25. Structure-Based Models [1]

INFO 202 - 24 November 2008

Bob Glushko

Plan for Today's Class

Why Structure-based Models?

Citation Analysis

Adapting Citation Analysis on the Web

Google Page Rank

Criticism of Page Rank

Web Crawling

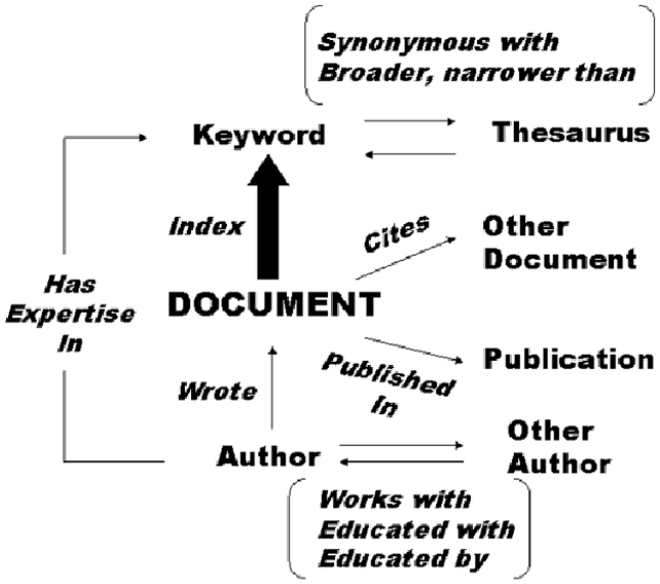
Structure-Based IR Models

Combine representations of terms with information about structures within documents (i.e., hierarchical organization) and between documents (i.e. hypertext links and other explicit relationships)

This structural information tells us what documents and parts of documents are most important and relevant

This information gives us additional justification for determining relevance and ordering a result set

The Index in Context



Equally Reliable and Authoritative?

NOVEMBER 19, 2008

the ONION
America's Finest News Source

HOME VIDEO SPORTS RADIO ELECTION 08 BUSINESS OUR DUMB WO

SEARCH

Supreme Court Upholds Bill Of Rights In 5-4 Decision

NOVEMBER 14, 2008 | ISSUE 44•46

WASHINGTON—In a landmark decision Monday, the U.S. Supreme Court narrowly ruled to uphold the Bill of Rights, the very tenets upon which American society is based. "After carefully considering the relevance of the 10 inviolable rights that comprise the ideological foundation on which our nation is built, the court finds that these basic freedoms remain important for the time being, and should not be overturned," read the majority opinion authored by Justice Anthony Kennedy, who cast the tie-breaking vote. "Until such time as it can be definitively proven that citizens no longer require the protections provided by the Bill of Rights, it shall remain the principal legal guidance for the United States of America." The Supreme Court's latest decision comes on the heels of last month's 6-3 ruling to abolish the pursuit of happiness from the three inalienable rights guaranteed by the Declaration of Independence.

ARTICLE TOOLS
Share This
Email This
Print This
Sponsored by

RELATED ARTICLES
Hussein Judge Hoping For Fair, Speedy Assassinatic
JUNE 23, 2006
Supremes Court Upholds Stopping In The Name Of Love In 2-1 Decision
OCTOBER 31, 2008

The New York Times

Late Edition

VOL. CLVIII ... No. 54,690 NEW YORK, WEDNESDAY, NOVEMBER 12, 2008 $1.50

Aide to Obama Seen as Choice For Justice Post

Decision Is Not Final on Attorney General

DEMOCRATS GAIN AS STEVENS LOSES HIS SENATE RACE

LATEST TALLY IN ALASKA

Vote Averts Expulsion — Lieberman Keeps Committee Post

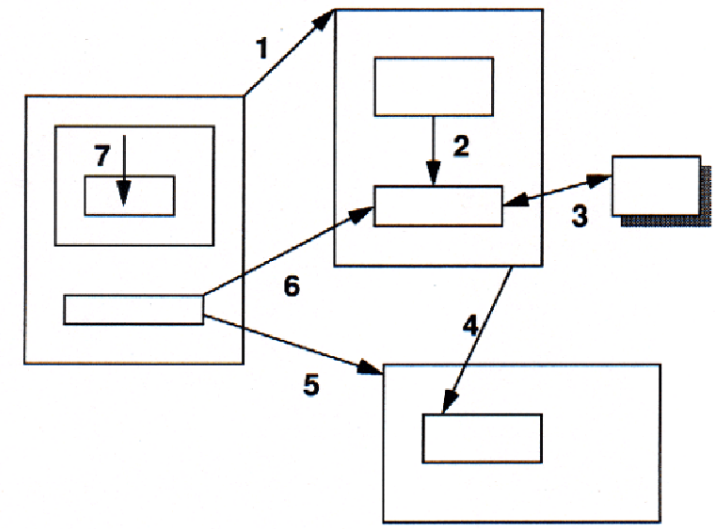
Detroit Chiefs Plead for Aid, To Little Avail

Britain Grapples With Role for Islamic Justice

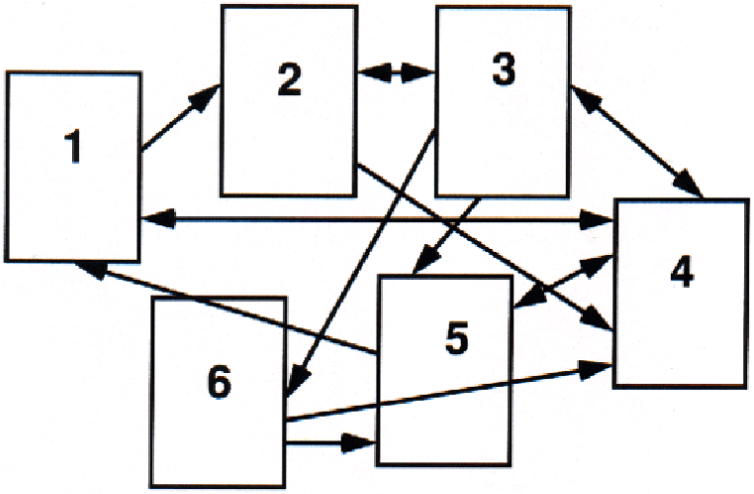
Fouled in the Act of Grunting

A Family Reunited in Congo

Link Structures and Anchors



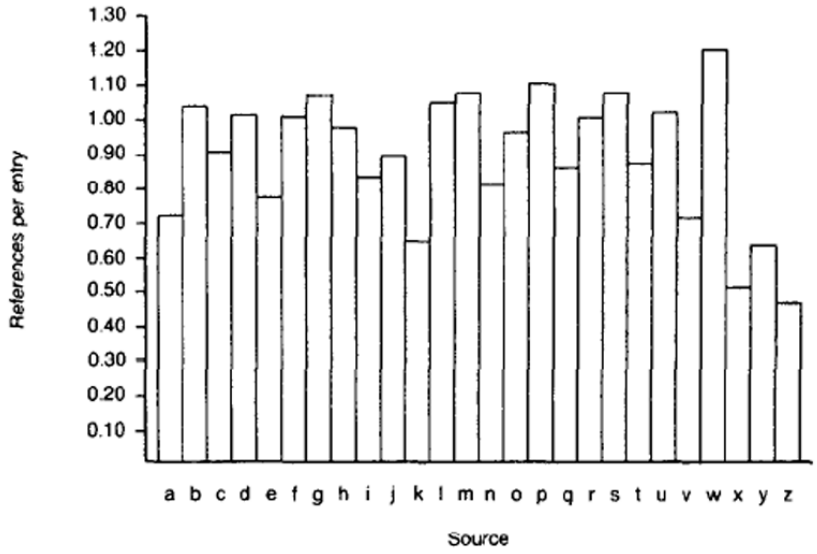
Link Network - Graphical View



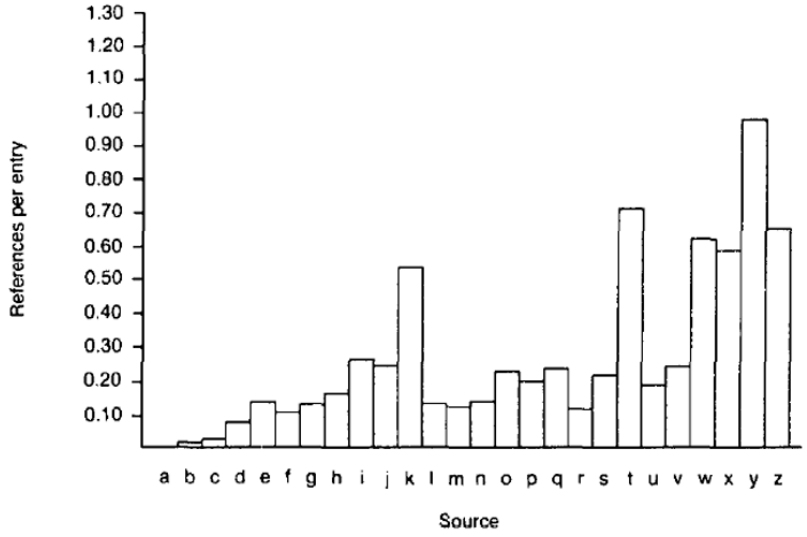
Link Network - Matrix View

	1	2	3	4	5	6
1		x				
2			x	x		
3		x		x	x	x
4	x		x		x	
5	x			x		
6				x	x	

Cross References in the OED [to same letter]



Cross References in the OED [to following letters]



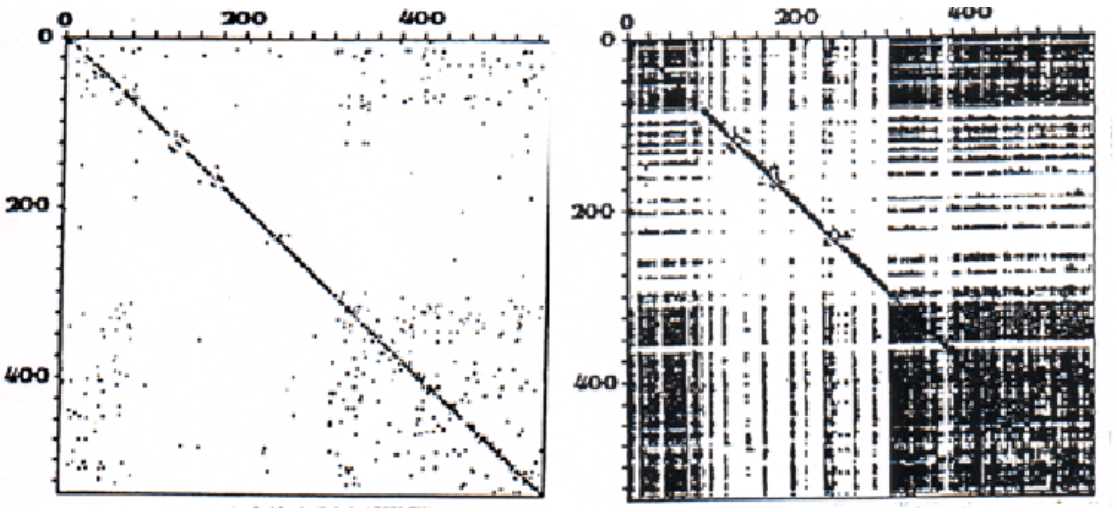
Transitive Closure

We can do some computing with a link matrix to understand its properties, the most important of which is "can you get there from here" or *reachability*

We can determine whether a path exists between any two nodes in a graph by calculating the *transitive closure*
of the graph; the most commonly used approach is Warshall's algorithm

Reachability and transitive closure analysis also makes for interesting games and visualizations

Reachability Analysis



Moving Beyond "Reachability"

We've been treating linking in purely structural terms - is one thing connected to another - but we can refine that into two perspectives:

- *Relational*
  analysis treats links as indicators of the amount of connectedness or the direction of flow between documents, people, groups, journals, disciplines, domains, organizations, or nations
- *Evaluative*
  analysis treats links as indicators of the level of quality, importance, influence or performance of documents, people, groups. ...

Bibilometrics (or "Scientometrics"): Structure of Scientific Citation

Analysis of scientific citation began in the 1920s as a way to quantify the influence of specific documents or authors in terms of their "impact factor"

It can also identify "invisible colleges" of scientists whose citations are largely self-referential

It can recognize the emergence of new scientific disciplines

It can measure (the "H-index")
the productivity and impact of a scientist or researcher (and predict Nobel Prize winners)

Garfield & de Solla Price

Eugene Garfield and Derek J. de Solla Price are two of the "founding fathers" of bibliometrics

Garfield's 1990 paper "The Most Cited Papers of All Time" reports a Zipf distribution for citation frequency

- A 1951 paper from the Journal of Biological Chemistry on "protein determination" was cited 187,652 times between 1951 and 1988!

de Solla Price's 1965 paper "Network of Scientific Papers" in the journal *Science* is also one the classic papers in the field

Citation Analysis and Journal Ranking

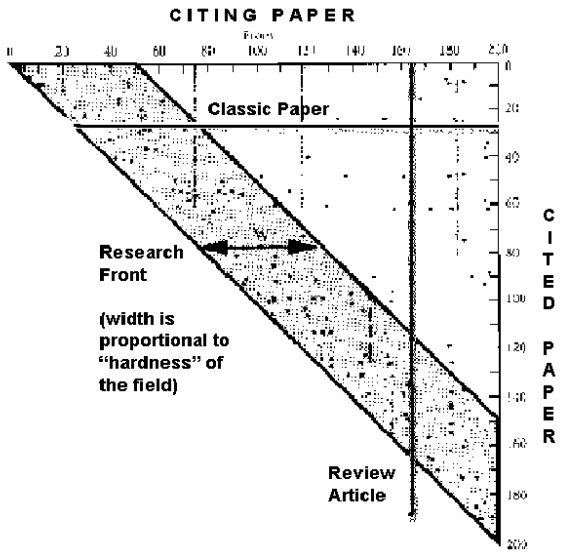
The concluding sentence of the de Solla Price article:

> ...a very large fraction of the alleged 35,000 journals now current must be reckoned as merely a distant background noise, and as very far from central or strategic in any of the knitted strips from which the cloth of science is woven.

Garfield devised the IMPACT FACTOR for journals by measuring the frequency with which the "average article" in a journal gets cited during a two-year period following its publication

But this has been widely criticized because it can be easily manipulated and because the short time window misses the impact of classic articles and others that were "ahead of their time"

Temporal Structure of Scientific Citation



Some Tricky Issues for Citation Analysis

How should "credit" be allocated for papers with multiple authors and institutions? Does each get full credit, or "adjusted" credit based on authorship order? Does the adjustment function depend on the relationships among the authors?

Should author self-citations be included in citation counts? Should they be weighted differently?

How should productivity (in terms of the number of papers authored) and citation analysis be combined in evaluating the impact or performance of an author?

Citation Signals and Polarity

A *citation signal*
indicates how a writer views the relationship of a citation to the text from which the citation is made

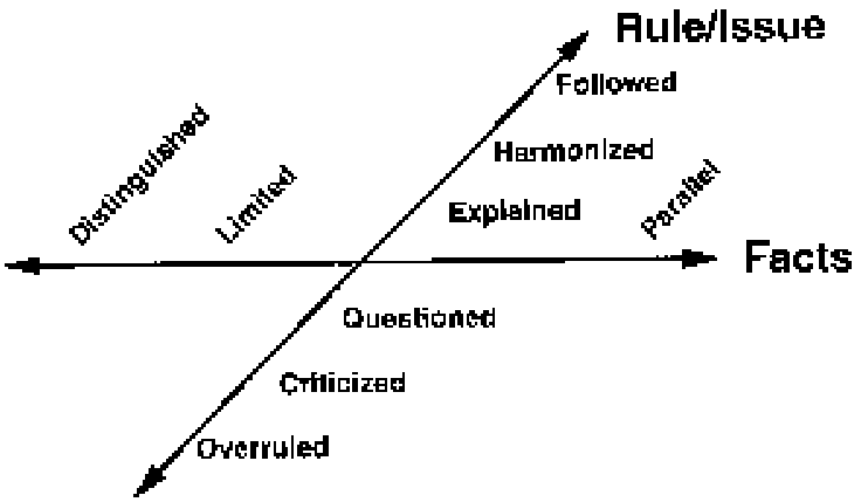
A citation or link without a signal suggests by default that the citation supports the current text

Explicit signals that indicate positive *polarity*
include "See," "See also," "See generally," and "Cf."

Signals that indicate negative polarity include "But see" and "Contra"

Shepardizing

Because of the use of legal cases as precedents, an entire industry in legal publishing exists to interpret any citations to cases; Shepard's is the oldest and checking a prior legal ruling to make sure it is still "good" is called "Shepardizing" a case



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

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

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

What value should we place on links from low importance or low relevance sites?

What value should we place on links to low importance or low relevance sites?

Should the number of outgoing links a site has influence how we value each link?

Google Page Rank

(from a classic paper by two Stanford graduate students)

If a web page A has N pages pointing to it, its Page Rank is:

PR(A) = (1-d) + d (*a proportion of the Page Rank of every page that links to it*)

- T1...Tn are the pages that point to A
- d is a "damping factor" usually set to 0.85.
- #Outgoing is the number of links from a page

This proportion is calculated recursively as follows:

$$\left(\frac{PageRank(T_1)}{\#Outgoing(T_1)} + \frac{PageRank(T_2)}{\#Outgoing(T_2)} + \ldots \frac{PageRank(T_n)}{\#Outgoing(T_n)} \right.$$

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. You don't give away your stock when you vote.

Note that the vote for the "linked to" site is divided by the number of outgoing links from the "voting" site

So is better for a site's Page Rank if it gets a vote from a site that only has a few of them, or to get a vote from a site that has a lot of them?

Weighting Page Rank

the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one

This quote from the Brin & Page paper is hinting that there is some weighting or scaling of the computed Page Ranks

So suppose that raw Page Ranks are then weighted using a logarithmic or other nonlinear function, giving most sites very low page ranks and a much smaller number of very important sites high page ranks

So is better for a site's Page Rank if it gets a vote from a site that only has a few of them, or to get a vote from a site that has a lot of them?

Page Rank and Relevancy

Because Page Rank measures the static structure of each web page, there is no concept of relevancy in the mathematical description of Page Rank, and 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

Following a link in the displayed results contributes to its subsequent ranking

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

"Google bombing" (or more generically, "link bombing") techniques attempt to associate negatively-connoted search terms with particular sites by using the search terms as anchor text; is this political activism, or web vandalism?

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... and interprets a link from page A to page B as a vote. But votes cast by pages that are themselves "important" weigh more heavily

The Page Rank algorithm favors older pages because a new page, regardless of its quality or relevance, will not have many incoming links

Similarly, pages from "big company" domains, with short URLs, and whose URLs contain the search terms are treated as more relevant

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

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

Is it "democratic" for the rich to have more votes and for their votes to count more? Or is this just "shareholder democracy" with "one dollar, one vote"

It may be democratic for the aggregated preferences of the majority to be put into practice, but is is democratic to allow only the majority's views to be heard?

Web Crawling

The size and unmanaged character of the web means there isn't and never will be an authoritative directory or catalog of its constituent pages

So a pre-requisite for any system that wants to index and search the web is to find the pages to index

This is harder than it seems

So even though the major search engines use more or less the same crawling techniques, they differ in how well they do it, and this in turn affects their indexing and retrieval approaches and performance

Web Crawling: Simplistic View

How do the web search engines get all of the items and links they use?

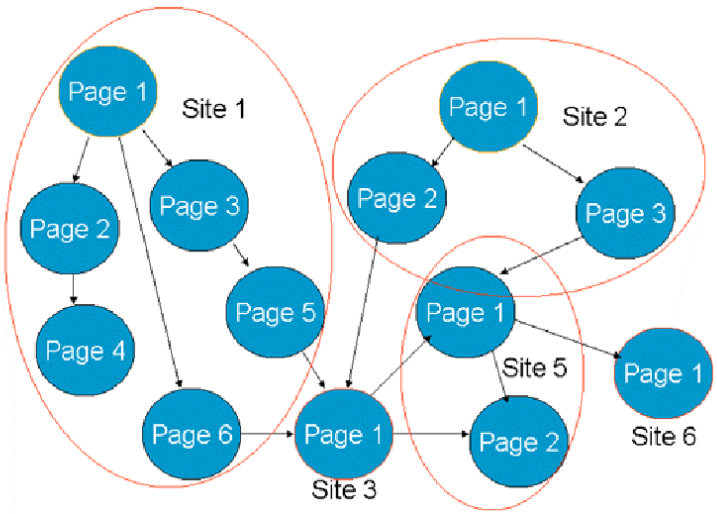
1. Start with known sites
2. Record information for these sites
3. Follow the links from each site
4. Record information from sites found by following links
5. Repeat

Crawling Order

Animations of breadth-first, depth-first, etc tree traversals at:
http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/SearchAnimations.htm

But the web is a graph, not a tree

Crawl Graphs, Not Trees



Web Crawling: Complications

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

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"

Readings for Lecture #26 on 26 November

Ron Bourret, "Native XML Databases in the Real World." XML 2005 (sections 1-6)

Introduction to Information Retrieval

- Chapter 10, "XML Retrieval" (skim or skip sections 10.3 and 10.4)