Plan for Today's Lecture(s)

- Dimensionality Reduction and LSA (runover from 11/12)
- History of web search
- Web crawling
- Using links to determine relevance

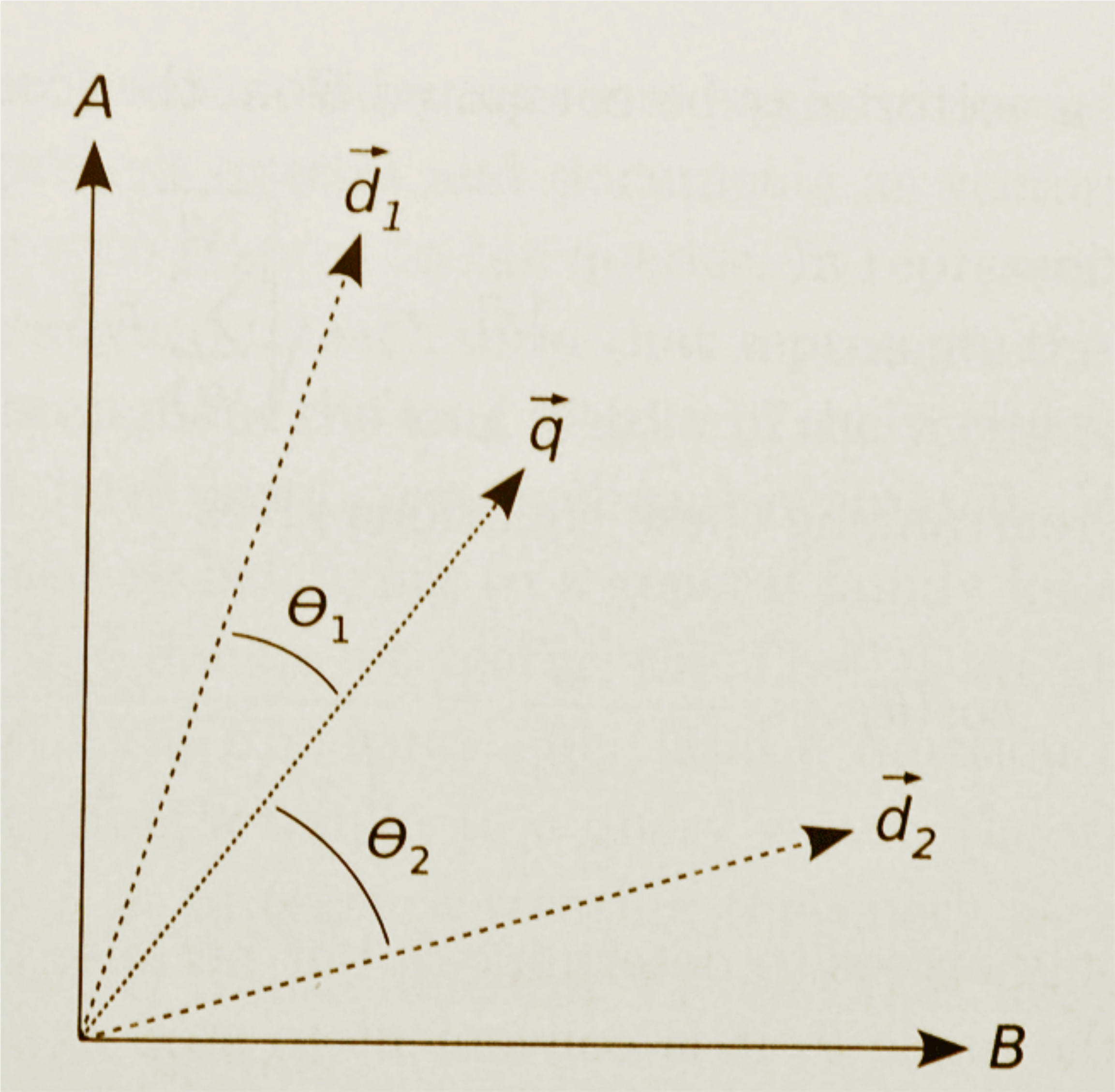
UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
“Information Organization & Retrieval”
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

12 November 2013
Lecture 22.5 – Dimensionality Reduction in the Vector Model

Similarity in Vector Models (Graphical Depiction)



Vector Model with Polysemy

- Because the vector model doesn't recognize that "BANK as in river" and "BANK as in money" are different senses, all occurrences of the term BANK are treated the same instead of being distinguished as separate dimensions in the space
- This overestimates the similarity of documents containing BANK

Vector Model with Synonymy

- The vector model can't recognize that "AUTO" and "CAR" are synonyms, and thus assigns them separate dimensions instead of counting them as additional occurrences of the same "semantic term”
- This underestimates the similarity of documents containing AUTO and CAR

Dimensionality Reduction - A Very Informal Motivation

- Reducing the number of dimensions in a description to the "principle" ones is a common goal in psychology or marketing
- For example, if you have lots of people answer the questions on a personality test, you want to reduce the "person x question" matrix to a "person x PersonalityDimension" one

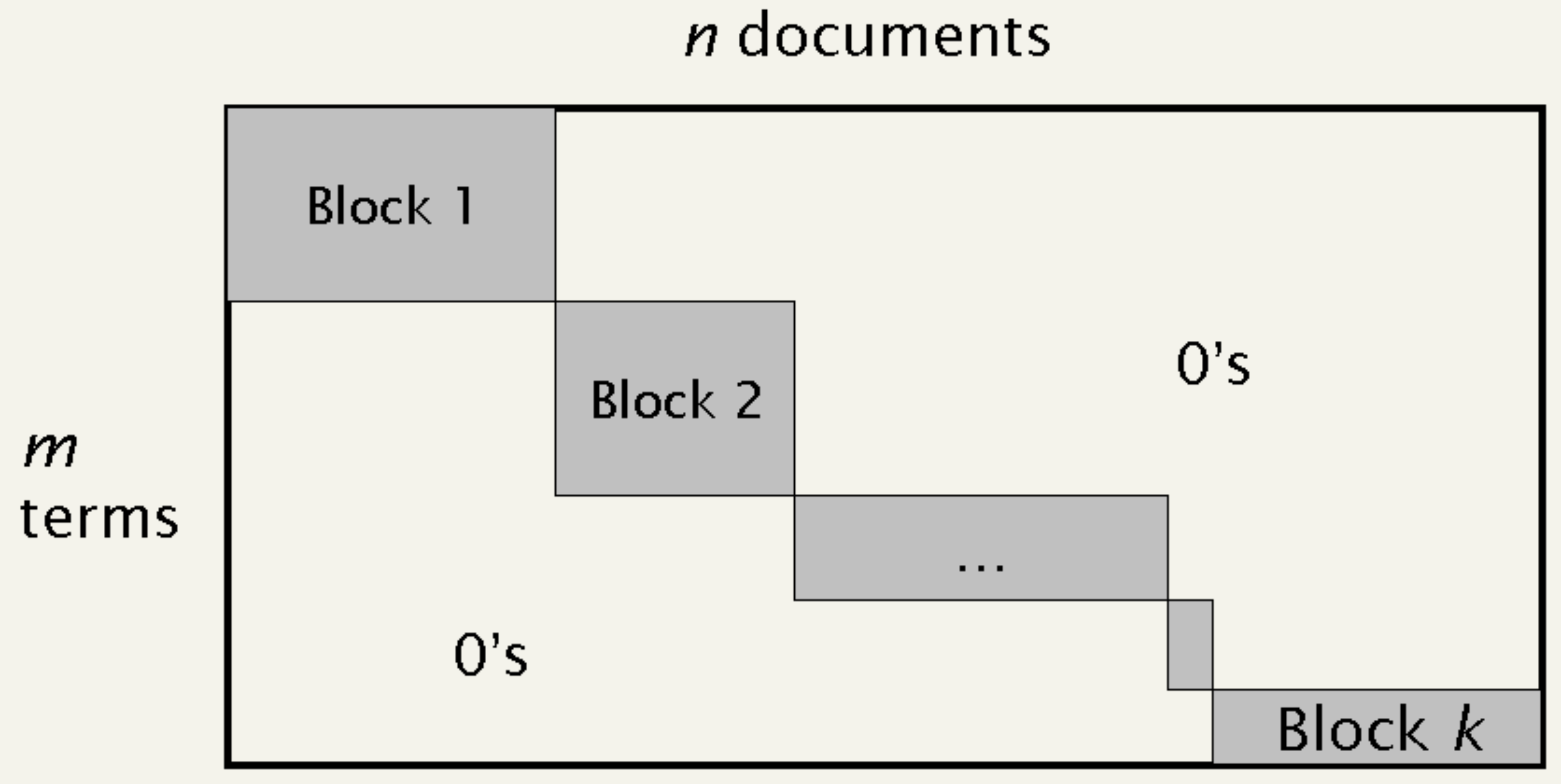
Dimensionality Reduction - A Very Informal Motivation

- If every resource described as “big” is also described as “red,” and every “small” resource is also “green,” this correlation between color and size means that either of these properties is sufficient
- With thousands of properties or descriptive terms, we need clever statistical analysis to choose the optimal descriptive terms
- We can synthesize new “logical terms” based on the correlations

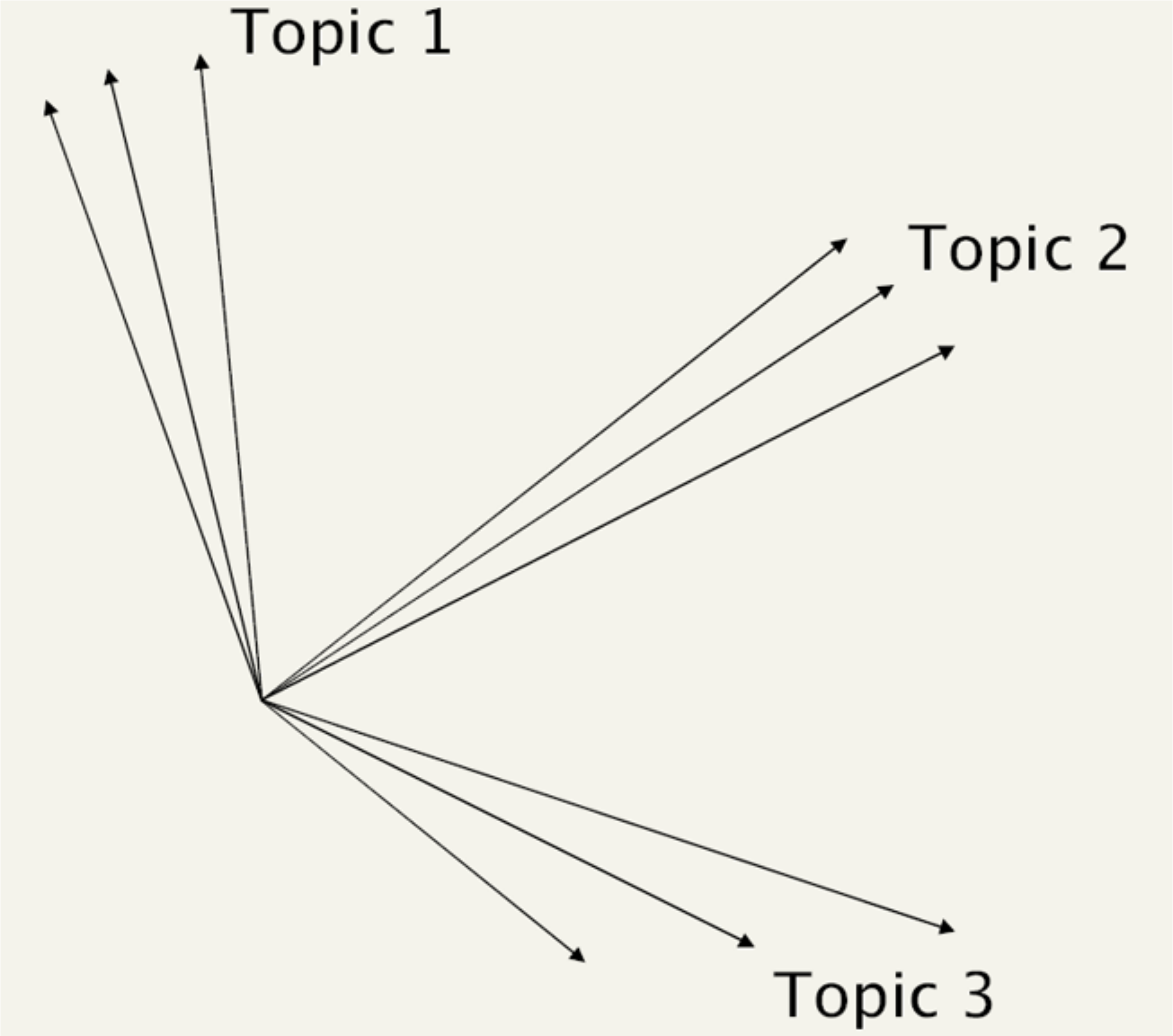
From Terms to Topics

- The dimensionality of the space in the simple vector model is the number of different terms in it
- But the "semantic dimensionality" of the space is the number of distinct topics represented in it
- The number of topics is much lower than the number of terms
- Documents can be similar in the topics they contain even if they have no words in common

An Intuitive Explanation for Dimensionality Reduction Techniques



"Topic Space," Not "Term Space"



Dimensionality Reduction with "Latent Semantic Analysis"

- The vectors in this reduced dimensionality space aren't directly identifiable, but they are "latently" semantic in that relationships between vectors in this lower dimensional space reflect semantic associations
- This dimensionality reduction is completely automatable

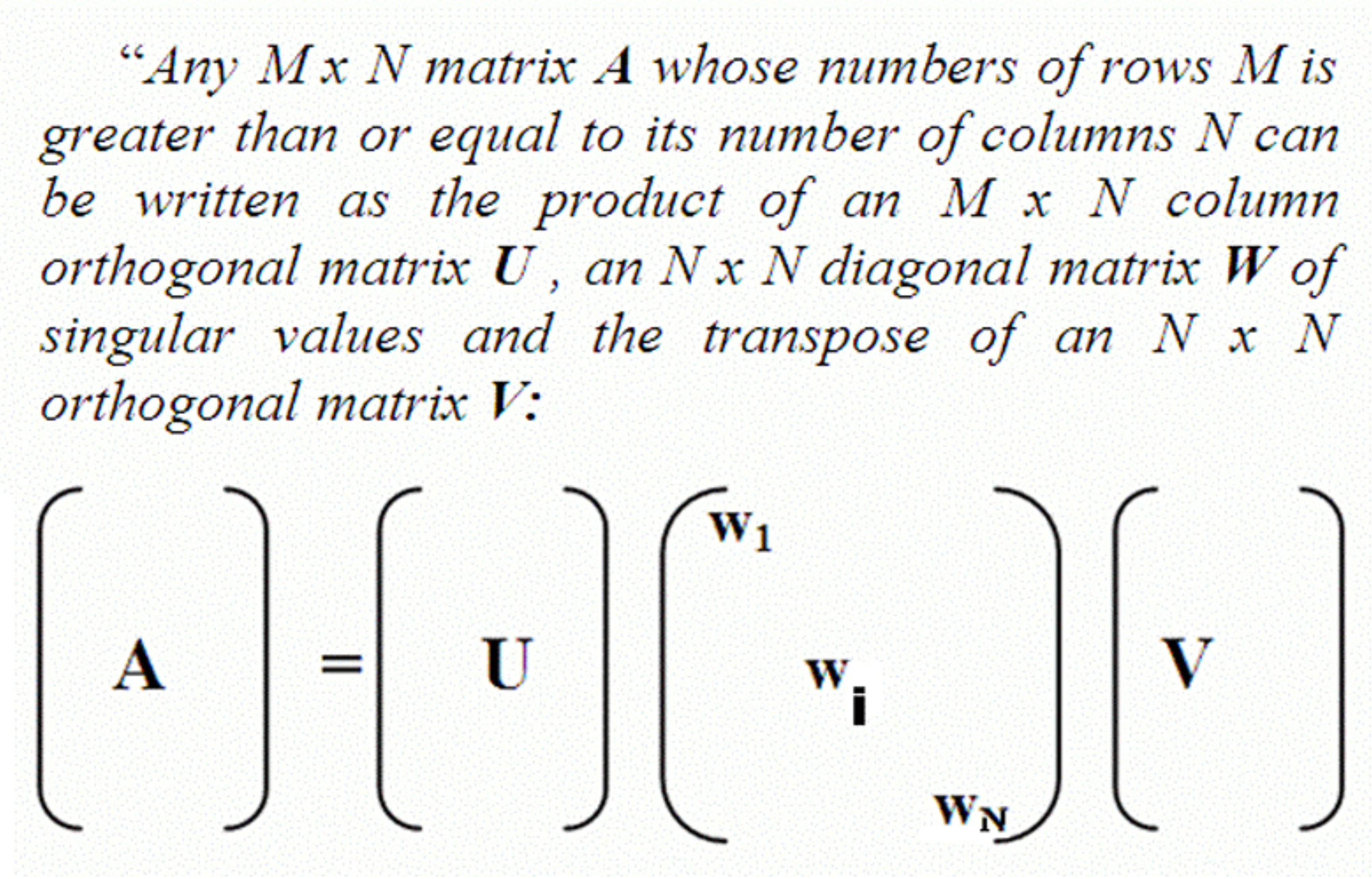
LSA as an IR Model

- Reducing the dimensionality of the Term x Document matrix might suggest that retrieval precision would suffer
- But we're not just discarding terms -- we are replacing sets of co-occurring (e.g., associated) terms with "superterms" or "topics" that represent meaning as a kind of average of all the terms that tend to occur in the same contexts
- We can compute document similarity based on the inner product / cosines in this latent semantic space just as we do with other vector models

LSA Applications

- LSA has been shown to be a practical technique for estimating the substitutability or semantic equivalence of words in larger text segments
- ...which makes it effective in IR, text categorization, and other NLP applications like question answering
- EXAMPLE: LSA to grade essay exams

Singular Value Decomposition (formal)



"*Any M x N matrix* ***A*** *whose numbers of rows M is greater than or equal to its number of columns N can be written as the product of an M x N column orthogonal matrix* ***U*** *, an N x N diagonal matrix* ***W*** *of singular values and the transpose of an N x N orthogonal matrix* ***V*****:**

$$\begin{pmatrix} \\ \mathbf{A} \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \mathbf{U} \\ \\ \end{pmatrix} \begin{pmatrix} w_1 & & \\ & w_i & \\ & & w_N \end{pmatrix} \begin{pmatrix} \\ \mathbf{V} \\ \\ \end{pmatrix}$$

Singular Value Decomposition (less formal)

- The original matrix is decomposed here into three matrices whose product exactly reproduces the original one
- ... which by itself doesn't help us
- ...but the rows and columns are now "orthogonal" or "independent" vectors

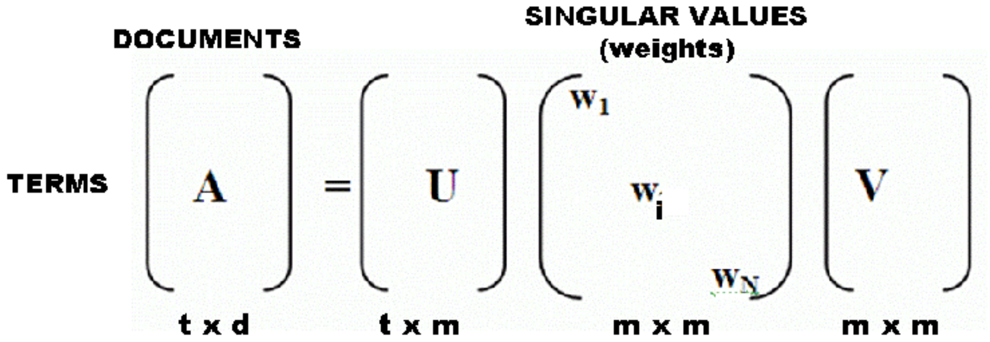
Singular Value Decomposition (less formal)

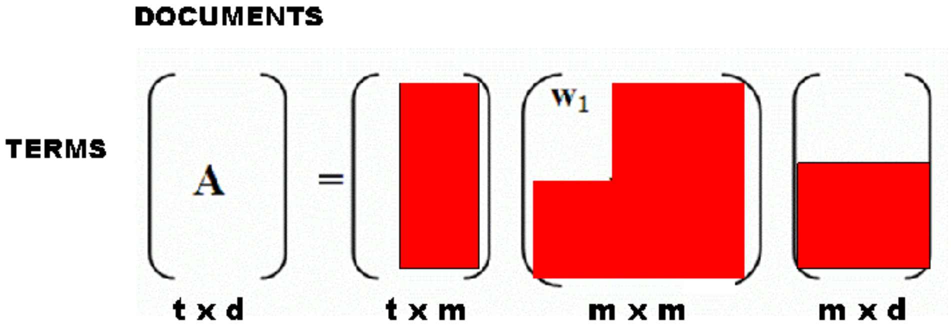
- ... and the "matrix in the middle" (the "singular values") is a diagonal matrix (all values other than the diagonal ones are 0) that is a set of "scaling" values for those new dimensions
- ... ordered in size so we can approximate our original matrix even if we toss out most of the new vectors because they don't explain much of the data

Dimensionality Reduction with "Latent Semantic Analysis"

- Once we have factored the original term x document matrix using SVD, we can then find a much smaller matrix that approximates it
- (...more or less by deleting coefficients from the diagonal matrix, starting with the smallest)
- These techniques in effect "squeeze down" the matrix to lower rank (typically 100-200) by bringing together terms that have similar co-occurrence patterns

Singular Value Decomposition Approximation





UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

19 November 2013
Lecture 24.1 – History of Web Search

History of Web Search

- Infographic showing the long and incremental evolution of systems for searching Internet resources, beginning around 1990 with Archie
- The earliest search engines indexed only the title of Internet resources
- TB-L created the Web to consolidate the diverse formats and protocols for Internet file exchange
- TB-L anticipated a web of academic and research institutions that knew each other, so search and site discovery were not top priorities

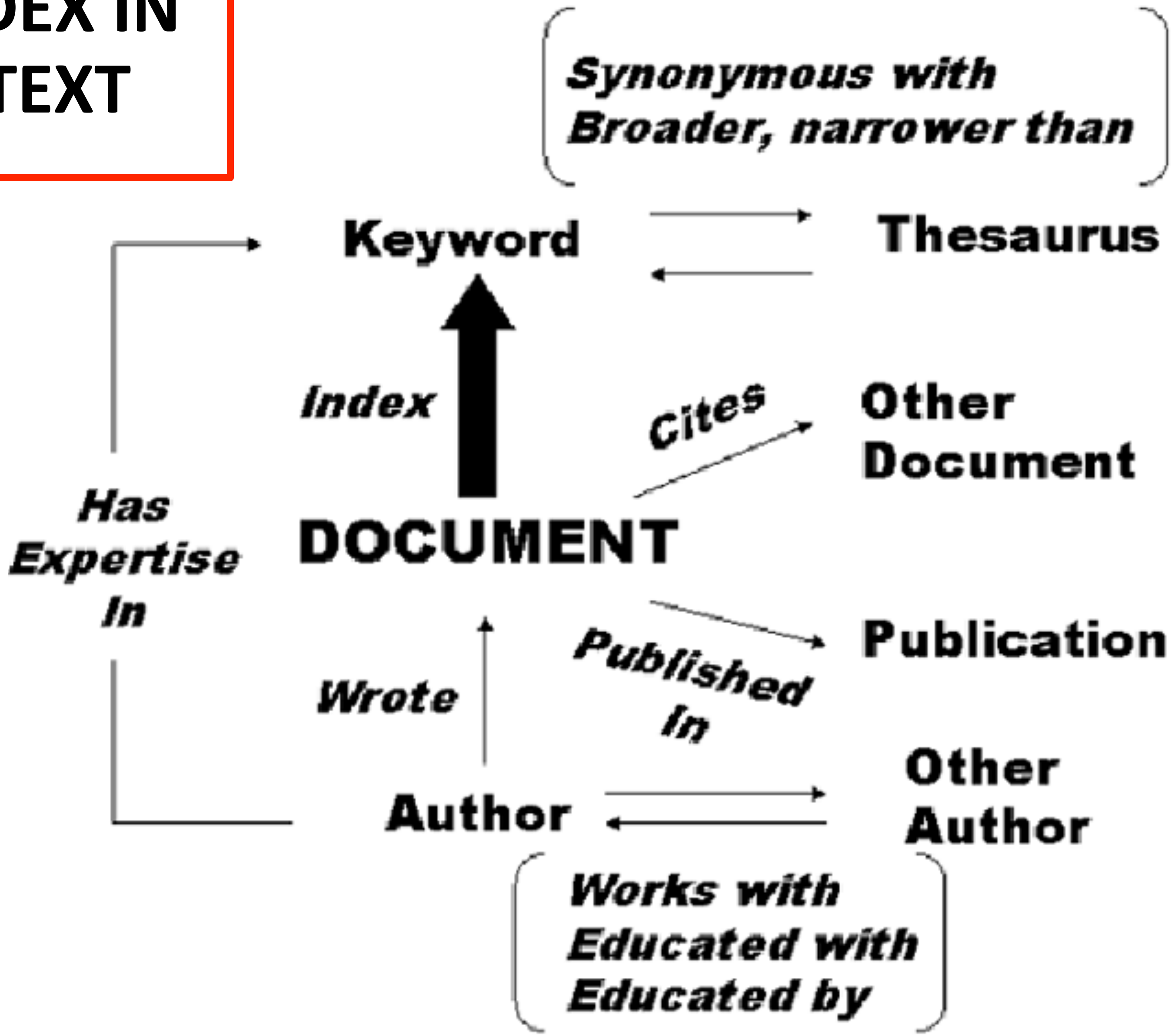
Listing, Indexing, Crawling

- When there were few web sites, each new one was announced on the NCSA “What’s New” page
- After NCSA’s Mosaic graphical browser was released in 1993, the number of sites grew rapidly, and NCSA’s list became inadequate
- Web Wanderer in 1993 was first to create an index of URLs
- Web Crawler in 1994 was first to create an index of web page content

The Index

- When we talked about Boolean model the index was the words and their locations in the document
- But more generally the index is “anything we know about the document” – its content, its metadata, its links, its access frequency, etc.

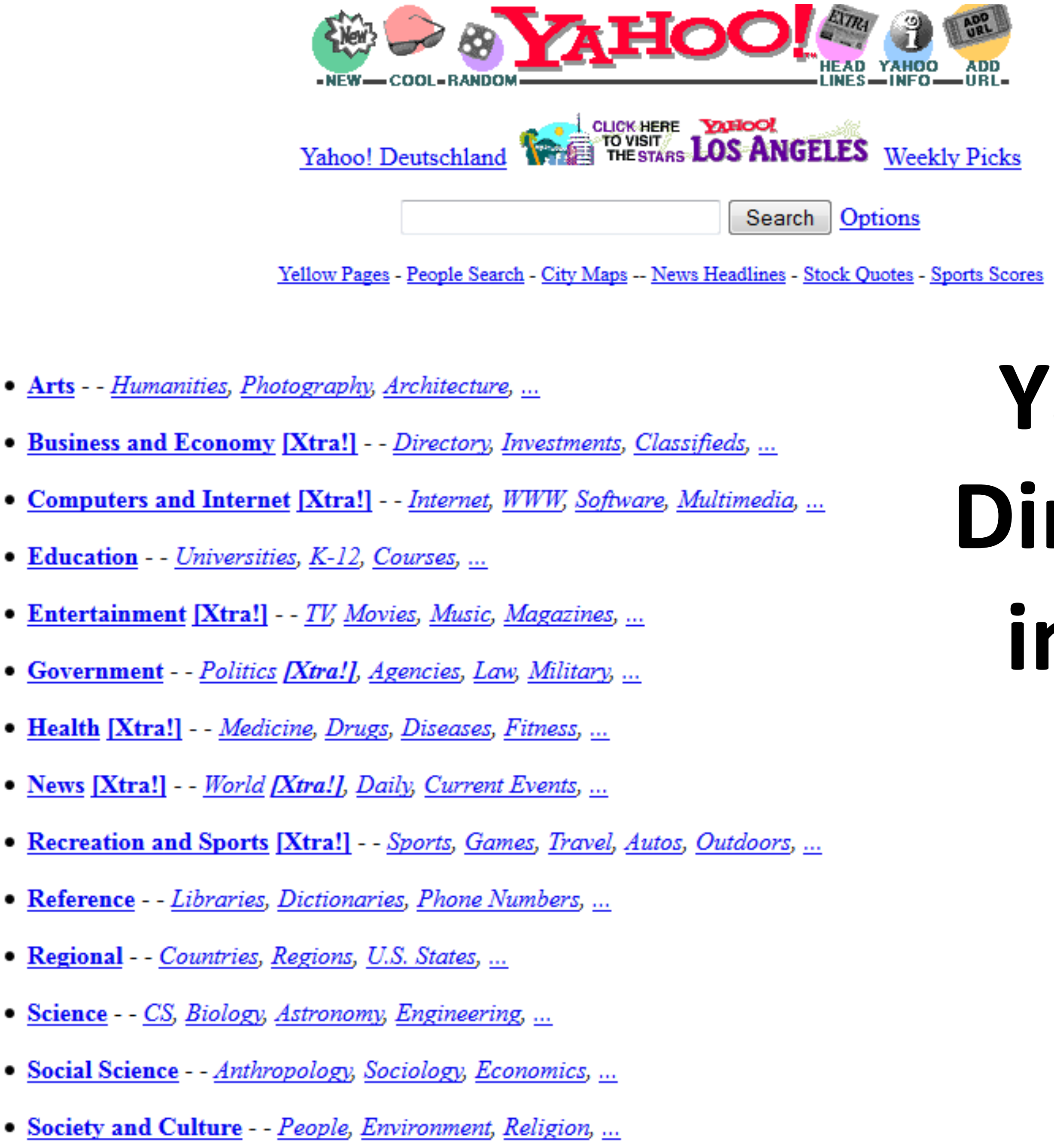
THE INDEX IN CONTEXT



Directories and Portals

- In 1994 two Stanford grad students (Yang and Filo) created a hierarchical directory that organized sites by category
- Searching this human-created organizing system was an effective and popular way to find relevant web sites
- Yahoo! went public in 1996 and became the leading “portal” site that aggregated web sites, games, news, email, and other interactions

Yahoo! Directory in 1996



Yahoo! Deutschland Weekly Picks

Search Options

Yellow Pages - People Search - City Maps -- News Headlines - Stock Quotes - Sports Scores

- Arts - - *Humanities, Photography, Architecture, ...*
- Business and Economy [Xtra!] - - *Directory, Investments, Classifieds, ...*
- Computers and Internet [Xtra!] - - *Internet, WWW, Software, Multimedia, ...*
- Education - - *Universities, K-12, Courses, ...*
- Entertainment [Xtra!] - - *TV, Movies, Music, Magazines, ...*
- Government - - *Politics [Xtra!], Agencies, Law, Military, ...*
- Health [Xtra!] - - *Medicine, Drugs, Diseases, Fitness, ...*
- News [Xtra!] - - *World [Xtra!], Daily, Current Events, ...*
- Recreation and Sports [Xtra!] - - *Sports, Games, Travel, Autos, Outdoors, ...*
- Reference - - *Libraries, Dictionaries, Phone Numbers, ...*
- Regional - - *Countries, Regions, U.S. States, ...*
- Science - - *CS, Biology, Astronomy, Engineering, ...*
- Social Science - - *Anthropology, Sociology, Economics, ...*
- Society and Culture - - *People, Environment, Religion, ...*

Commercialization

- In the mid-90s the web drew commercial interest, leading to substantial debate about whether this was an acceptable use... and the commercial interests won
- Commercial use of the web grew so fast that the directory sites couldn’t keep up, and Altavista, Excite, Infoseek, Inktomi, and Lycos emerged as keyword-based search engines
- The uncontrolled essence of the web enabled the emergence of spam by the mid-1990s

Fighting Spam with Authoritative Relevance

- Vector IR models that worked well in controlled collections were not effective against spam
- Page & Brin at Stanford developed BackRub, used incoming links to estimate "authoritative" relevance
- This "Page Rank" method was substantially better than vector IR models at eliminating spam sites from search results
- Google was founded in 1998 with goal of licensing Page Rank to other search engines, but this strategy wasn't successful

Advertising

- GoTo (soon renamed Overture) was founded in 1998 and invented paid search results placement using keyword auctions; no "organic" results
- Page & Brin were adamantly opposed to this
  - *advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers… it is crucial to have a competitive search engine that is transparent and in the academic realm.*

Ads Pay for It!

- But their advisors and investors convinced Page and Brin to change their minds, and Google added paid search results, segregating them from the results produced by their search engine
- Ad revenue pays for investment in search engine infrastructure, which makes the web more useful, which incents the creation of more web sites, which makes search more necessary…

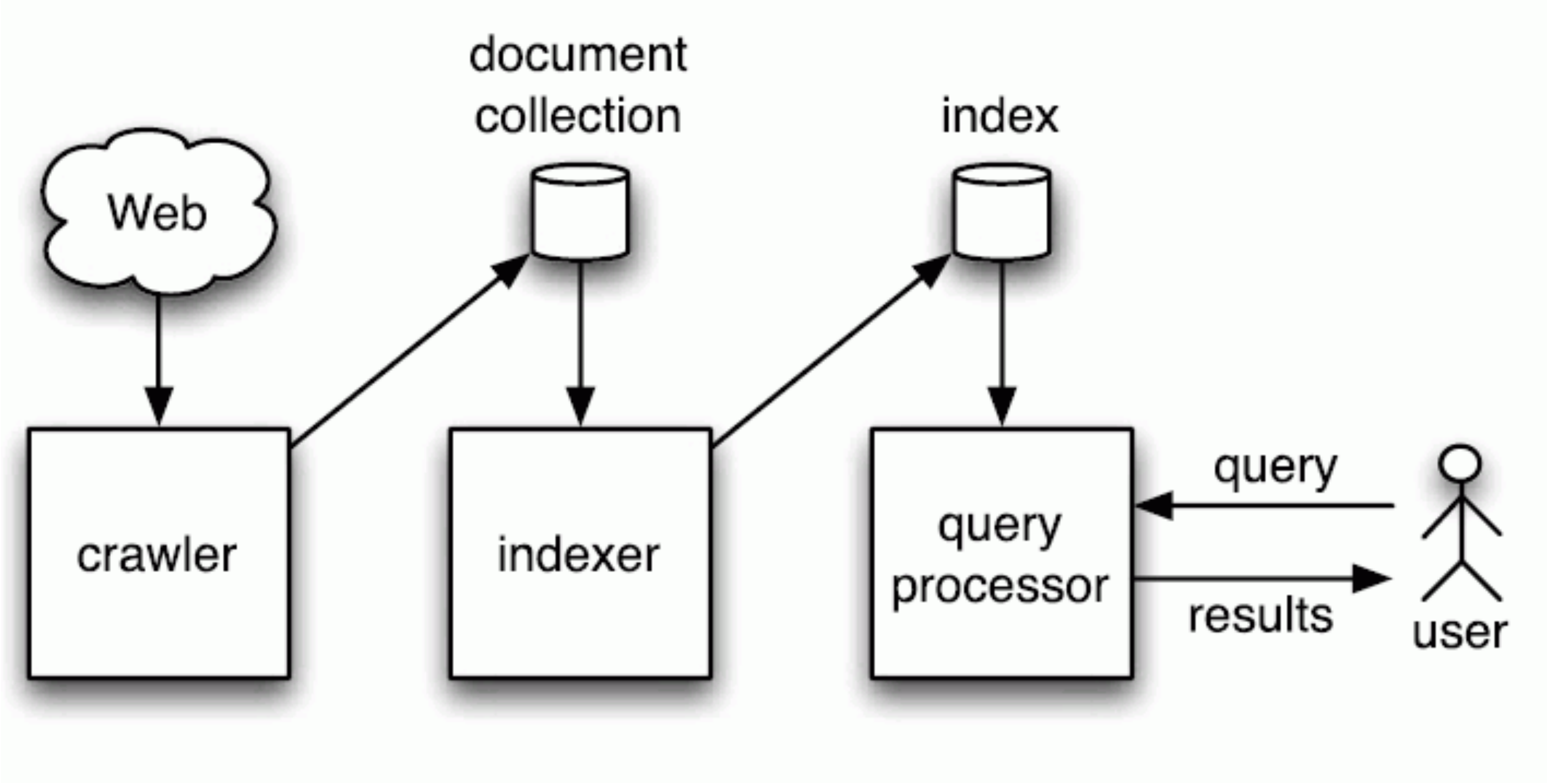
UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

19 November 2013
Lecture 24.2 – Web Crawling

Simplified View of Web Search Engine



Web Crawling: Simplistic View

- How do search engines find web pages to index?
- “Crawling” is the conventional name for this activity but it is misleading
- It suggests that some program that is moving around the web… rather than “staying at home” and sending queries to web sites

Web Crawling: Simplistic View

- Start with known sites
- Record information for these sites
- Follow the links from each site
- Record information from sites found by following links
- Repeat
- (Put "to be processed" pages in a queue)

Web Crawling: Complications

- "Individual" web pages are often highly complex structures with menus, images, advertising, and lots of dynamically generated content
- Some sites are linked to by many pages, and we don't want to process them whenever we find them
- Some sites change a lot, some rarely change
- Duplicate pages
- A great deal of the web is "deep" or "hidden" and not directly accessible to crawlers

Web Crawling: Complications

- Link loops (A links to B, B links to C, C links to A)
- Invalid HTML
- Some sites don't want to be indexed
- Some sites don't deserve to be indexed because they are "link farms" - these just impose useless work on the crawler

The "Deep Web"

- Much of the dynamic web is also the "deep" or "hidden" or "invisible" web whose pages are generated in response to queries submitted to an underlying database or repository
- Using overlap analysis between pairs of search engines, it was estimated in that 43,000–96,000 “deep Web sites” and an informal estimate of 7,500 terabytes of data exist - 500 times larger than the surface Web

He, Patel, Zhang, and Chang. "The Deep Web," Communications of the ACM, May 2007

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
“Information Organization & Retrieval”
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

19 November 2013
Lecture 24.3 – Using Links to Determine Relevance

The Need for Web Citation Analysis

- The primary reason for using web links is because relying only on the content of web pages just doesn't work well enough
- The typical short queries (1 or 2 words) create short query vectors that would bias toward the retrieval of short documents
- Unlike documents in controlled collections, the metadata associated with web pages is often missing or misleading

Adapting Citation Analysis to the Web

- The concepts and techniques of citation analysis seem applicable to the web since we can view it as a network of interlinked articles
- But not everything applies because the web is different in numerous ways
- Google's Page Rank is the exemplar of how citation analysis influences relevance ranking

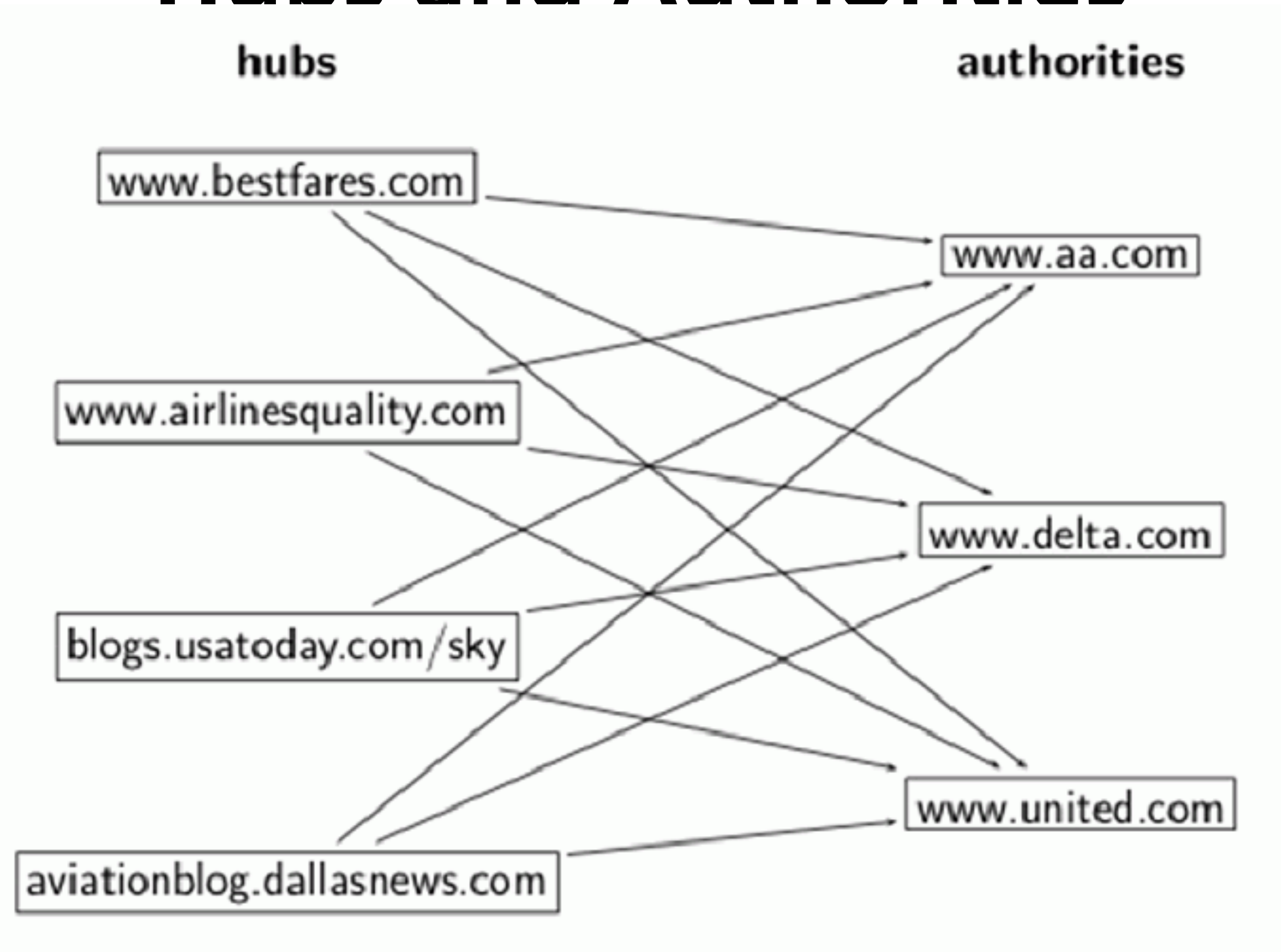
Google Page Rank

- Google is extremely secretive about how it computes relevance, and it uses hundreds of “signals” to select and rank web sites in response to a query
- But the basic ideas about exploiting link structure to estimate relevance have probably not changed much

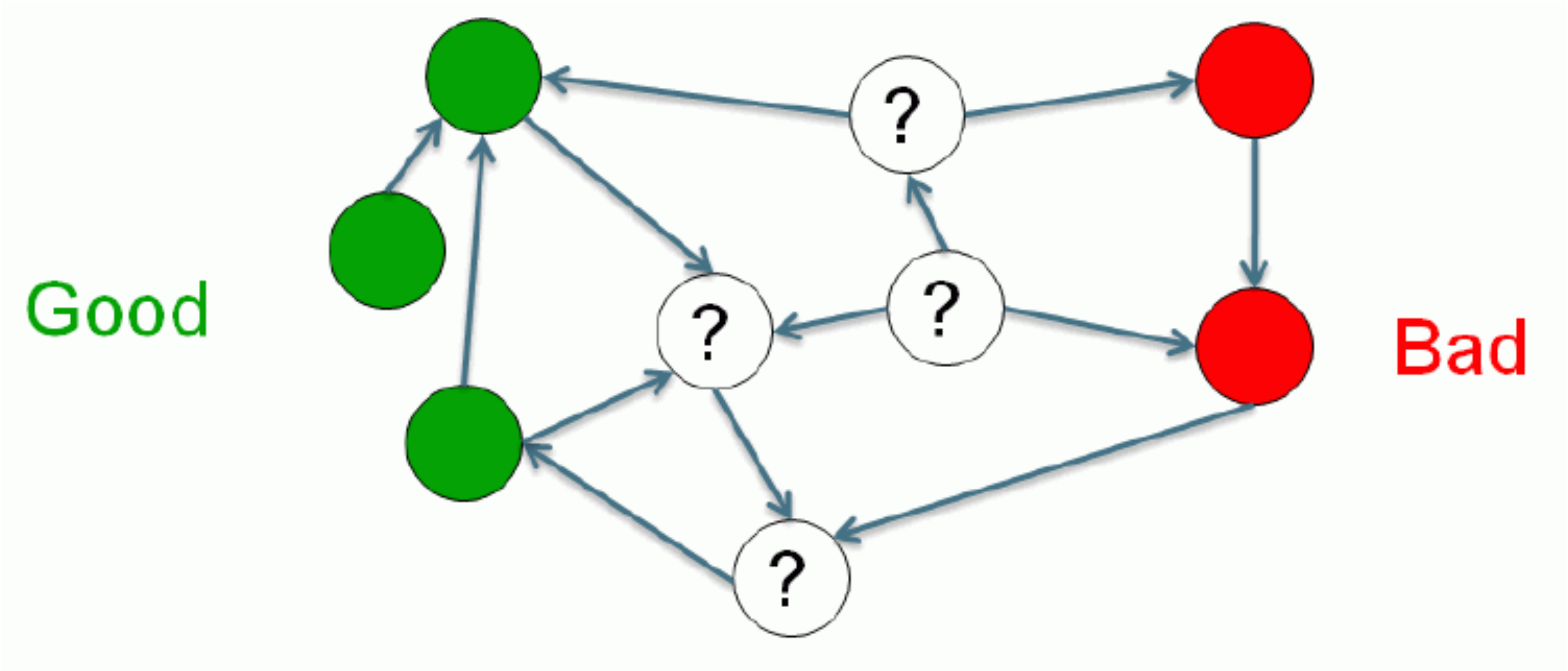
Using Links to Assess Relevance

- Using links to assess the relevance of a web site seems intuitively sensible
- Sites that are the "official" or "authoritative" or "gateway" site for an enterprise or organization will attract links from the Es & Os that have relationships with them -> we should value incoming links
- Sites that these sites then link to are being endorsed by them -> we should value outgoing links from high relevance sites

Hubs and Authorities



Evaluating Incoming and Outgoing Links



- A "good" site only points to other "good" sites, endorsing them...what about the other types?

Evaluating Incoming and Outgoing Links

- If you point to a known bad site, you are either bad or showing poor judgment; either way we should lower our evaluation of you
- If a known bad site points to a known bad site, maybe they are cooperating
- If a known bad site points to a known good site, it isn't the good site's fault
- Should the number of outgoing links a site has influence how we value each link?

The Page Rank "Voting" Calculation

- A site doesn't lose any of its own Page Rank by linking - it is just voting, as in a company shareholders meeting where you get as many votes as you have shares of stock
- The vote for the "linked to" site is divided by the number of outgoing links from the "voting" site
- So it is better for a site's Page Rank if it gets a vote from a site that only has a few of them than from a site that has a lot of them

Page Rank and Relevancy

- Page Rank measures the static structure of each web page, and there is no concept of relevancy in the mathematical description of Page Rank, so it is not dependent on any aspect of a query
- Page Rank comes into play only after some set of relevant documents has been identified by other retrieval models that more directly use the search terms
- The simplest approach would be to order the results by descending Page Rank

Manipulating Page Rank

- There are issues with web links that are analogous to concerns in scientific citation that self-citation and citations to classics distort the "true" link structure and relevance measures
- "Search engine optimization" techniques claim to increase a page or site's page rank
- The simpler the IR model, the easier it is to manipulate

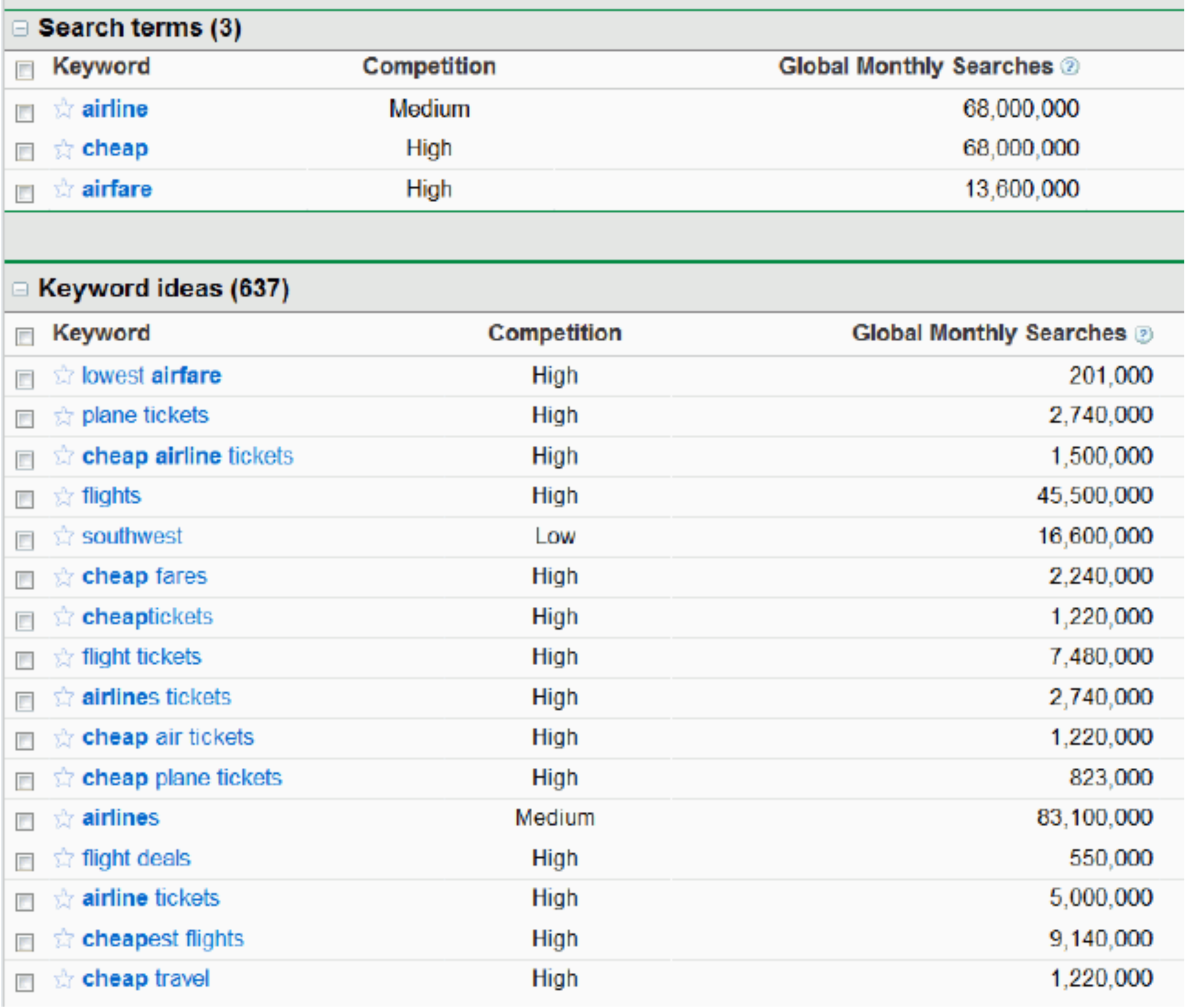
"Good" SEO Techniques

- CONTENT IS KING; good content deserves a higher relevance ranking
- Design your site with a clear hierarchy that is implemented using static text links
- Validate the HTML and make sure that links aren't broken
- Choose informative page titles and section headings that accurately describe the content

"Good" SEO Techniques

- Recognize the vocabulary problem and use the same words as your target customers (http://adwords.google.com)
- Avoid jargon or scientific terms if you aren't aiming for scientists
- BUT BECAUSE EVERYONE COMPETES FOR THE SAME "GOOD" KEYWORDS, GOOGLE GETS RICH SELLING THEM

Keyword Analysis: Google Adwords



⊟ **Search terms (3)**

□	Keyword	Competition	Global Monthly Searches
□	☆ airline	Medium	68,000,000
□	☆ cheap	High	68,000,000
□	☆ airfare	High	13,600,000

⊟ **Keyword ideas (637)**

□	Keyword	Competition	Global Monthly Searches
□	☆ lowest airfare	High	201,000
□	☆ plane tickets	High	2,740,000
□	☆ cheap airline tickets	High	1,500,000
□	☆ flights	High	45,500,000
□	☆ southwest	Low	16,600,000
□	☆ cheap fares	High	2,240,000
□	☆ cheaptickets	High	1,220,000
□	☆ flight tickets	High	7,480,000
□	☆ airlines tickets	High	2,740,000
□	☆ cheap air tickets	High	1,220,000
□	☆ cheap plane tickets	High	823,000
□	☆ airlines	Medium	83,100,000
□	☆ flight deals	High	550,000
□	☆ airline tickets	High	5,000,000
□	☆ cheapest flights	High	9,140,000
□	☆ cheap travel	High	1,220,000

"Good" Linking Techniques

- Provide links to "Resources" that visitors to your site might find useful
- Don't link to sites, or ask for links from sites, that are "bad actors" or not credible

"Good" Linking Techniques

- Use informative link anchor text; it is often a better description of a page than its own title!
- Participate meaningfully in social or professional communities where reciprocal linking is a side effect

"Bad" SEO Techniques

- Some techniques are not intuitive and aren't as related to the IR model; most of these are easily detected by search engines
- Link farms, especially those created by automated techniques - essentially a form of spam directed toward search engines
- Submitting a site to a directory service is rarely a good idea since most of them are little better than link farms in terms of content quality

The Relevance "Arms Race"

- In response to "bad" SEO techniques, search engines adapt their relevance algorithms to ignore or penalize sites that use those strategies
- But Google makes most of its money from selling ads to companies that want to manipulate relevance...
- and these sites then that take up some of the space on the results page that might have gone to the truly relevant and deserving pages that didn't have the money to pay to get seen

Sociopolitical Criticism of Page Rank and Google Relevance Heuristics [1]

- Google says Page Rank relies on the uniquely democratic nature of the web by using its vast link structure
- The Page Rank algorithm favors older pages because a new page, regardless of its quality or relevance, will not have many incoming links

Diaz, Alejandro. "Through the Google goggles: Sociopolitical bias in search engine design."

Sociopolitical Criticism of Page Rank and Google Relevance Heuristics [2]

- Put another way, Page Rank treats popularity as a substitute for relevance
- Similarly, pages from "big company" domains, with short URLs, and whose URLs contain the search terms are treated as more relevant
- So "conventional" authority is favored
- Google also seems to vastly overweight Wikipedia articles

Sociopolitical Criticism of Page Rank and Google Relevance Heuristics [3]

- Does Page Rank systematically disfavor or suppress new, underrepresented, or other voices that are critical of the "mainstream" point of view?

Readings for Next Lecture

- Geller, Tom. “Talking to machines.” Communications of the ACM 55, no. 4 (2012): 14-16.
- Knees, Peter, Tim Pohle, Markus Schedl, and Gerhard Widmer. “A music search engine built upon audio-based and web-based similarity measures.”
- Schmitz, Patrick and Black, Michael – “The Delphi Toolkit: Enabling Semantic Search for Museum Collections”