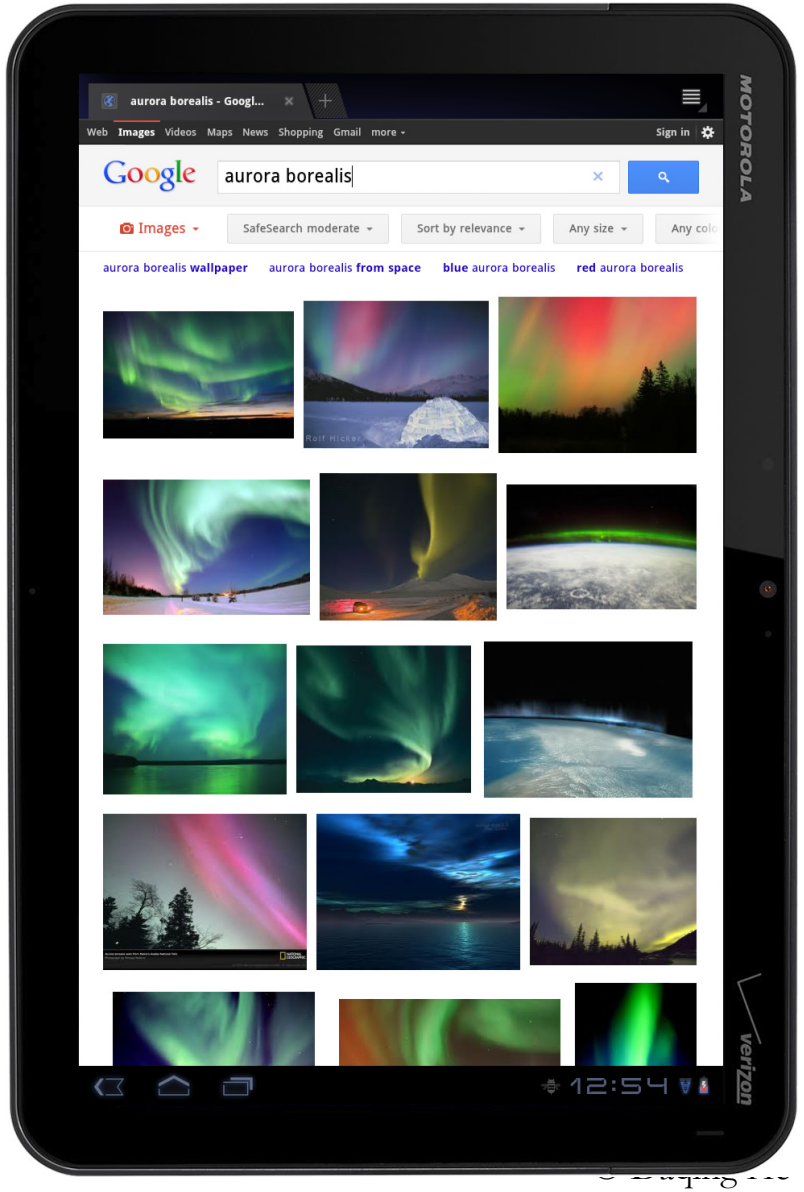
Old Google Interface



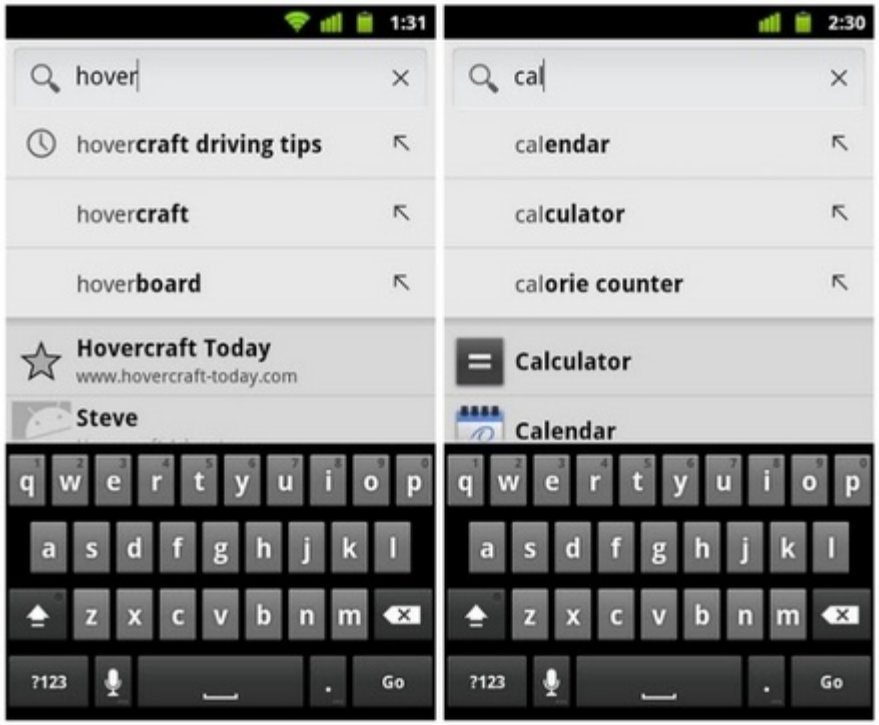
Google Search Interface

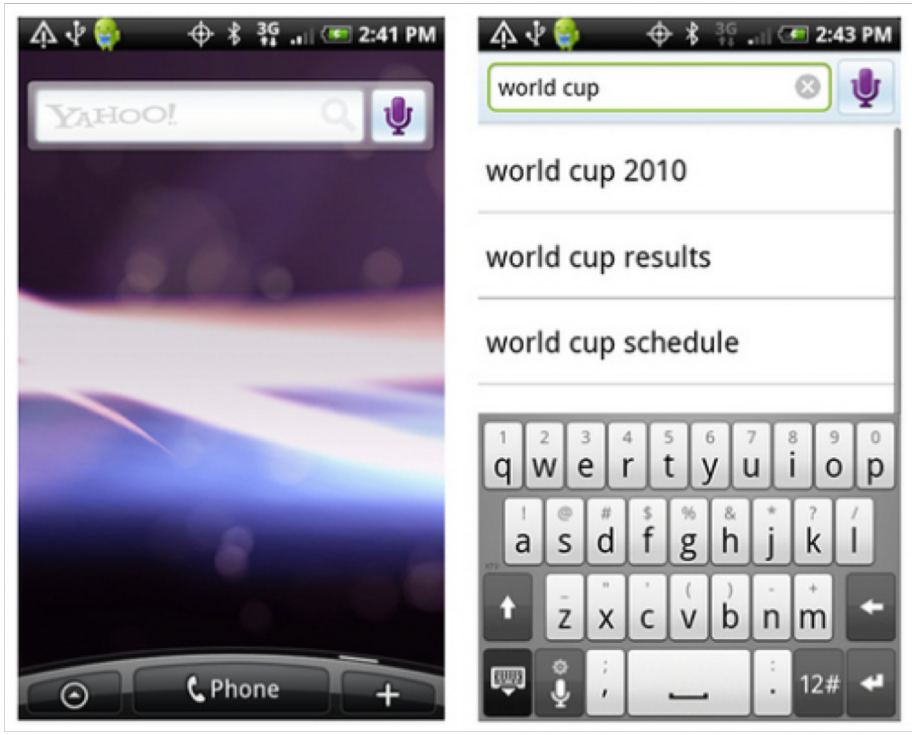



aurora borealis - Googl...
Web Images Videos Maps News Shopping Gmail more
Sign in
Google
aurora borealis
Search
Any time
All results
Aurora (astronomy) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Aurora_(astronomy)
In northern latitudes, the effect is known as the aurora borealis (or the northern lights), named after the Roman goddess of dawn, Aurora, and the Greek ...
Auroral mechanism - Forms and magnetism - Solar wind and the magnetosphere
Images for aurora borealis - Report images
Aurora FAQ
odin.gi.alaska.edu/FAQ/
Explanations of auroral phenomena. Where can you see aurora? When can you see aurora? What makes the color of the aurora? Do auroras occur on other planets?
Northern Lights - Aurora Borealis
fairbanks-alaska.com/northern-lights-alaska.htm
May 3, 2011 – Northern Lights, also known as aurora borealis, can be seen most of the year in Fairbanks, Alaska.
The Aurora Page
www.geo.mtu.edu/weather/aurora/
Jun 22, 2010 – This is the Aurora Page from Michigan Tech featuring Aurora images and links to other Aurora Resources.
Aurora Borealis - YouTube
www.youtube.com/watch?v=YJBrMXSn-hU
1 min - Mar 4, 2006 - Uploaded by katanafx
http://www.hemmy.net has more interesting stuffs as well.
Northern lights (Aurora Borealis) in beautiful Northern Norway ...
www.visitnorway.com/en/Articles/...to.../Let-there-be-northern-lights/
Jun 7, 2011
Watch nature's own theatre unfold above you as the most spectacular light
MOTOROLA
verizon
6:27

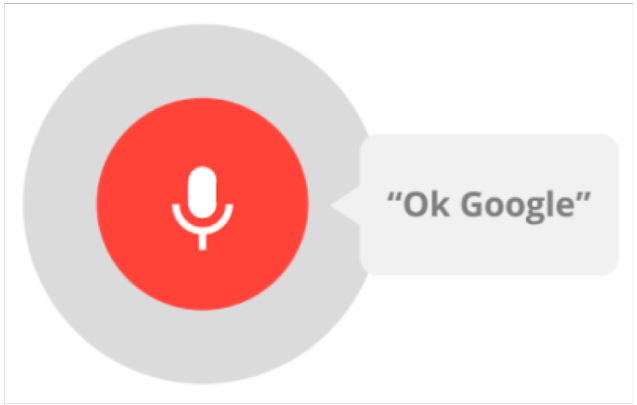

aurora borealis - Googl...
Web Images Videos Maps News Shopping Gmail more
Sign in
Google
aurora borealis
Images
SafeSearch moderate
Sort by relevance
Any size
aurora borealis wallpaper
aurora borealis from space
blue aurora borealis
red aurora borealis
MOTOROLA
verizon
12:54

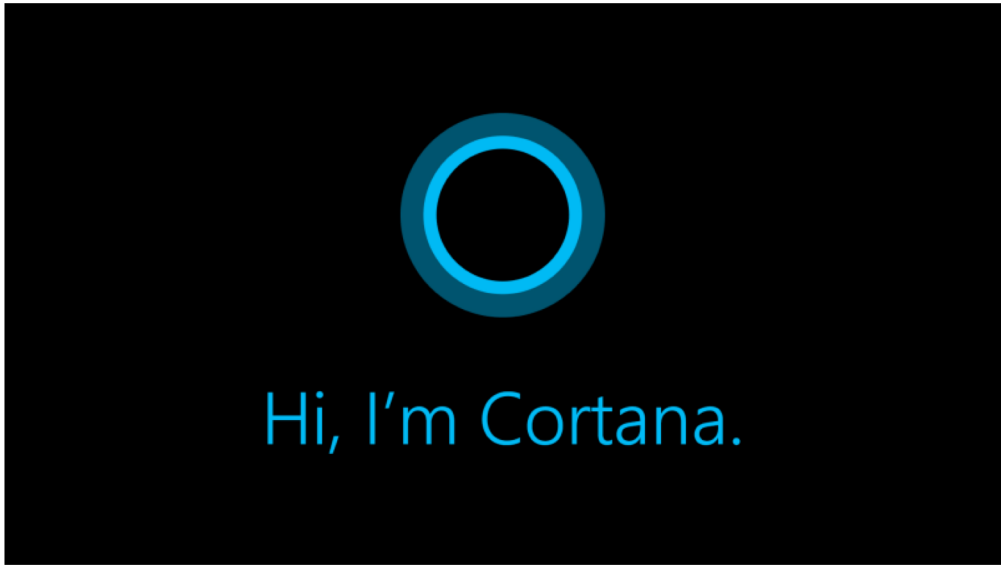
Query Auto-Completion



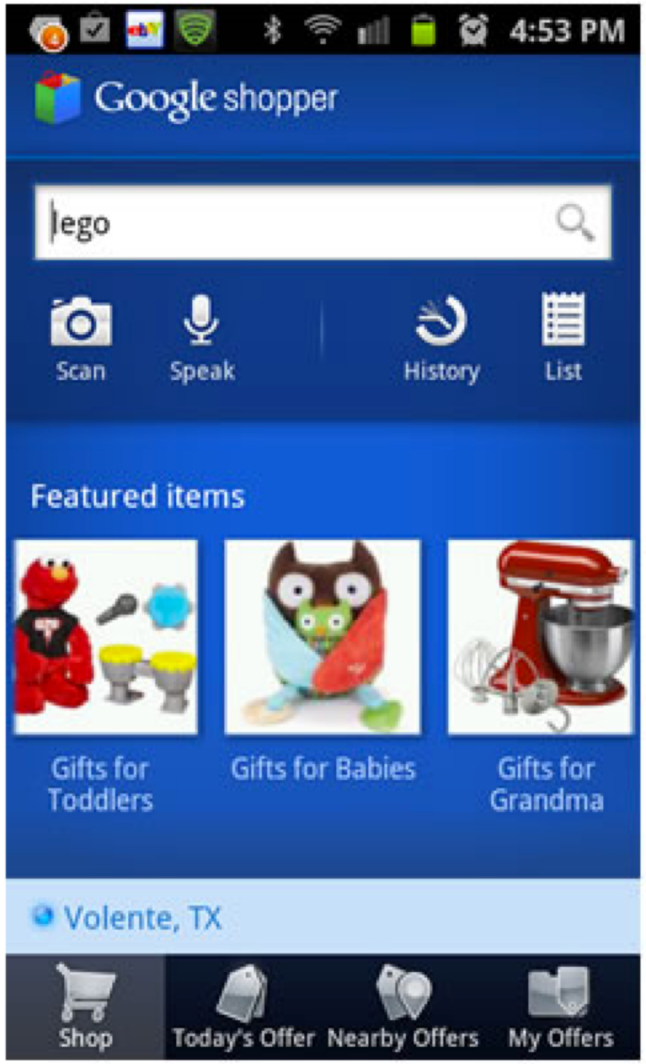


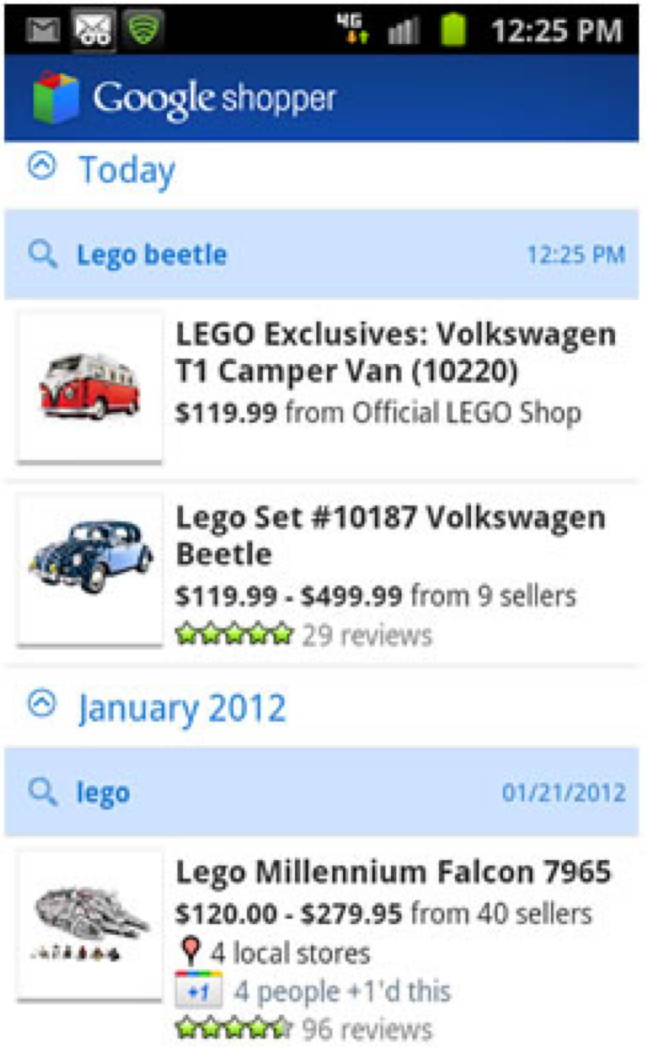
Speech based Search Input



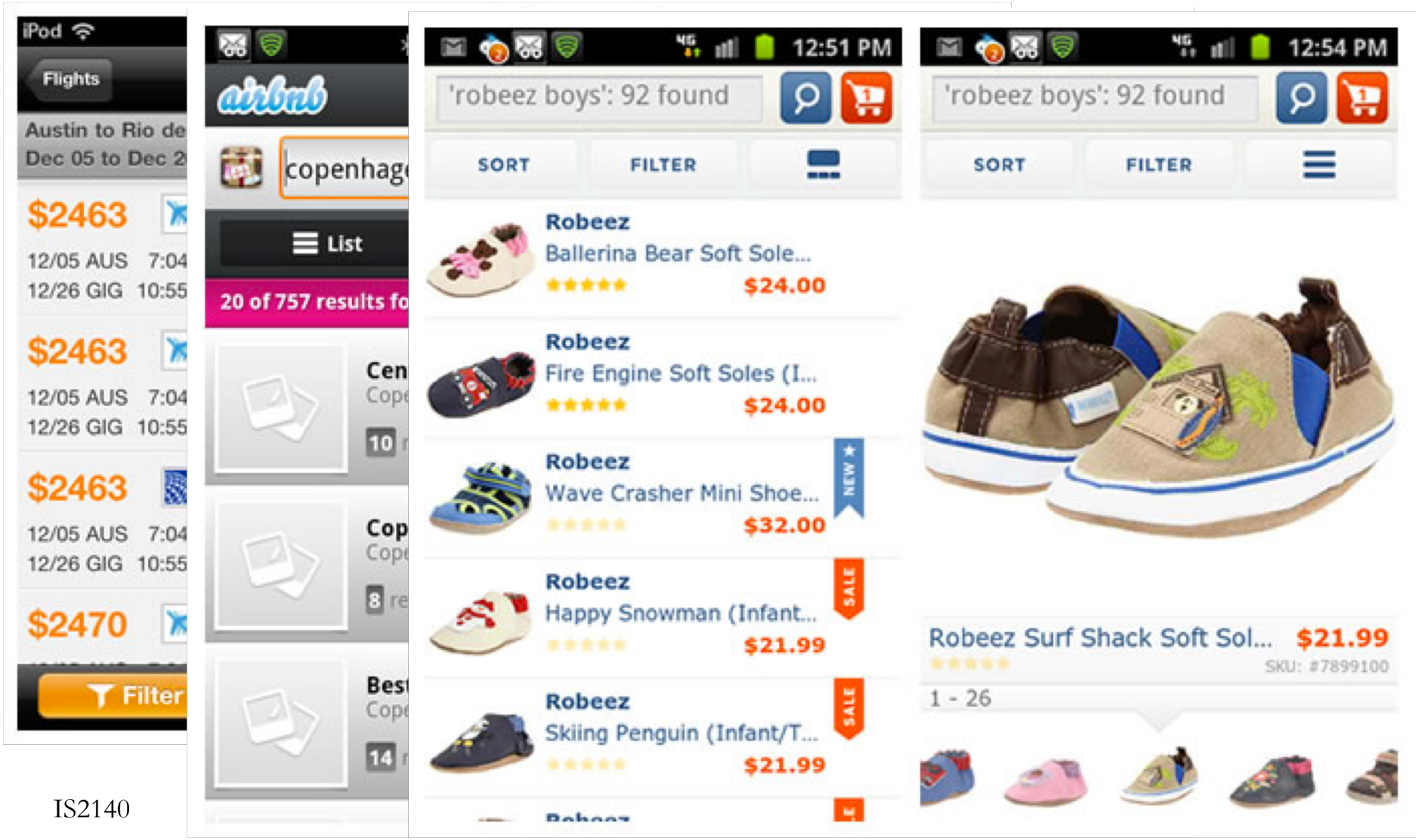


Context-based Search: Location

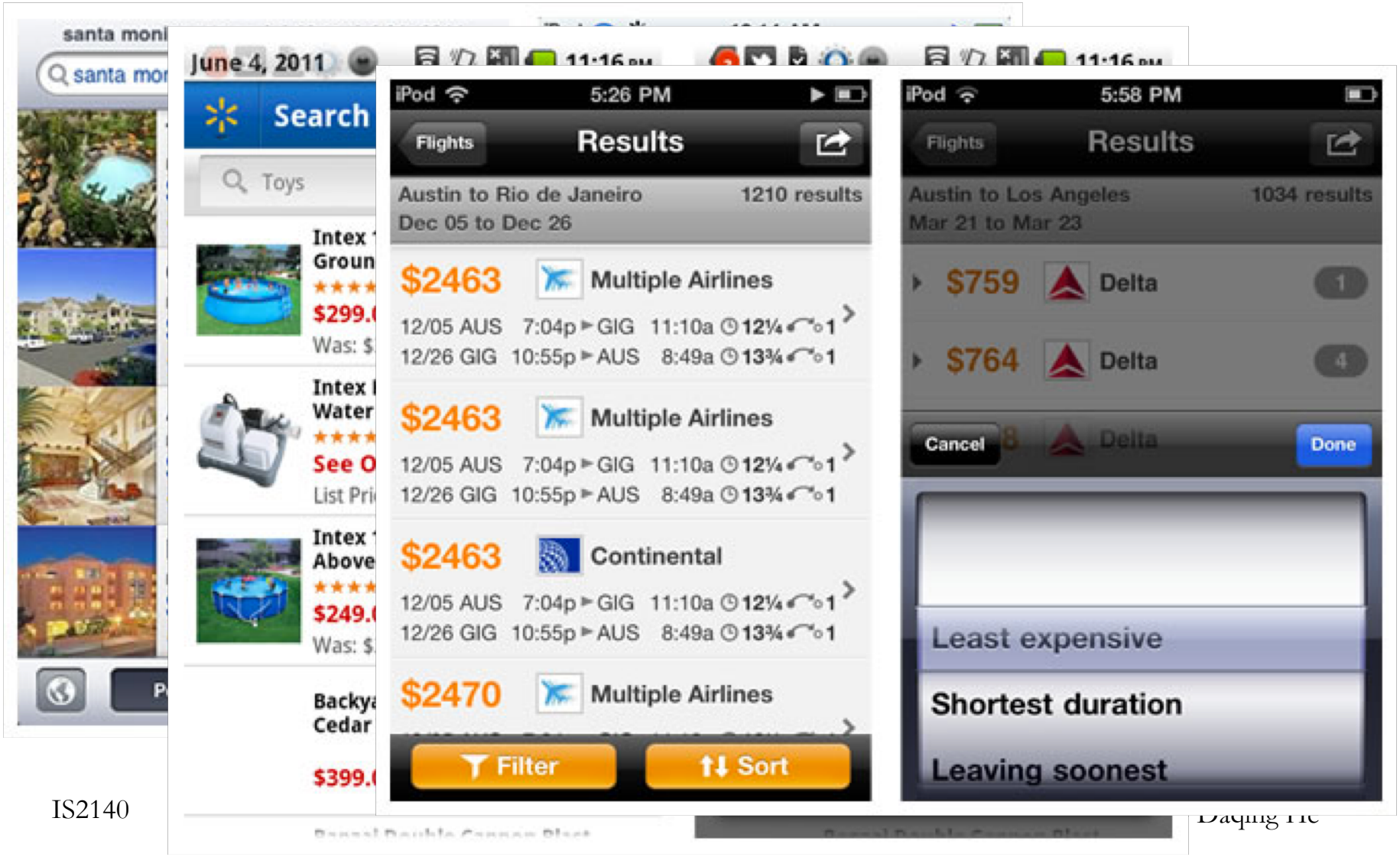




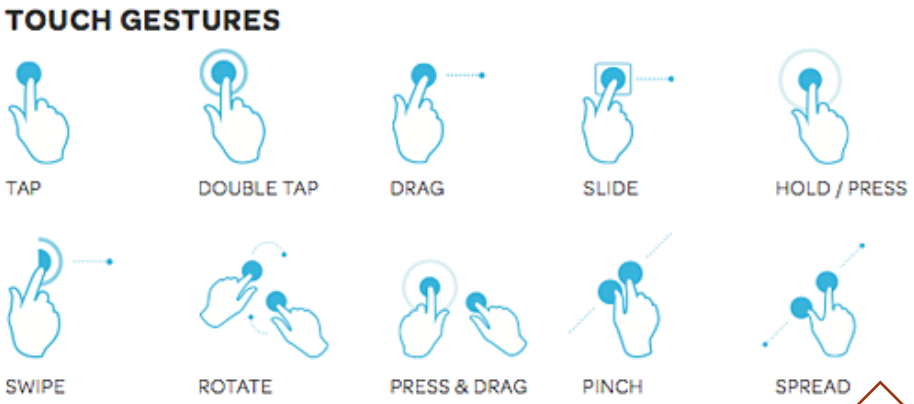
Search Results

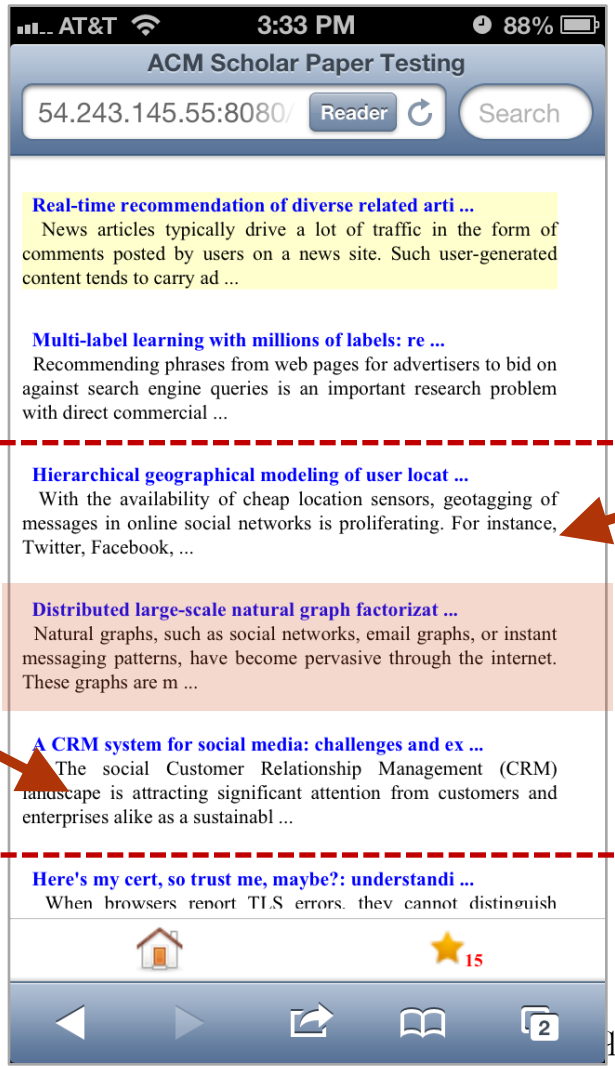


Sorting Results



Mobile Touch Interactions





Further Topics on Mobile Search

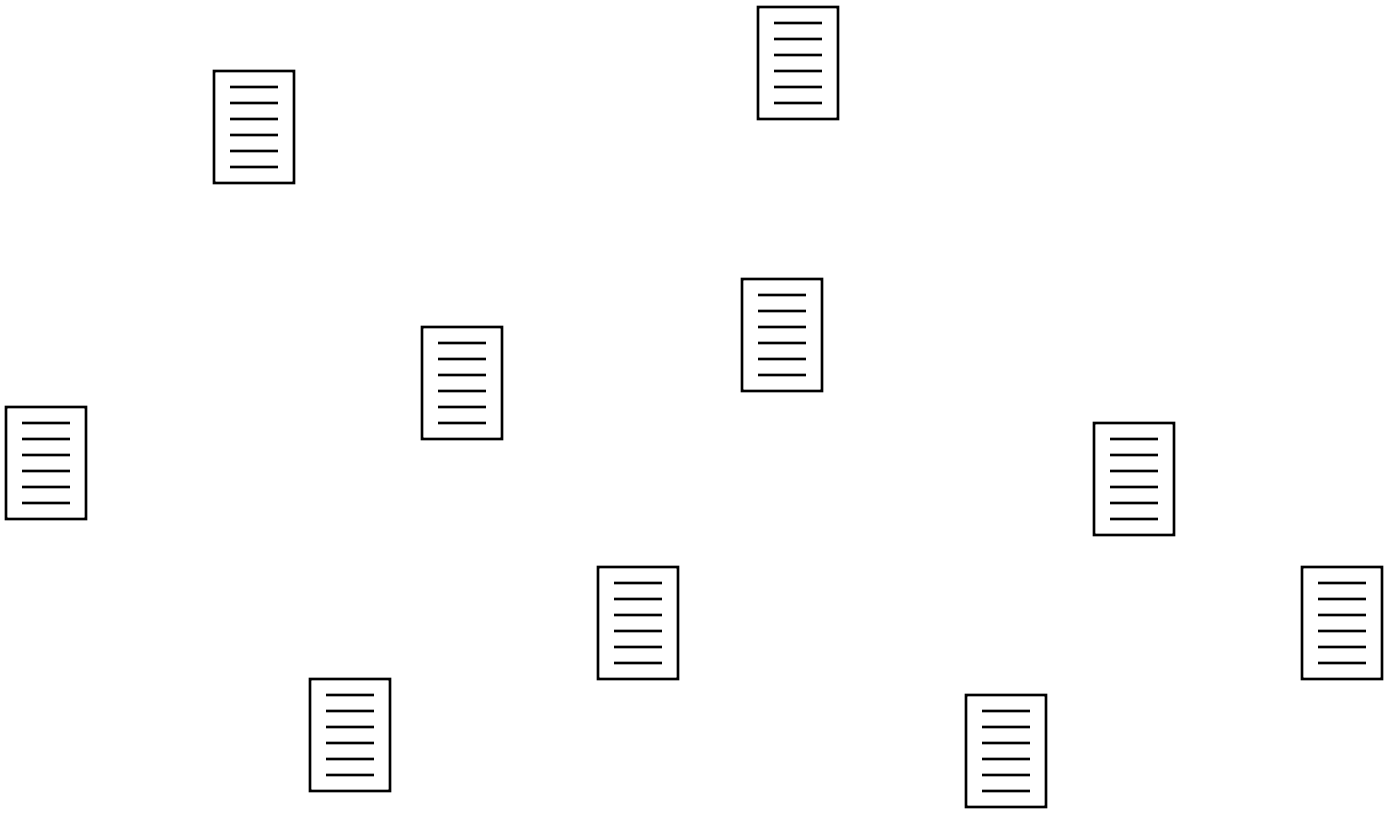
- Mobile information seeking behaviors
  - Small screen/keyboard size make it difficult to type keywords. Studies found that people did not issue shorter queries, why ?
  - Spent much more time on reading a web page
  - Clicked more high-quality documents
- Cross-device web search
  - People often migrated tasks from one device to another
- Utilizing unique information in mobile web search
  - Contextual information
  - Mobile touch interactions

Synergy between Human and IR system

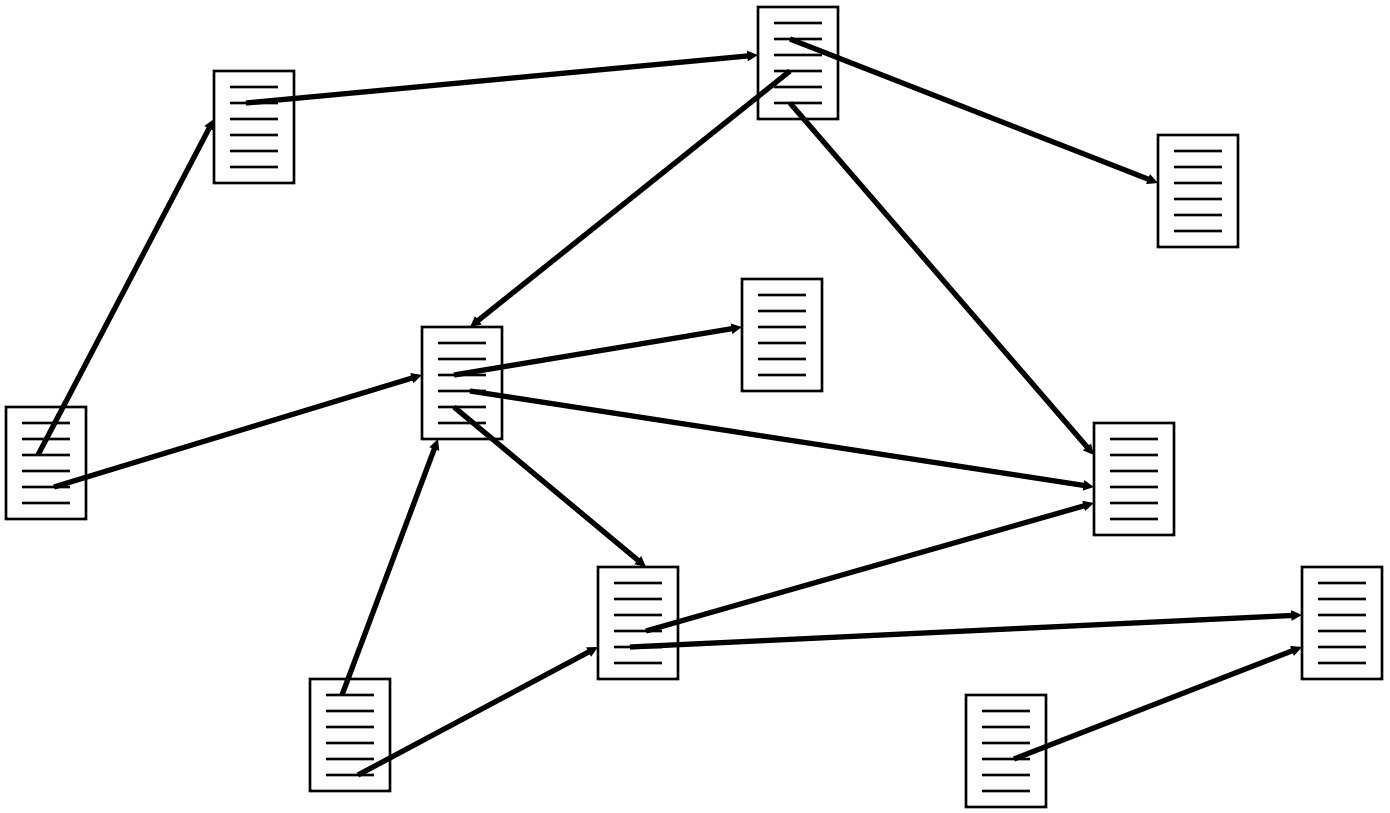
- The strength of one covers the weakness of the other
  - Humans do things that human are good at
  - Computers do things that computers are good at
- What human good at?
  - Sense low level stimuli, Recognize patterns, Reason inductively, Communicate with multiple channels, Apply multiple strategies, Adapt to changes or unexpected events
- What machine good at?
  - Sense stimuli outside human's range, Calculate fast and mechanical, Store large quantities and recall accurately, Response rapidly and consistently, Perform repetitive actions reliably, Maintain performance under heavy load and extended time

Web and Web Search

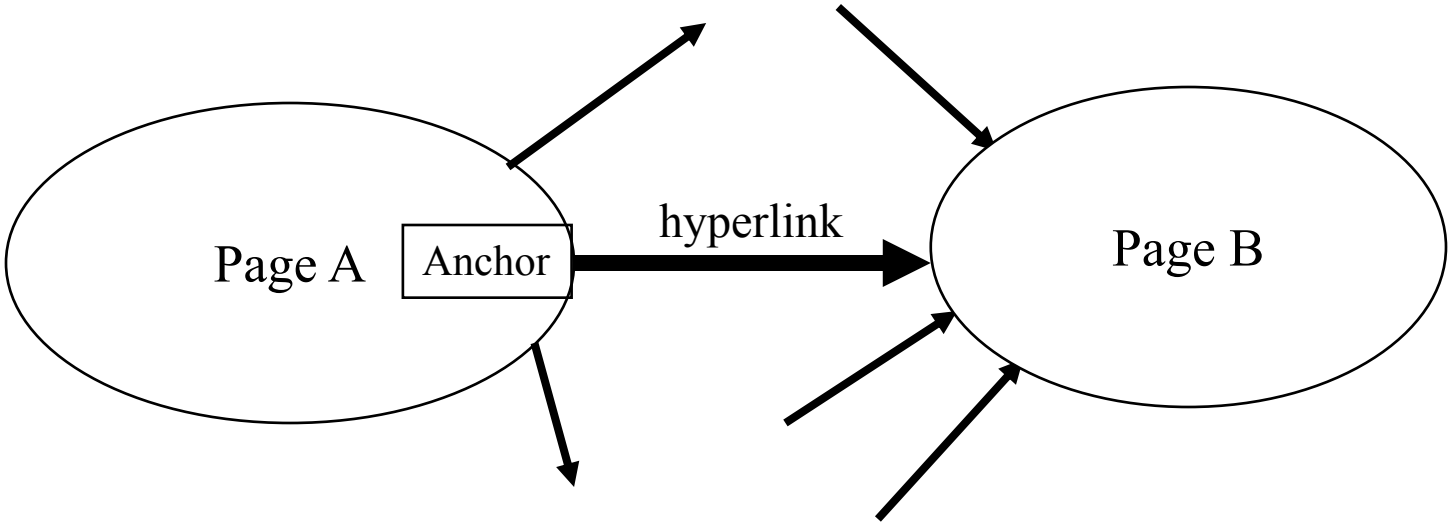
Web of Interconnected Pages



Web of Interconnected Pages



The Web as a Directed Graph



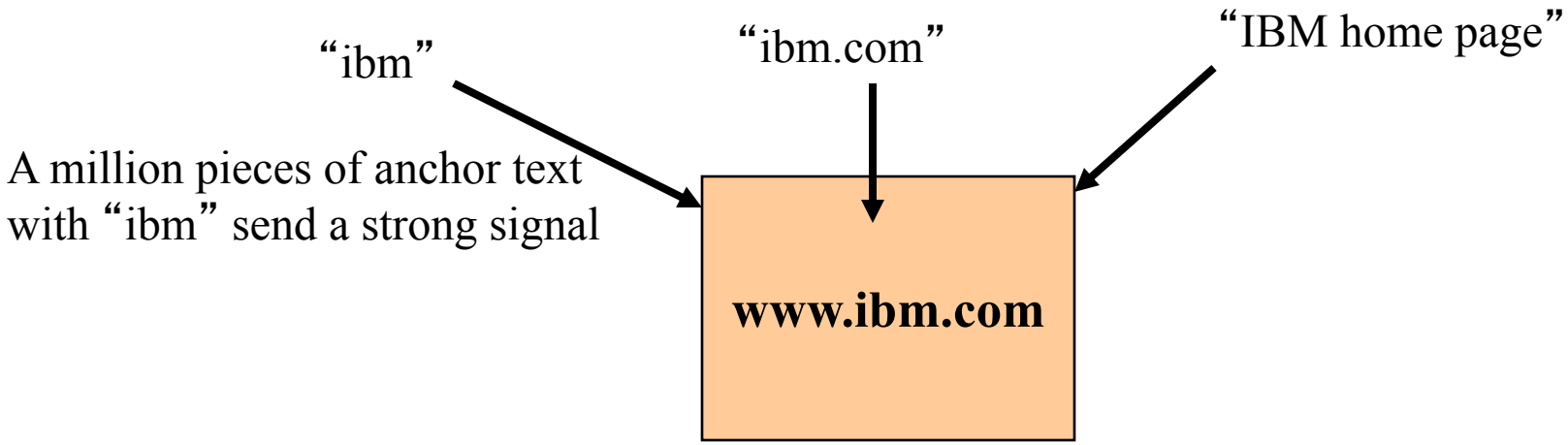
Hypothesis 1: A hyperlink between pages denotes a conferral of authority (quality signal)

Hypothesis 2: The text in the anchor of the hyperlink on page A describes the target page B. In Feb, 1997, Yanhong Li (Scotch Plains, NJ) filed a hyperlink based search patent. The method uses words in anchor text of hyperlinks.

Anchor Text

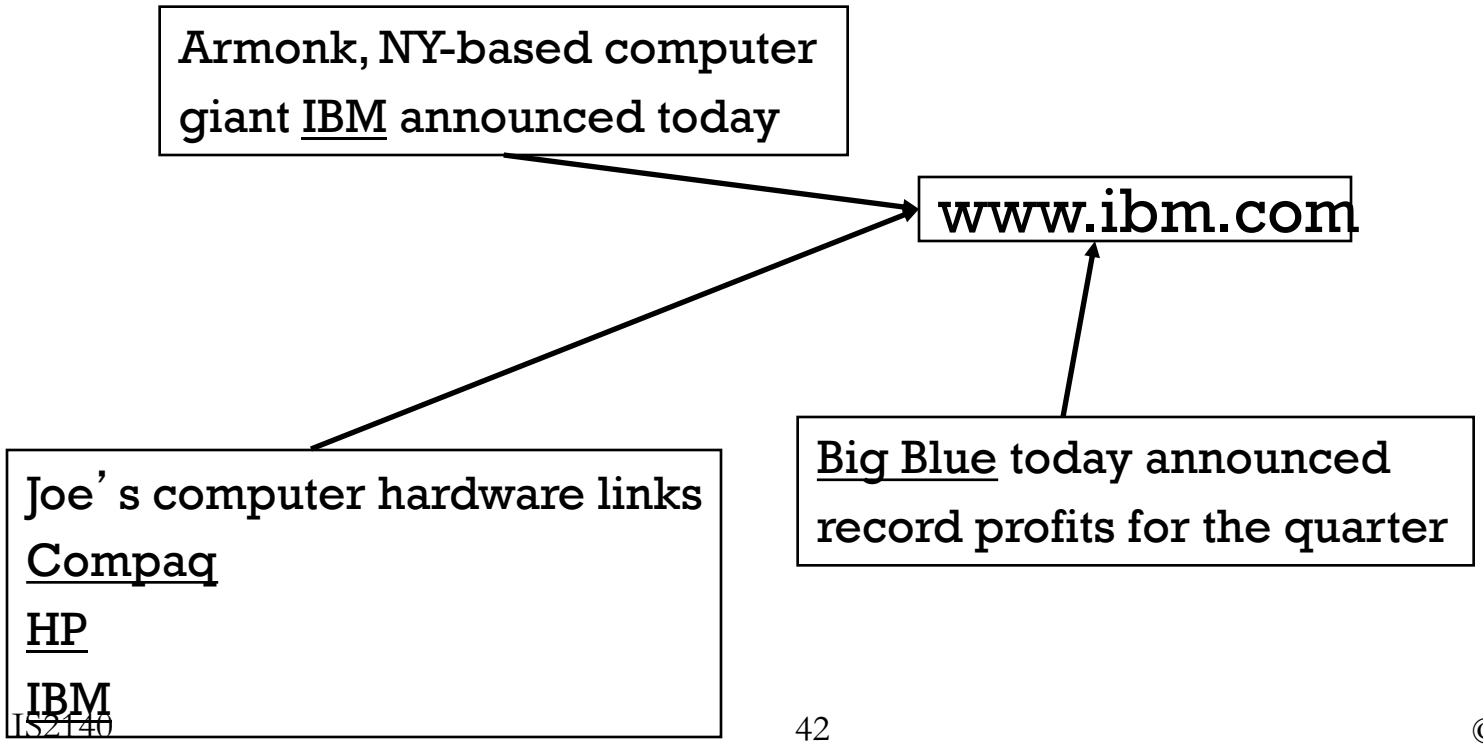
WWW Worm - McBryan [Mcbr94]

- For ***ibm*** how to distinguish between:
  - IBM's home page (mostly graphical)
  - IBM's copyright page (high term freq. for 'ibm')
  - Rival's spam page (arbitrarily high term freq.)



Indexing anchor text

- When indexing a document *D*, include anchor text from links pointing to *D*.



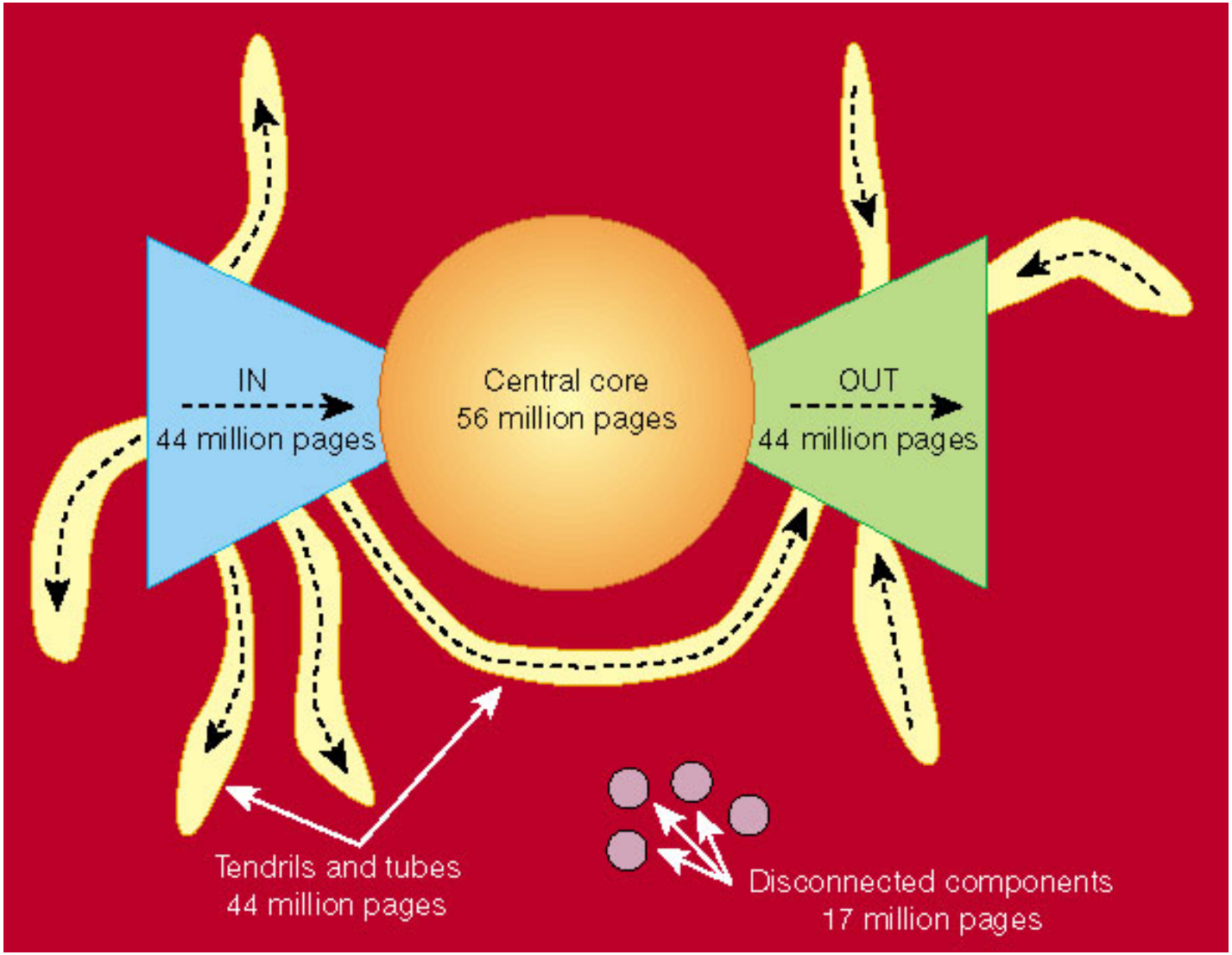
Indexing anchor text

- Can sometimes have unexpected side effects - *e.g.*, ***french military victories***
- Helps when descriptive text in destination page is embedded in image logos rather than in accessible text.
- Many times anchor text is not useful:
  - "click here"
- Increases content more for popular pages with many in-coming links, increasing recall of these pages.
- May even give higher weights to tokens from anchor text.

Nature of the Web

- Over one billion pages by 1999
  - Growing at 25% per month!
  - Google indexed about 3 billion pages in 2003
- Over 12 billion surface Web pages by 2005
  - Google indexed about 8.1 billion pages in 2005
- Unstable
  - Changing at 1% per week
- Redundant
  - 30-40% (near) duplicates
    - e.g., unix man page tree

Bowtie Theory of the Web

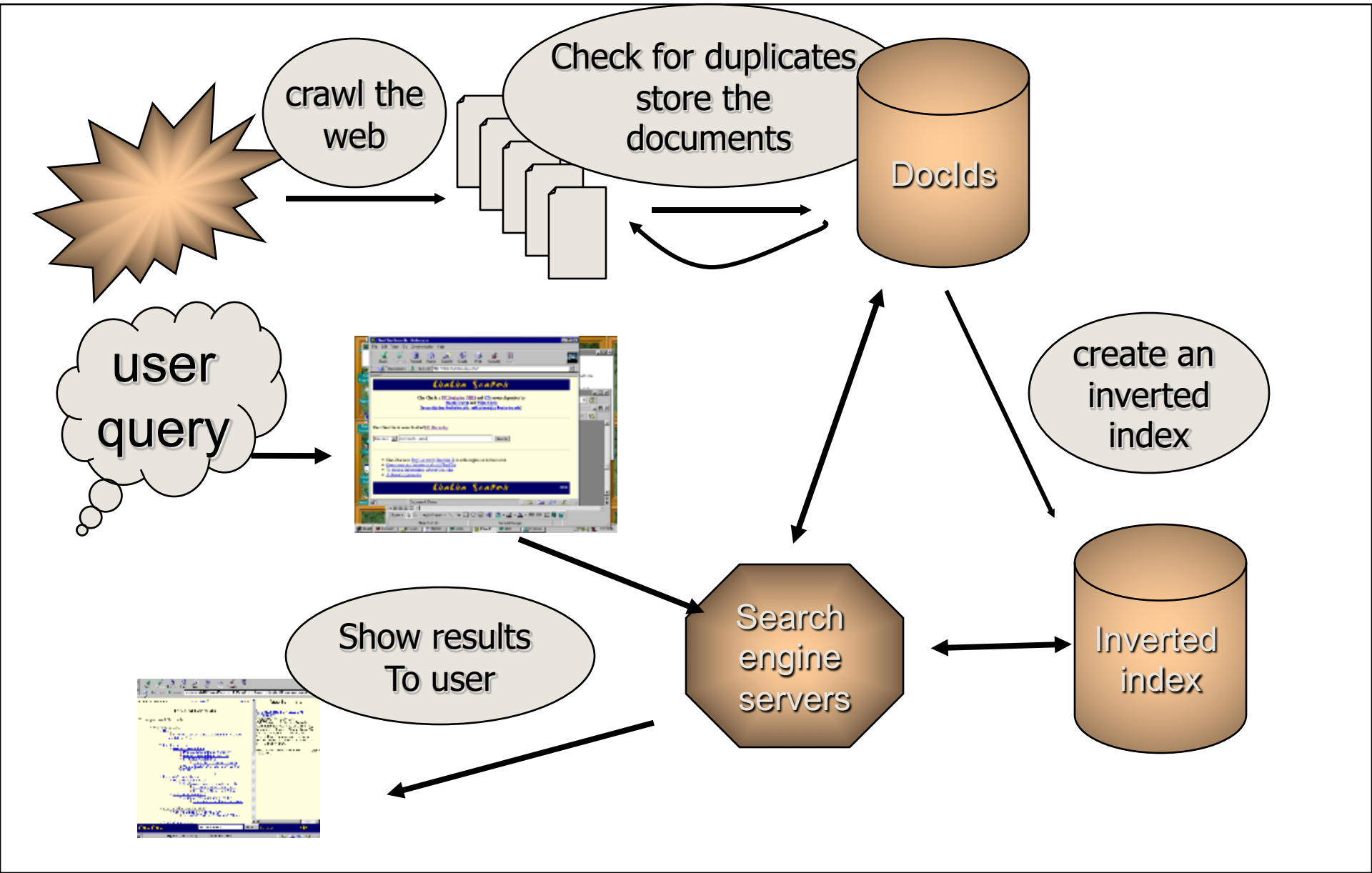


http://www.almaden.ibm.com/cs/k53/www9.final/

Web Searches

- Comparing to traditional search, Web search tasks are the same at
  - Match queries against a collection of documents
  - Help people find relevant information
  - Content matching is still very important
- Comparing to traditional search, Web search tasks are the different at
  - Users
    - Information needs
    - Search skills and behaviors
  - Collection
    - Location of documents: scattered, distributed, changing
    - Nature of documents: hypertext, multimedia, duplications, quality/reputation

Standard Web Search Engine Architecture



Inverted Indexes for Web Search Engines

- Inverted indexes are still used, even though the web is so huge.
- Some systems partition the indexes across different machines. Each machine handles different parts of the data.
- Other systems duplicate the data across many machines; queries are distributed among the machines.
- Most do a combination of these.

Web Search Engines

- Over 3,200 search engines in 2002
- Lots of consolidation recently
  - Overture bought altavista and alltheweb.com
  - Yahoo bought inktomi and overture
- Five major search engines in 2002
  - Google (29.5%), Yahoo (28.9%), MSN (27.6%), AOL (18.4%), Ask (9.9%)

Top Search Engines for 2010

2010	Google	Yahoo!	Bing	Ask	AOL Search	Total
2010-08-28	71.59%	14.28%	9.87%	2.28%	1.21%	99.23%
2010-07-31	71.43%	14.43%	9.86%	2.32%	1.19%	99.23%
2010-06-26	71.65%	14.37%	9.85%	2.19%	1.15%	99.21%
2010-05-22	72.00%	14.58%	9.20%	2.18%	1.06%	99.02%
2010-05-08	71.56%	14.79%	9.31%	2.27%	1.07%	99.00%
2010-03-06	71.07%	14.46%	9.55%	3.01%	0.98%	99.07%
2010-02-06	71.35%	14.60%	9.56%	2.55%	1.06%	99.12%
2010-01-02	72.25%	14.83%	8.91%	2.53%	0.77%	99.29%

http://www.seoconsultants.com/search-engines/

Web Query Logs Studies

- Why?
  - Directly and unobtrusively observing Web searchers is difficult, especially in large scale
- Web query logs
  - An electronic record of interactions that have occurred between a user and a system (Web search engine)

1748	237ACEDD326E2B74	pepsi
2138	237ACEDD326E2B74	PEPSI
2200	237ACEDD326E2B74	PEPSI
2421	237ACEDD326E2B74	PEPSI
2725	*237ACEDD326E2B74*	*NBA.COM*

Web Query Log –An Example

[10/09 06:39:25] Query: holiday decorations [1-10]
[10/09 06:39:35] Query: [web]holiday decorations [11-20]
[10/09 06:39:54] Query: [web]holiday decorations [21-30]
[10/09 06:39:59] Click: [webresult][q=holiday decorations][21]
http://www.stretcher.com/stories/99/991129b.cfm
[10/09 06:40:45] Query: [web]halloween decorations [1-10]
[10/09 06:41:17] Query: [web]home made halloween decorations [1-10]
[10/09 06:41:31] Click: [webresult][q=home made halloween decorations][6]
http://www.rats2u.com/halloween/halloween_crafts.htm
[10/09 06:52:18] Click: [webresult][q=home made halloween decorations][8]
http://www.rpmwebworx.com/halloweenhouse/index.html
[10/09 06:53:01] Query: [web]home made halloween decorations [11-20]
[10/09 06:53:30] Click: [webresult][q=home made halloween decorations][20]
http://www.halloween-magazine.com/

collected on October 9, 2000 for 24 hours from excite.com users who accepted cookies.

Available Web Logs

Search engine	Excite	Excite	AlltheWeb	Excite	AlltheWeb	AltaVista
Data collection	Tuesday, 16 sep 1997	Wednesday, 1 dec, 1999	Tuesday, 6 feb 2001	Monday, 30 april 2001	Tuesday 28 may 2002	Sunday, 8 sep 2002
Sessions	210,590	325,711	153,297	262,025	345,093	369,350
Queries	545,206	1,025,910	451,551	1,025,910	957,703	1,073,388
Terms	1,224,245	1,500,500	1,350,619	1,538,120	2,225,141	3,132,106

These logs can be obtained from
http://ist.psu.edu/faculty_pages/jjansen/academic/transaction_logs.html

Some More Recent Logs

- **Sogou Query Logs (2008) [44M]**: Queries from a Chinese search engine http://www.sogou.com/labs/dl/q.html

- **Yandex Query Logs (2009) [341M]**: From the most popular Russian search engine. http://switchdetect.yandex.ru/en/datasets

- **MSN Query Logs (2006 and 2007) [14M and 100M]**: http://research.microsoft.com/en-us/um/people/nickcr/wscd09/

Web Crawler

collaborative filtering

Search

Advanced Search
Preferences

Web Scholar

Results **1** - **10** of about **622,000** for **collaborative filtering**. (**0.07** seconds)

Collaborative filtering - Wikipedia, the free encyclopedia - 12:02pm
Collaborative filtering (CF) is the process of **filtering** for information or patterns using techniques involving collaboration among multiple agents, ...
en.wikipedia.org/wiki/**Collaborative_filtering** - 53k - Cached - Similar pages - Note this

Collaborative Filtering
Collaborative filtering systems can produce personal recommendations by computing the similarity between your preference and the one of other people.
pespmc1.vub.ac.be/collfilt.html - 11k - Cached - Similar pages - Note this

Collaborative Filtering Research Papers
User-updated directory of **collaborative filtering** research papers with abstracts , links to the full papers, and reader comments.
jamesthornton.com/cf/ - 100k - Cached - Similar pages - Note this

Using **collaborative filtering** to weave an information tapestry
Dhruv Gupta , Mark Digiovanni , Hiro Narita , Ken Goldberg, Jester 2.0 (poster abstract): evaluation of an new linear time **collaborative filtering** algorithm ...
portal.acm.org/citation.cfm?id=138867 - Similar pages - Note this
by D Goldberg - 1992 - Cited by 1072 - Related articles - All 4 versions

[PPT] **Collaborative Filtering**: A Tutorial
File Format: Microsoft Powerpoint - View as HTML
Empirical Analysis of Predictive Algorithms for **Collaborative Filtering** Breese, ... vs **collaborative filtering**, recommendation is based on properties of the ...
www.cs.cmu.edu/~wcohen/collab-**filtering**-tutorial.ppt - Similar pages - Note this

Sponsored Links

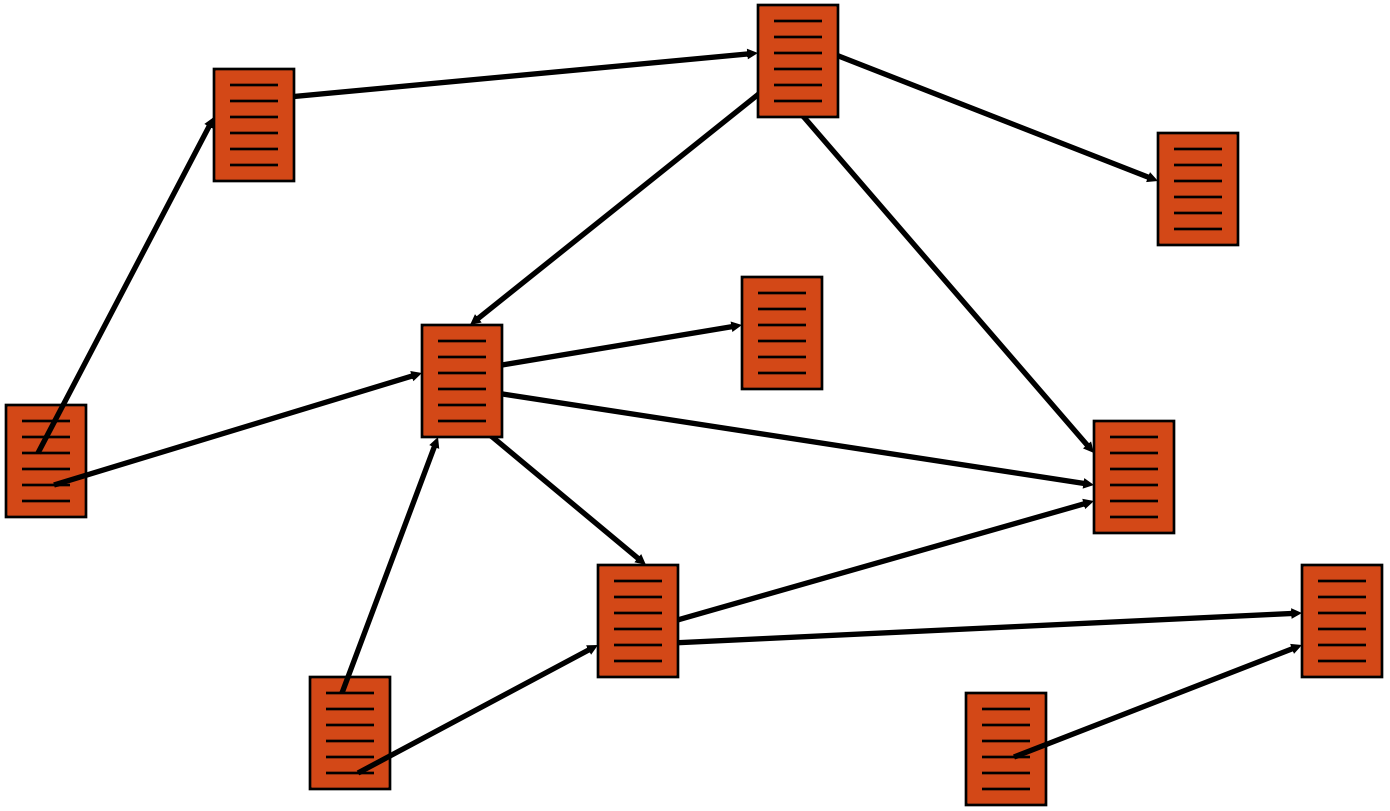
Recommendation Engine
Show relevant products/content with online discovery. Find out how...
www.aggregateknowledge.com

Neptuny ContentWise
The first recommendation engine for IPTV and WebTV
www.neptuny.com

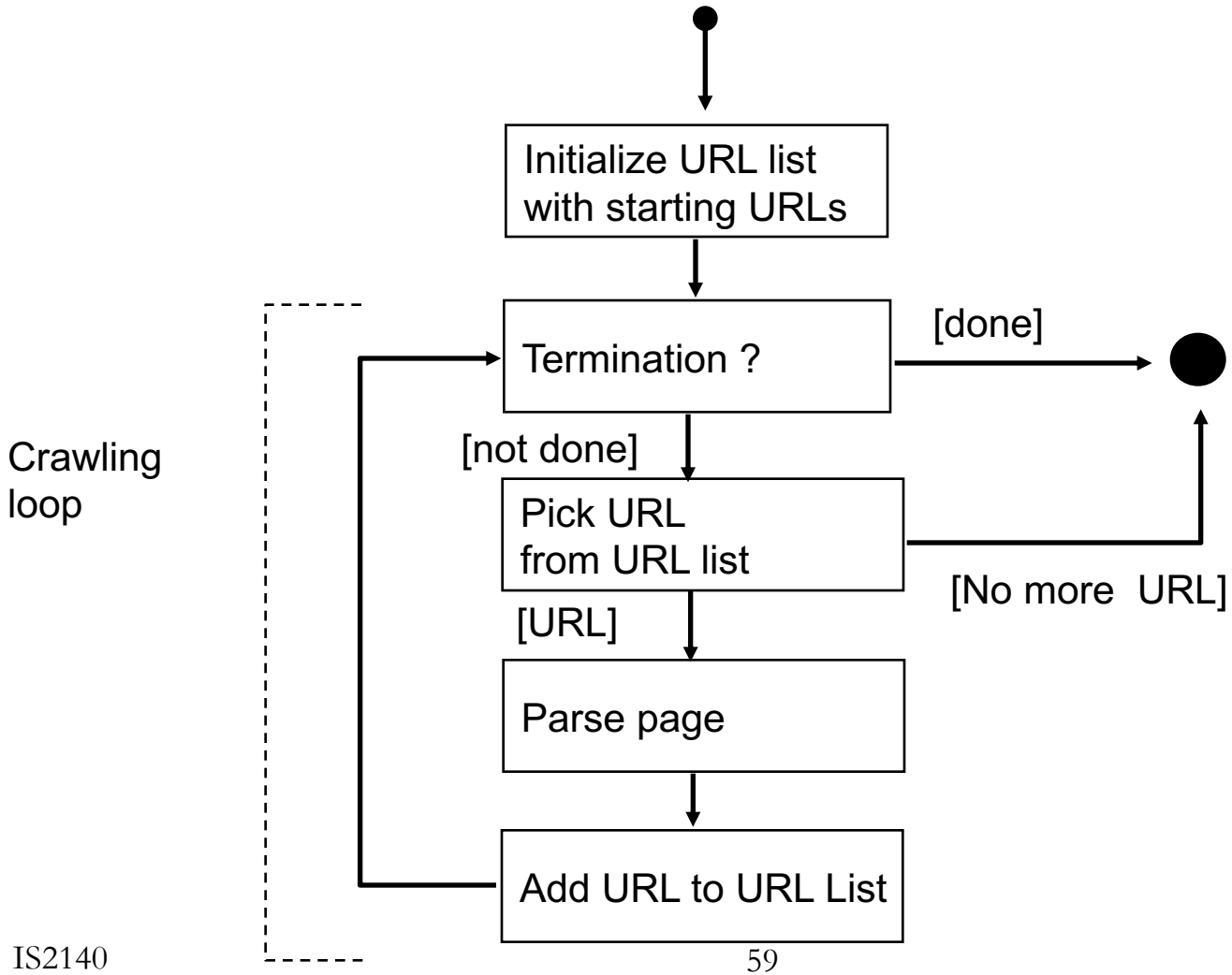
Crawling the Web

- Web search engines needs a way to collect all the documents for indexing ➔ Crawler
  - **Also know as web spider, web robot, web scutter**
- Main idea:
  - 1. Use known sites as the seeds or starting points
  - 2. download information from these sites
  - 3. Follow the links from each site to other sites
  - 4. Repeat 2 to 4

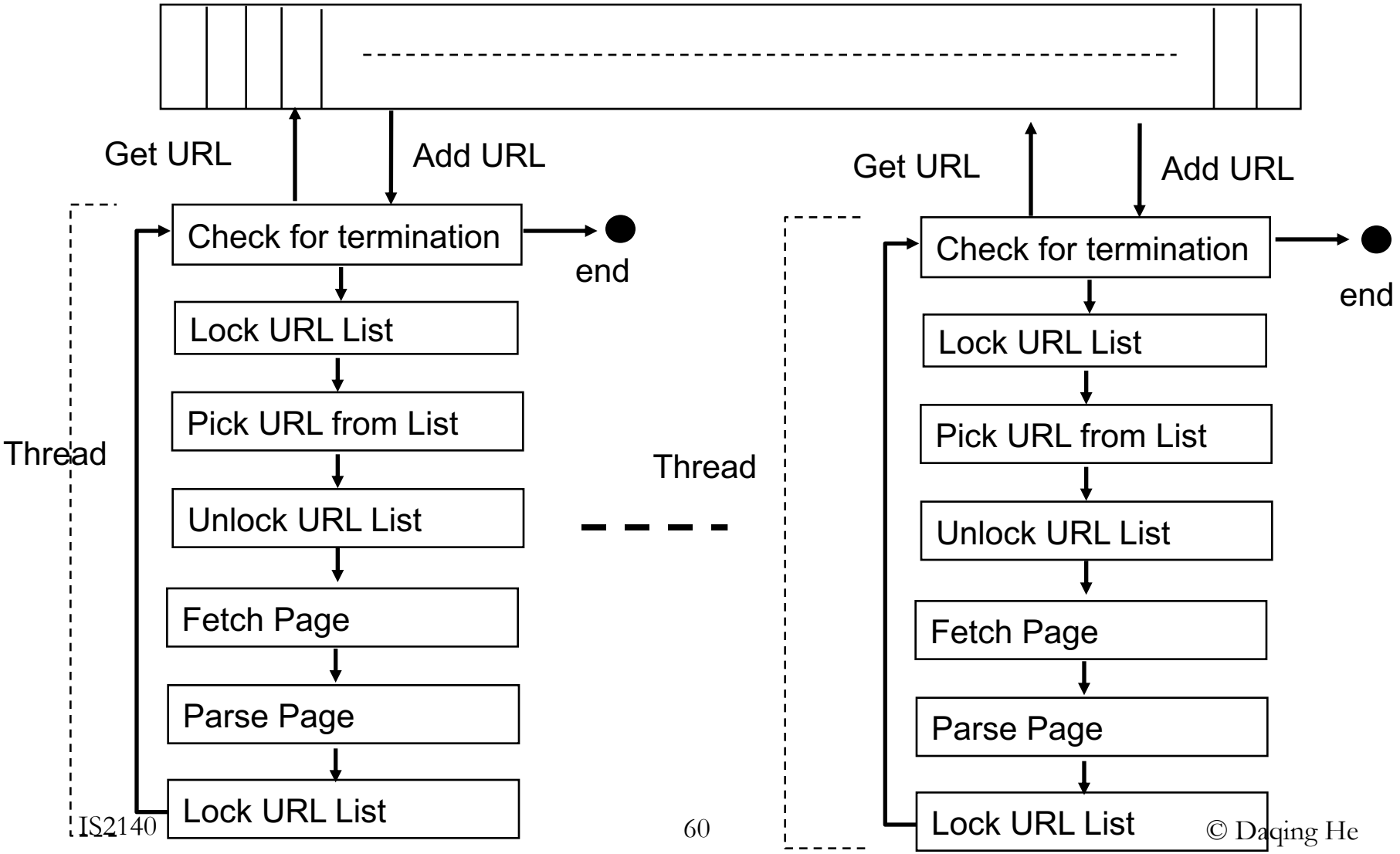
Crawling the Web



Crawler basic algorithm

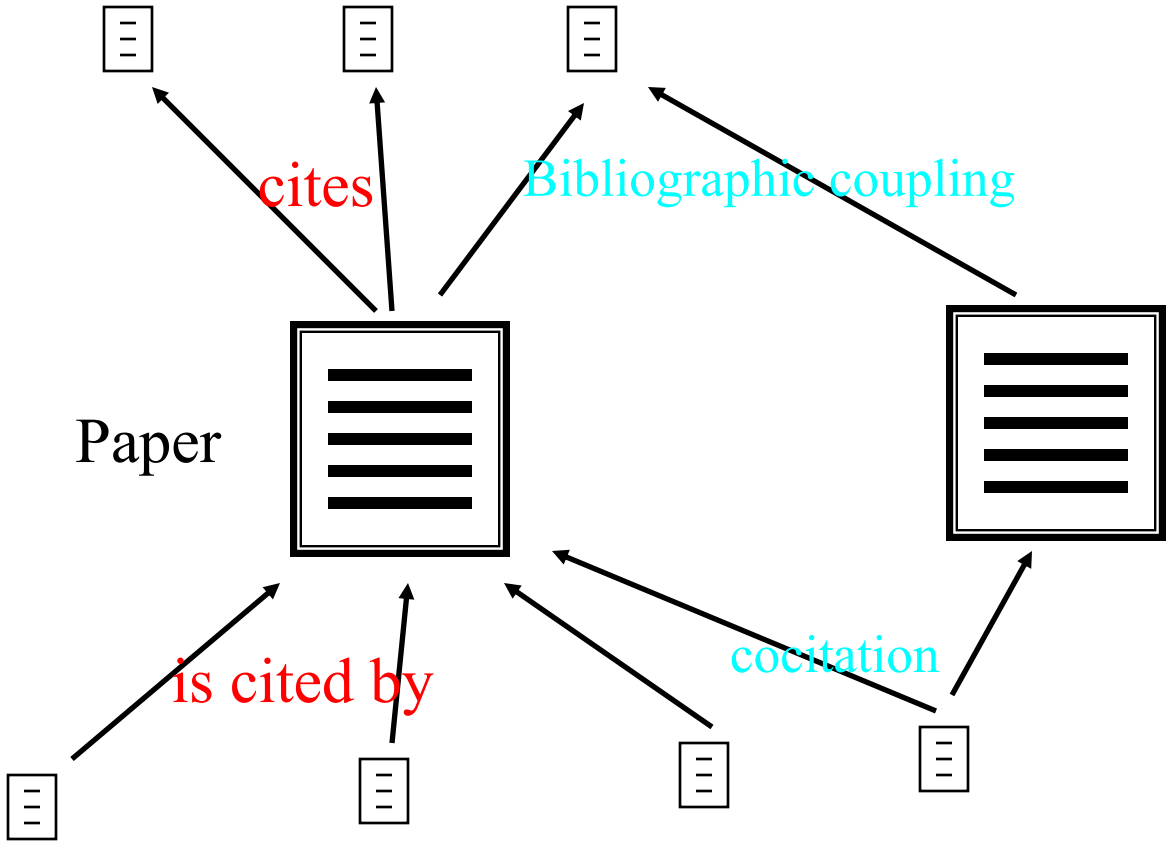


Multithreaded Crawler



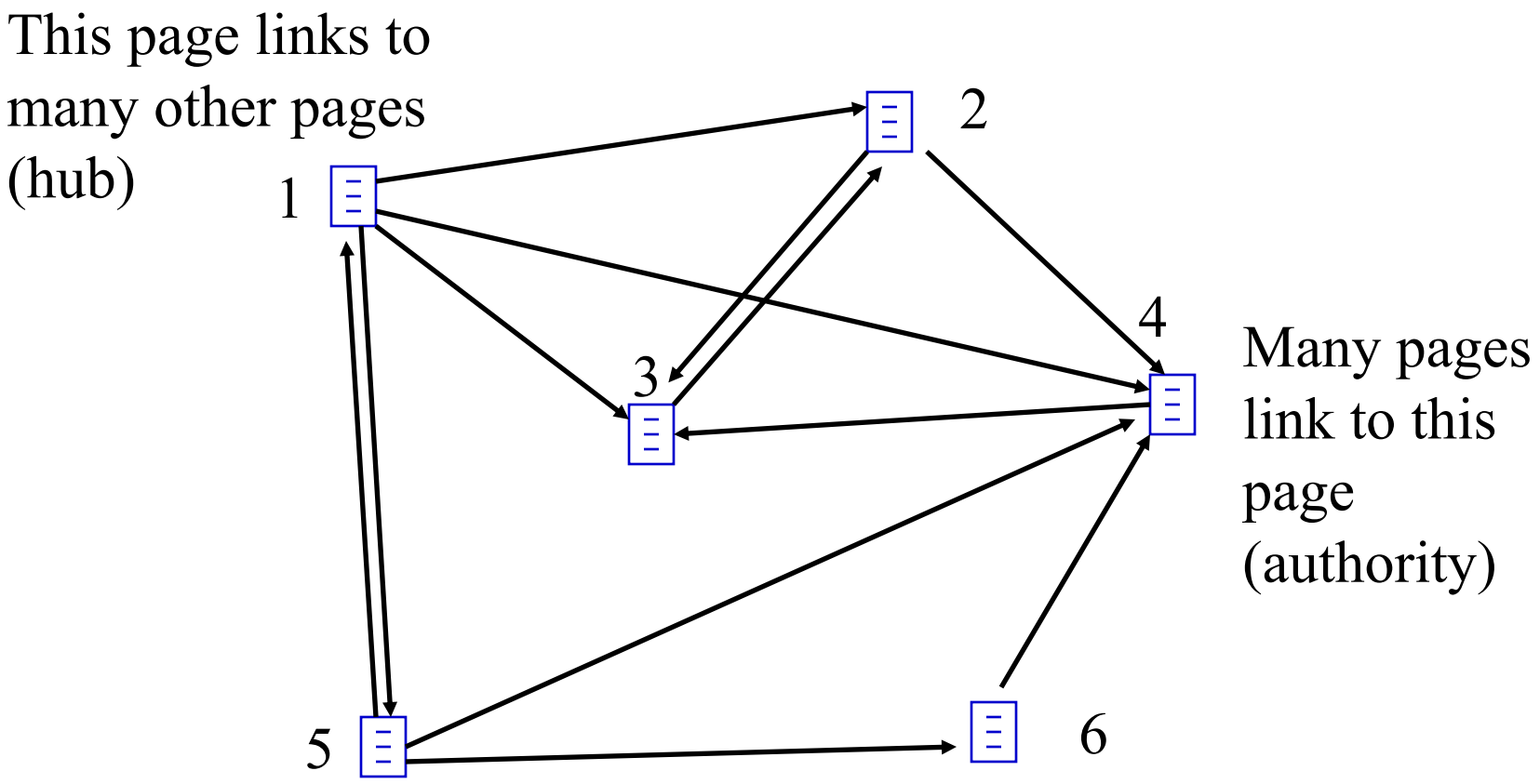
Link Analysis

Citation Graph



Note that academic citations nearly always refer to earlier work.

Graphical Analysis of Hyperlinks on the Web



Impact Factor

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

Link Analysis

- Exploiting hyperlink structure of web pages to find relevant and importance pages for a user query
- Assumptions :
  - Hyperlink from page A to page B is a recommendation of page B from the author of page A
  - If page A and page B are connected by a hyperlink , they might be on the same topic.
- Used for crawling, ranking, computing the geographic scope of a web page, finding mirrored hosts , computing statistics of web pages and search engines, web page categorization.

Link Analysis

- Most popular methods :
  - Hypertext Induced Topic Search (HITS) (1998)
    By Jon Kleinberg
  - PageRank (1998)
    By Lawrence Page & Sergey Brin
    Google's founders

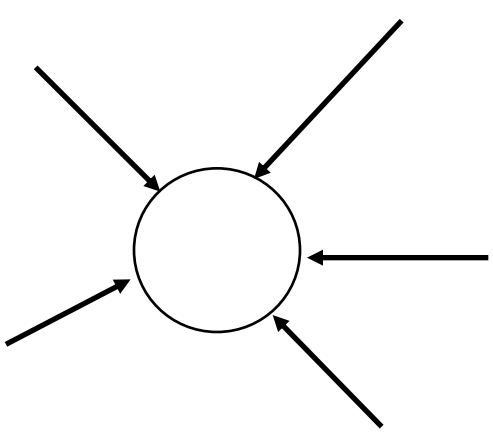
Authority and Hub

Authority: Roughly, a authority is a page with many in-links.

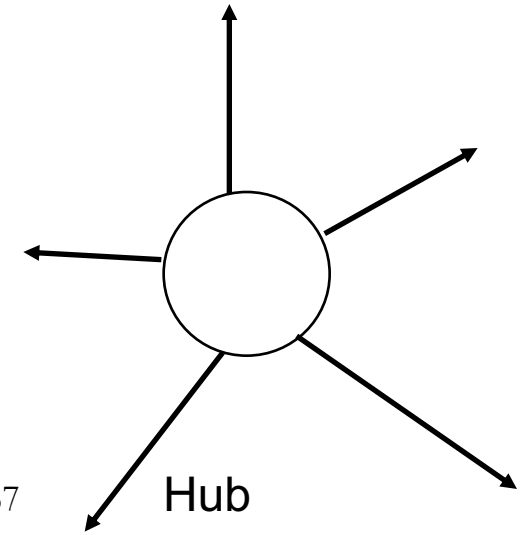
- the page may have good or authoritative content on some topic
- thus many people trust it and link to it.

Hub: A hub is a page with many out-links.

- The page serves as an organizer of the starting point for a particular topic , especially points to many good authority pages on the topic.



Authority



Hub

The key idea of HITS

- A good hub points to many good authorities, and
- A good authority is pointed to by many good hubs.
- Authorities and hubs have a **mutual reinforcement relationship**. Fig. 8 shows some densely linked authorities and hubs (a **bipartite sub-graph**).

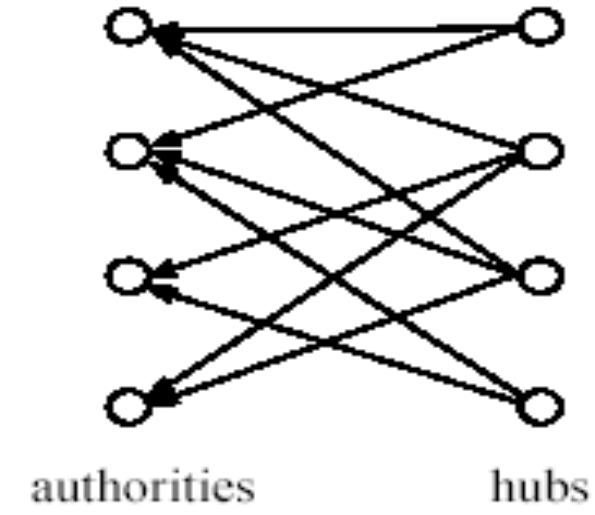
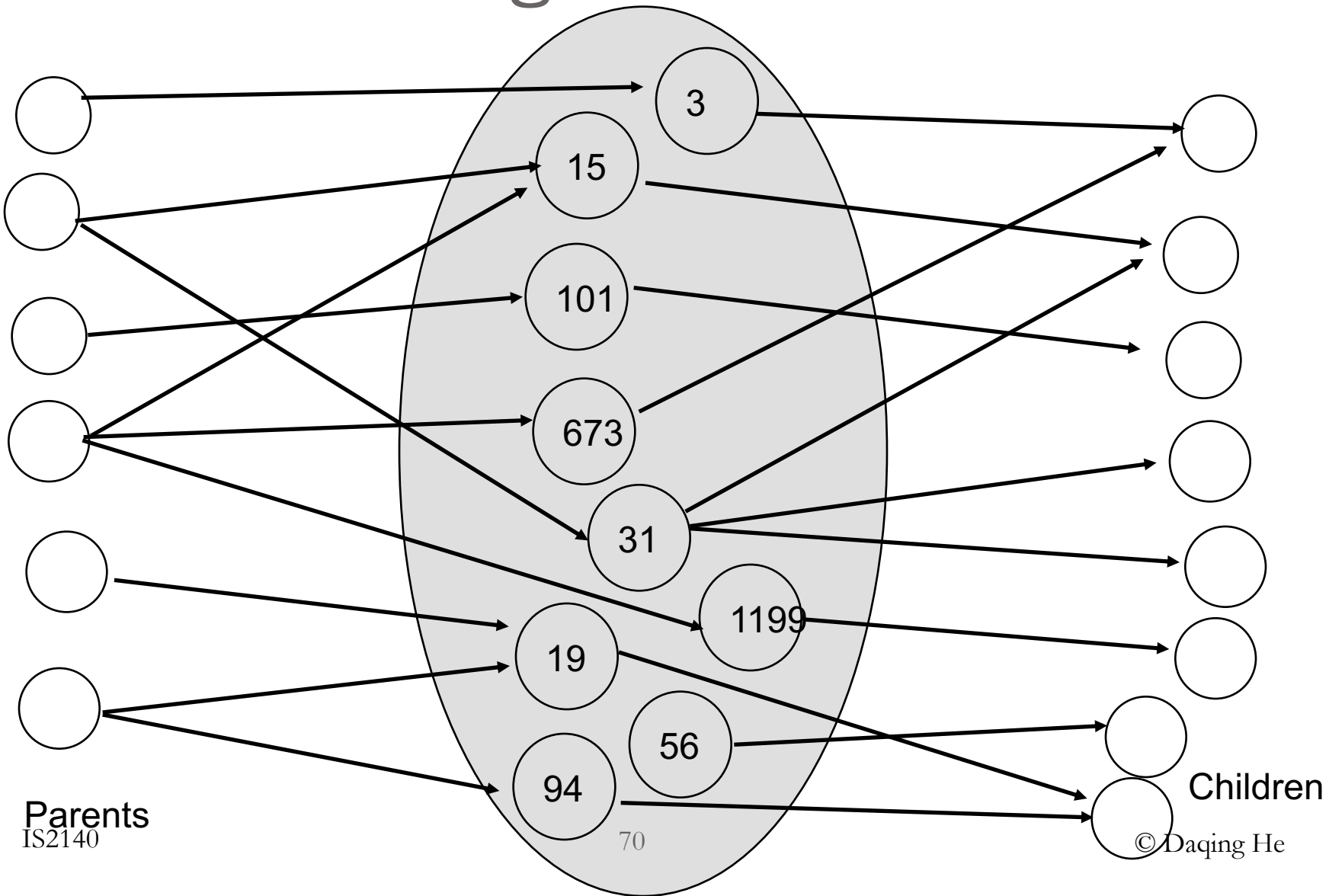


Fig. 8. A densely linked set of authorities and hubs

The HITS algorithm: Grab pages

- Given a search query, q, HITS collects a set of pages as follows:
  - Among the returned pages for the query q from a search engine.
  - collects t ($t = 200$ is used in the HITS paper) highest ranked pages. This set is called the **root set** W.
  - then grows W by including any page pointed to by a page in W and any page that points to a page in W. This gives a larger set S, **base set**.

HITS – Grab Pages

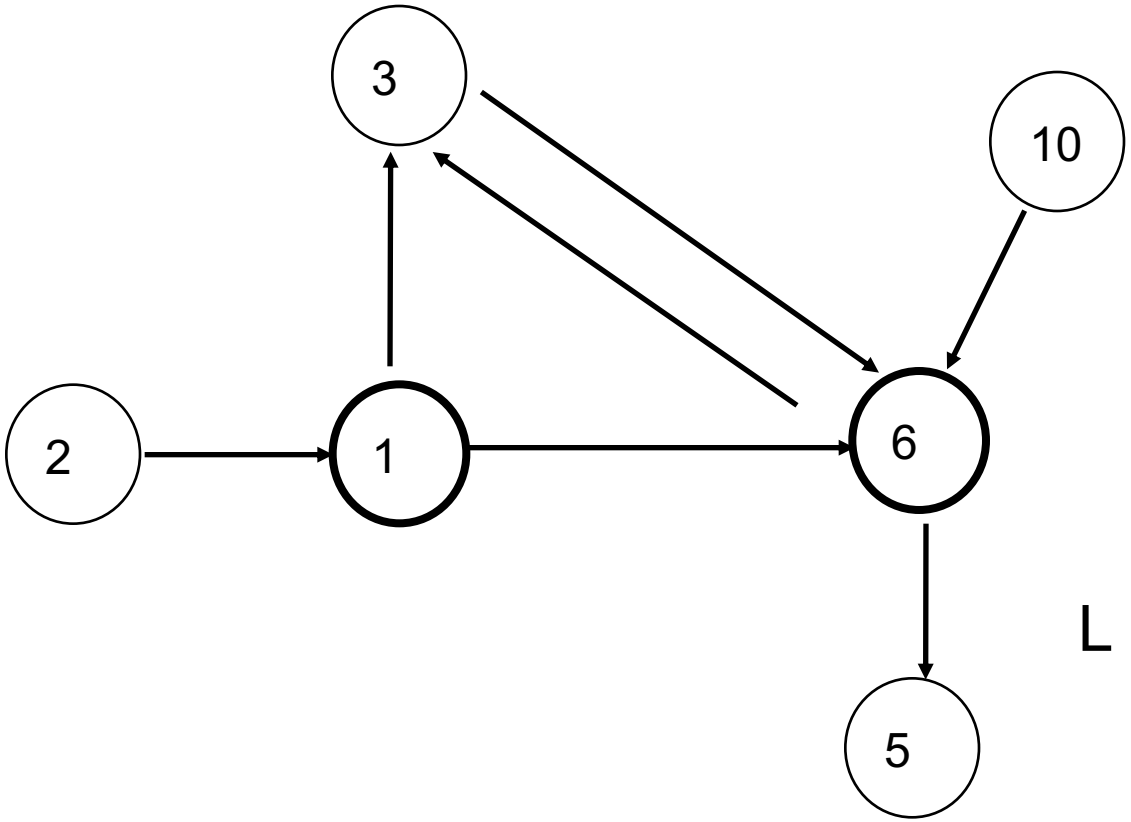


The link graph G

- We use G = (V, E) to denote the hyperlink graph of S.
- We use L to denote the adjacency matrix of the graph.

$$L_{ij} = \begin{cases} 1 & if\ (i,j) \in E \\ 0 & otherwise \end{cases}$$

A Simple Link Graph and L Matrix



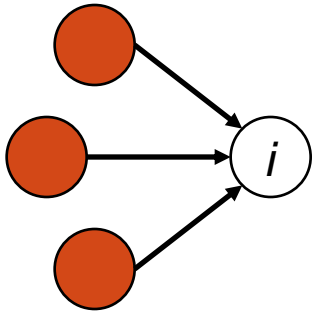
L =

	1	2	3	5	6	10
1	0	0	1	0	1	0
2	1	0	0	0	0	0
3	0	0	0	0	1	0
5	0	0	0	0	0	0
6	0	0	1	1	0	0
10	0	0	0	0	1	0

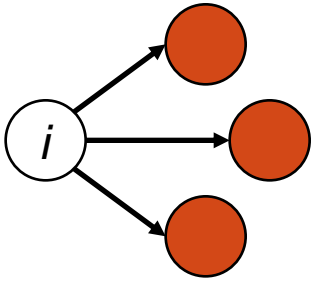
The HITS algorithm

- Let the authority score of the page i be $a(i)$, and the hub score of page i be $h(i)$.
- The mutual reinforcing relationship of the two scores is represented as follows:

$$a(i) = \sum_{(j,i)\in E} h(j)$$



$$h(i) = \sum_{(i,j)\in E} a(j)$$



HITS in matrix form

- HITS works on the pages in S, and assigns every page in S an authority score and a hub score.
- Let the number of pages in S be n.
- We use $\boldsymbol{a}$ to denote the column vector with all the authority scores,

  $\boldsymbol{a} = (a(1), a(2), \ldots, a(n))^T$, and
- use $\boldsymbol{h}$ to denote the column vector with all the hub scores,

  $\boldsymbol{h} = (h(1), h(2), \ldots, h(n))^T$,
- Then,

  $\boldsymbol{a} = \boldsymbol{L}^T\boldsymbol{h}$

  $\boldsymbol{h} = \boldsymbol{L}\boldsymbol{a}$

Computation of HITS

- The computation of authority scores and hub scores uses power iteration.
  - If we use $\boldsymbol{a}_k$ and $\boldsymbol{h}_k$ to denote authority and hub vectors at the kth iteration, the iterations for generating the final solutions are

$$\boldsymbol{a}_k = \boldsymbol{L}^T\boldsymbol{L}\boldsymbol{a}_{k-1}$$

$$\boldsymbol{h}_k = \boldsymbol{L}\boldsymbol{L}^T\boldsymbol{h}_{k-1}$$

starting with

$$\boldsymbol{a}_0 = \boldsymbol{h}_0 = (1, 1, \ldots, 1),$$

- In practice, ~5 iterations get you close to stability.

The algorithm

HITS-Iterate(G)

$\boldsymbol{a}_0 = \boldsymbol{h}_0 = (1, 1, \ldots, 1)$;

$k = 1$

Repeat

$\boldsymbol{a}_k = \boldsymbol{L}^T \boldsymbol{L} \boldsymbol{a}_{k-1}$;

$\boldsymbol{h}_k = \boldsymbol{L}\boldsymbol{L}^T \boldsymbol{h}_{k-1}$;

normalize $\boldsymbol{a}_k$;

normalize $\boldsymbol{h}_k$;

$k = k + 1$;

until $\boldsymbol{a}_k$ and $\boldsymbol{h}_k$ do not change significantly;

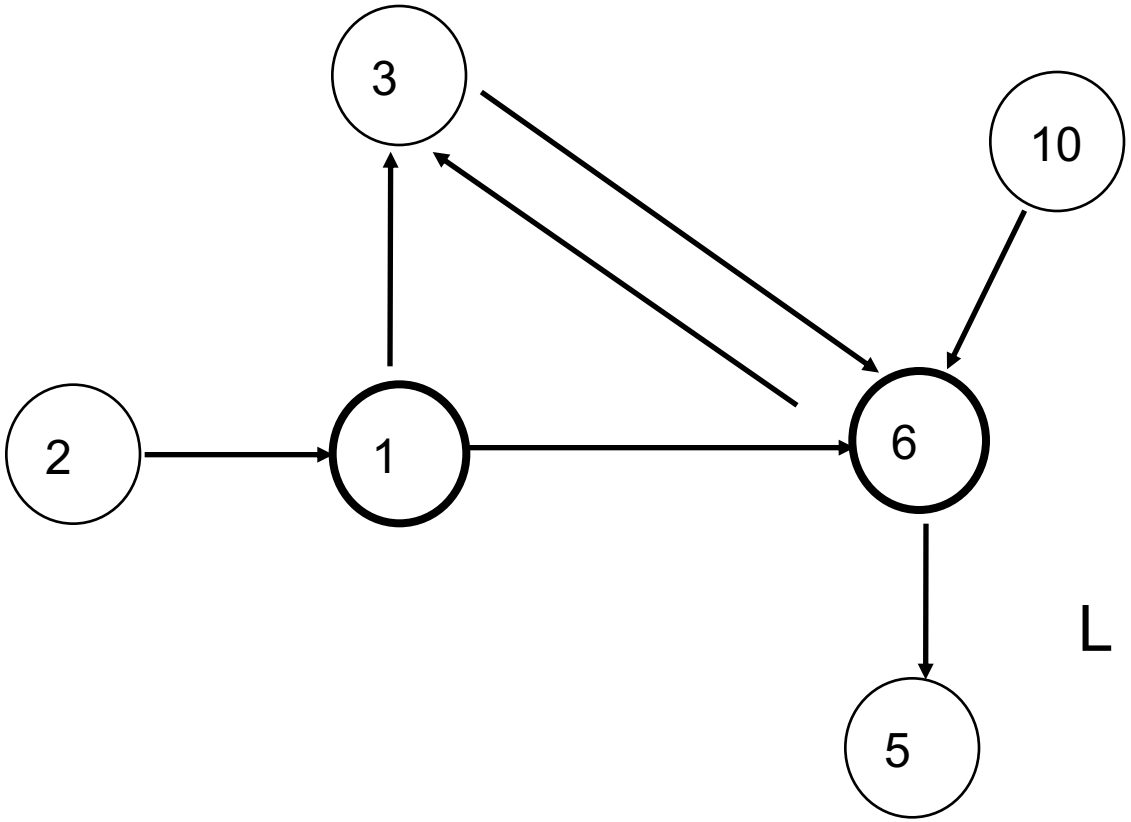
return $\boldsymbol{a}_k$ and $\boldsymbol{h}_k$

$L^T L$ = authority matrix

$L L^T$ = hub matrix

Matrix Multiplication

HITS Example



$$L = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

HITS Example

$L^TL =$

	1	2	3	5	6	10
1	1	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	2	1	1	0
5	0	0	1	1	0	0
6	0	0	1	0	3	0
10	0	0	0	0	0	0

Authorities matrix

$LL^T =$

	1	2	3	5	6	10
1	2	0	1	0	1	1
2	0	1	0	0	0	0
3	1	0	1	0	0	1
5	0	0	0	0	0	0
6	1	0	0	0	2	0
10	1	0	1	0	0	1

Hub matrix

HITS Example

The normalized principles eigenvectors with the Authority score x and Hub y are :

A^T = (0.0002 0 0.5823 0.2163 0.7836 0)

H^T = (0.7054 0.0004 0.4219 0 0.4219 0.4219)

Authority Ranking = (6 3 5 1 2 10)

Hub Ranking = (1 3 6 10 2 5)

Strengths and Weaknesses of HITS

- Strength: its ability to rank pages according to the query topic, which may be able to provide more relevant authority and hub pages.
- Weaknesses:
  - It is easily spammed. It is in fact quite easy to influence HITS since adding out-links in one's own page is so easy.
  - Topic drift. It is possible that a very authoritative yet off-topic document be linked to a document containing the query terms .
  - Inefficiency at query time: The query time evaluation is slow. Collecting the root set, expanding it and performing eigenvector computation are all expensive operations

PageRank

- Introduced by Page et al (1998)
- relies on the democratic nature of the Web by using its vast link structure as an indicator of an individual page's value or quality.
  - PageRank interprets a hyperlink from page *x* to page *y* as a vote, by page *x*, for page *y*. However, PageRank looks at more than the sheer number of votes; it also analyzes the page that casts the vote.
    - Votes casted by "important" pages weigh more heavily and help to make other pages more "important."
  - The pagerank is proportional to its parents' rank, but inversely proportional to its parents' outdegree
- Difference with HITS
  - HITS takes Hubness & Authority weights
  - HITS is query dependent

More specifically

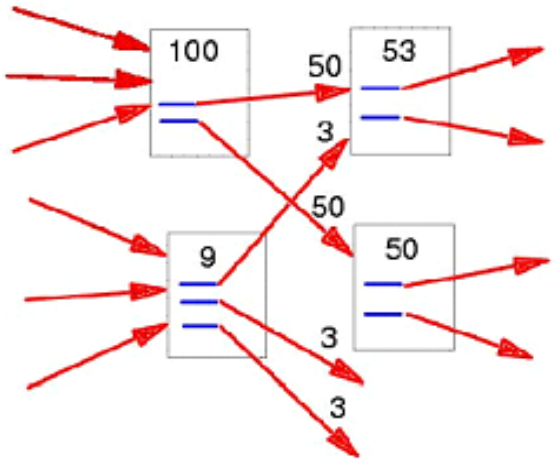
- A hyperlink from a page to another page is an implicit conveyance of authority to the target page.
  - The more in-links that a page *i* receives, the more prestige the page *i* has.
- Pages that point to page *i* also have their own prestige scores.
  - A page of a higher prestige pointing to *i* is more important than a page of a lower prestige pointing to *i*.
  - In other words, a page is important if it is pointed to by other important pages.

Simple View of PageRank

- page i's PageRank score is the sum of the PageRank scores of all pages that point to i.
- Since a page may point to many other pages, its prestige score should be shared.
- The Web as a directed graph G = (V, E). Let the total number of pages be n. The PageRank score of the page i (denoted by P(i)) is defined by:

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j},$$

O_j is the number of out-link of j



Matrix notation

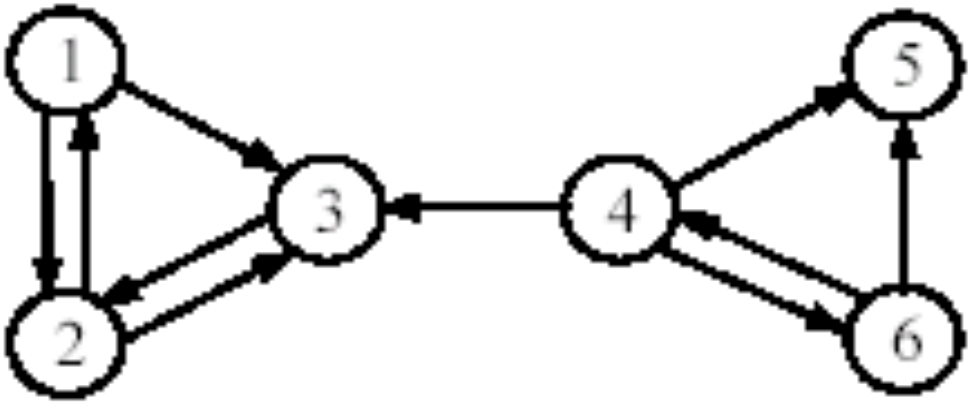
- Let $\boldsymbol{P}$ be a n-dimensional column vector of PageRank values, i.e., $\boldsymbol{P} = (P(1), P(2), \ldots, P(n))^T$.
- The Web as a directed graph G = (V, E). Let the total number of pages be n. Let $\boldsymbol{A}$ be the adjacency matrix of our graph with

$$A_{ij} = \begin{cases} \dfrac{1}{O_i} & if\ (i,j) \in E \\ 0 & otherwise \end{cases}$$

- We can write the calculation of PageRank

$$\boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P}$$

An example Web hyperlink graph



$$\boldsymbol{A} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Solve the PageRank equation

$$\boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P}$$

- This is the characteristic equation of the **eigensystem**, where the solution to $\boldsymbol{P}$ is an **eigenvector** with the corresponding **eigenvalue** of 1.
- It turns out that if ***some conditions*** are satisfied, 1 is the largest **eigenvalue** and the PageRank vector $\boldsymbol{P}$ is the **principal eigenvector**.
- A well known mathematical technique called **power iteration** can be used to find $\boldsymbol{P}$.
- **Problem**: the above Equation does not quite suffice because the Web graph does not meet the conditions.

Random surfing

- Recall we use O_i to denote the number of out-links of a node i.
- Each transition probability is $1/O_i$ if we assume the Web surfer will click the hyperlinks in the page i uniformly at random.
  - The "back" button on the browser is not used and
  - the surfer does not type in an URL.

Dangling Pages

- Dangling Pages: Web pages have no out-links, which are reflected in transition matrix $\boldsymbol{A}$ by some rows of complete 0's.
- Solutions:
  - Idea 1: Remove those pages with no out-links during the PageRank computation as these pages do not affect the ranking of any other page directly.
  - Idea 2: Add a complete set of outgoing links from each such page i to all the pages on the Web.
  - Assume idea 2, then

$$\overline{\boldsymbol{A}} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Improved PageRank

- After this augmentation, at a page, the random surfer has two options
  - With probability d, he randomly chooses an out-link to follow.
  - With probability 1-d, he jumps to a random page
- the improved model,

$$\boldsymbol{P} = ((1-d)\frac{\boldsymbol{E}}{n} + d\boldsymbol{A}^T)\boldsymbol{P}$$

where $\boldsymbol{E}$ is $\boldsymbol{e}\boldsymbol{e}^T$ ($\boldsymbol{e}$ is a column vector of all 1's) and thus $\boldsymbol{E}$ is a $n \times n$ square matrix of all 1's.

Follow our example

$$(1-d)\frac{\boldsymbol{E}}{n}+d\boldsymbol{A}^T=\begin{pmatrix} 1/60 & 7/15 & 1/60 & 1/60 & 1/6 & 1/60 \\ 7/15 & 1/60 & 11/12 & 1/60 & 1/6 & 1/60 \\ 7/15 & 7/15 & 1/60 & 19/60 & 1/6 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 1/60 \end{pmatrix}$$

Where d= 0.9, n = 6

The final PageRank algorithm

- If we scale the formula in slide 58 so that $\boldsymbol{e}^T\boldsymbol{P} = n$,

$$\boldsymbol{P} = (1-d)\boldsymbol{e} + d\boldsymbol{A}^T\boldsymbol{P}$$

- PageRank for each page i is

$$P(i) = (1-d) + d\sum_{j=1}^{n} A_{ji}P(j)$$

- This is equivalent to the formula given in the PageRank paper

$$P(i) = (1-d) + d\sum_{(j,i)\in E} \frac{P(j)}{O_j}$$

- The parameter d is called the **damping factor** which can be set to between 0 and 1. $d = 0.85$ was used in the PageRank paper.

PageRank Algorithm

- Use the power iteration method

PageRank-Iterate(G)

$P_0 \leftarrow e/n$

$k = 1$

repeat

$P_{k+1} \leftarrow (1-d)e + dA^T P_k$;

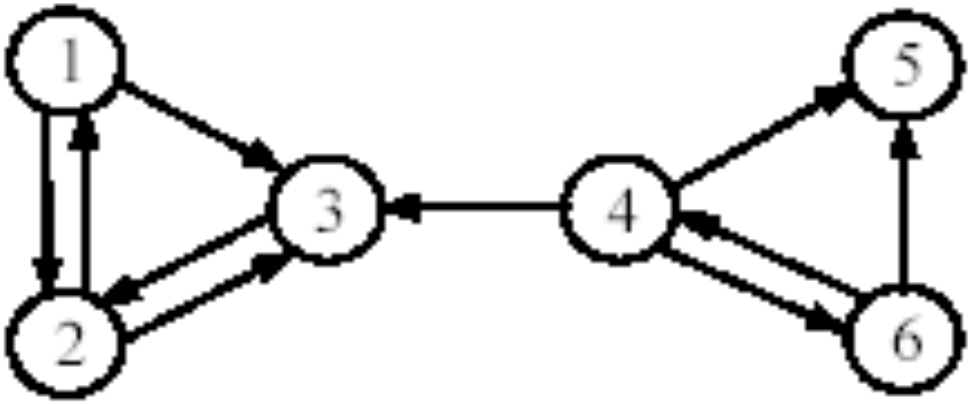
$k = k + 1$;

until $||P_{k+1} - P_k||_1 < \varepsilon$

return P_{k+1}

Fig. 6. The power iteration method for PageRank

Still the example



- After 4 iterations, the PageRank vector P is

  (0.1878 0.3605 0.2859 0.0515 0.0687 0.0456)
- So the ranking of pages based on their PageRank scores is

  2, 3, 1, 5, 4, 6

Advantages of PageRank

- **Fighting spam**. A page is important if the pages pointing to it are important.
  - Since it is not easy for Web page owner to add in-links into his/her page from other important pages, it is thus not easy to influence PageRank.
- **PageRank is a global measure and is query independent**.
  - PageRank values of all the pages are computed and saved off-line rather than at the query time.
- Criticism: Query-independence. It could not distinguish between pages that are authoritative in general and pages that are authoritative on the query topic.

Summary

- Web as a collection is different to traditional document collections
  - The size, the documents, the status, the locations of documents
- Web Search Engines then
  - Need to crawl the Web to build their collections
  - Take advantage of link structures to look for more evidence of authority, hubness, prestige/reputation etc.