An Introduction to Google's PageRank

Search Engine Rankings

Kendall Giles

Department of Computer Science
Johns Hopkins University
and
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University

UCLA IPAM Tutorial, September 11, 2007

Outline

Outline

Everyone knows Google...

- founded in 1998 by Larry Page and Sergey Brin
- 2006 revenue: $10.6 billion
- 2006 net income: $3.07 billion
- 13,748 employees (as of June 2007)

because they are good at what they do.

Web Images Video News Maps Gmail more ▼ Sign in

Google university Search Advanced Search Preferences

New! View and manage your web history

Web Results 1 - 10 of about 514,000,000 for university [definition]. (0.08 seconds)

University - Wikipedia, the free encyclopedia
A university is an institution of higher education and research, which grants academic degrees at all levels (bachelor, master, and doctorate) in a variety ...
en.wikipedia.org/wiki/University - 74k - Cached - Similar pages

Welcome to Harvard University
The Harvard University Homepage: an overview of academic programs, campus life, resources, news and events, with extensive links to other web sites located ...
www.harvard.edu/ - 22k - Cached - Similar pages

University of Delaware
The University of Delaware, founded in 1743, offers over 100 academic majors; its distinguished faculty includes internationally known scientists, ...
www.udel.edu/ - 16k - Cached - Similar pages

Google University Search
Google's University Search enables you to narrow your search to a specific school website. Try it for things like admissions information, course schedules, ...
www.google.com/options/universities.html - 100k - Cached - Similar pages

University of Virginia Home Page
The University of Virginia in Charlottesville, VA was founded in 1819 by Thomas Jefferson. The cornerstone of the University's first building was laid in ...
www.virginia.edu/ - 35k - Cached - Similar pages

Yale University
Yale University comprises three major academic components: Yale College (the undergraduate program), the Graduate School of Arts and Sciences, ...
www.yale.edu/ - 14k - Cached - Similar pages

Boston University
An independent, co-educational, and non-sectarian institution of higher education and research located along the banks of the Charles River in Boston, ...
www.bu.edu/ - 11k - Cached - Similar pages

The University of Texas at Austin - Web Central
A comprehensive university with a broad mission of undergraduate and graduate education, research, and service to society. Enrollment of over 48000.
www.utexas.edu/ - 21k - Cached - Similar pages

U.S. Universities, by State
A list of regionally-accredited US universities organized by state.
www.utexas.edu/world/univ/state/ - 187k - Cached - Similar pages

Stanford University
Stanford University is one of the world's leading research and teaching institutions. It is located in Palo Alto, California.
www.stanford.edu/ - 18k - Cached - Similar pages

Sponsored Links

University of Phoenix
Get a high quality education online at University of Phoenix.
www.phoenixdegrees.com

Find an Online School
Get connected with online schools and continue your education.
www.exchangeplace.com

Online University Degrees
Earn a Degree in as few as 2 yrs! Find the Right University Degree.
www.University.ClassesUSA.com

Online Degree Programs
Take classes online this fall. Get all the information that you need.
www.universityofphoenix-online.com

Schools in Your Area
Find a School Close to You. Directions, Maps & Local Search.
MapQuest.com

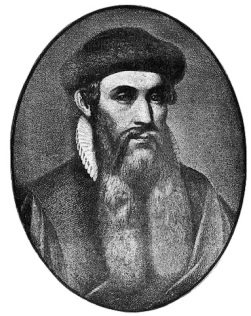
Find Top Online Schools
Request Info from The Top Online Schools. Financial Aid Available.
www.EducationDegreeSource.com

Earn over $60K a year
Average salary of our graduates. Get a B.S. degree in only 24 months
www.Neumont.edu

Top Online Universities
Search 6500+ Online Courses. Find Online Colleges in 60 seconds.
www.eLearners.com/OnlineUniversity

A Progression of Technologies

printing press: Johann Gutenberg, 1450



public library: Benjamin Franklin, 1731

WWW: Sir Tim Berners-Lee, 1989

photo: Uldis Bojars

Traditional Information Retrieval

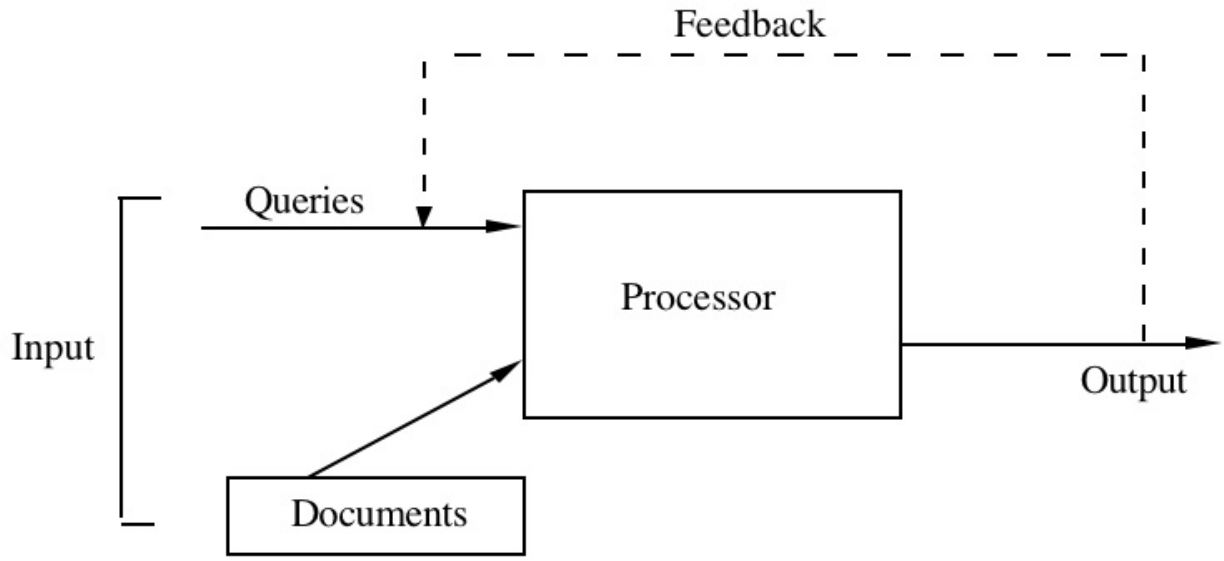


figure: Information Retrieval, by C. J. van Rijsbergen

- nonlinked documents
- collection is mostly static
- organized and categorized by specialists

Information Retrieval Approaches

find results based on *content* of the documents

- *boolean models*
  - use boolean logic to frame returned document sets
  - example: (*t1* AND *t2*) OR *t3*
- *vector space models*:
  - each document represented by keywords, with (doc_i, $keyword_j$) given some weight
  - document proximity measured with a similarity or dissimilarity score (e.g., $\cos\theta = XY/(||X||||Y||)$)
  - example: Latent Semantic Indexing
- *probabilistic models*:
  - estimate the probability a document is relevant to a query term

Simple Reverse Index

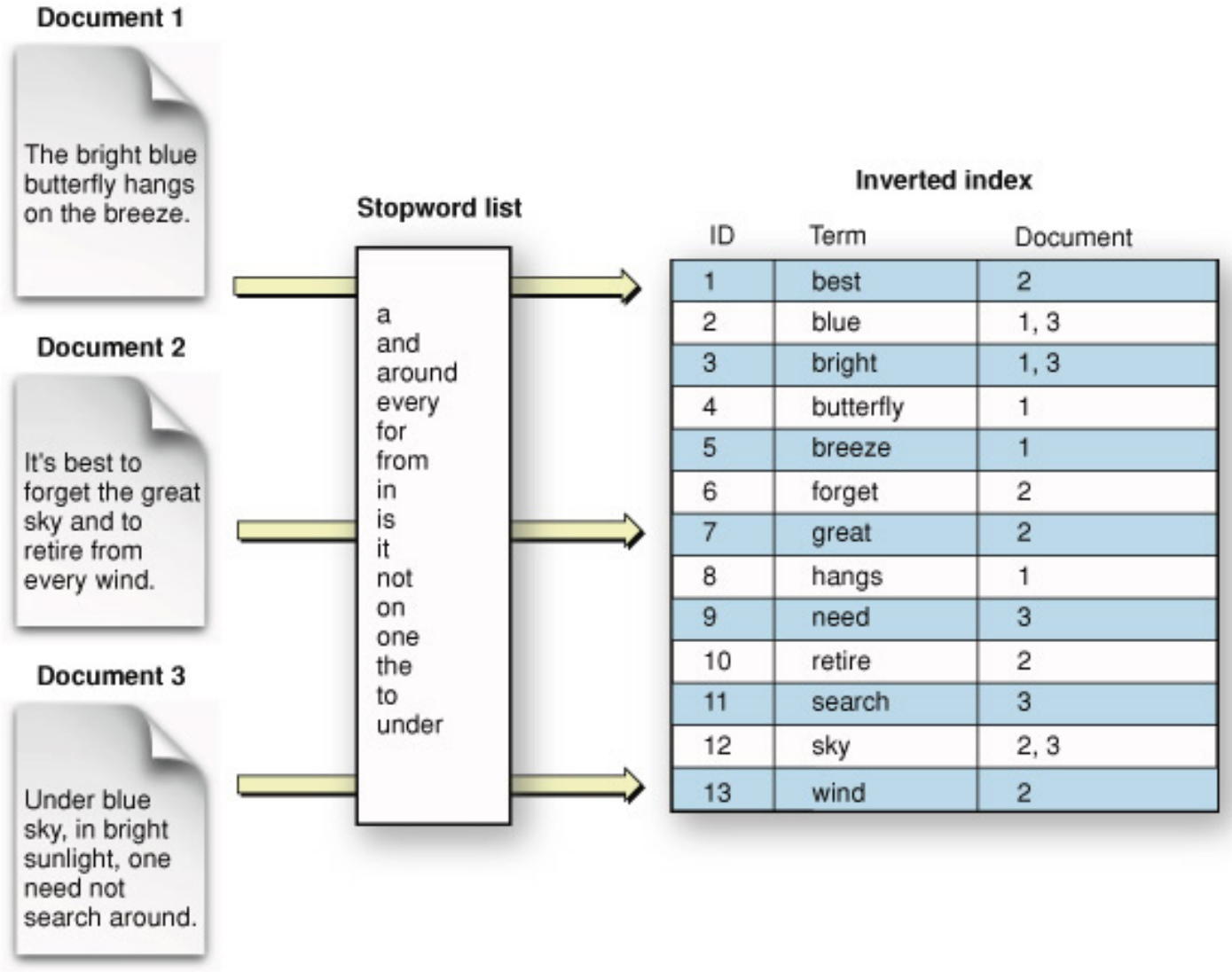
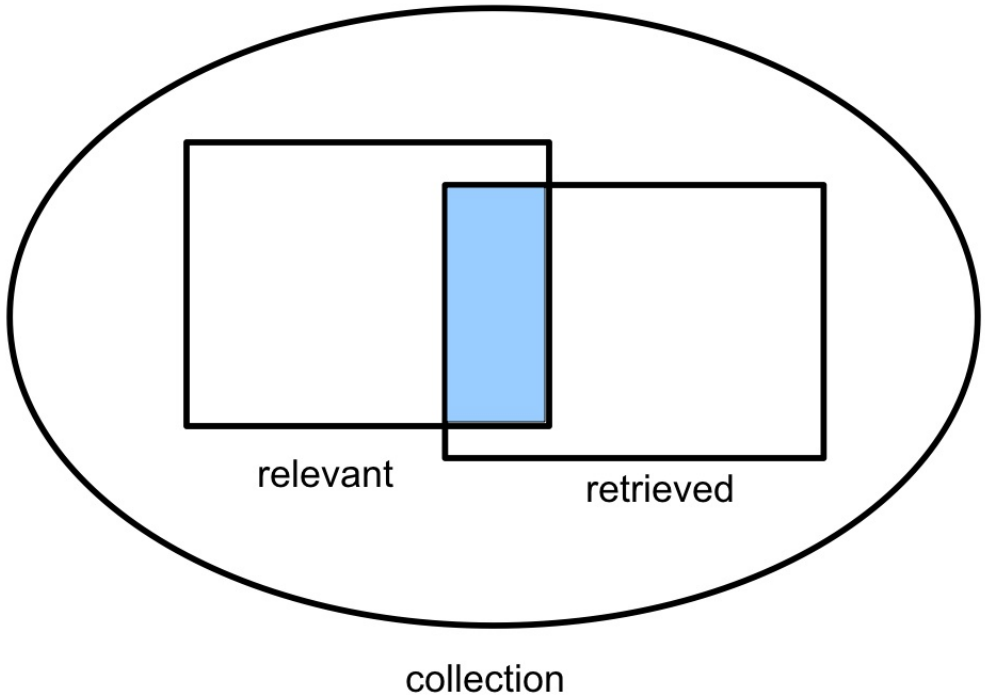


figure: Apple Search Kit Programming Guide

Comparing Search Engines.

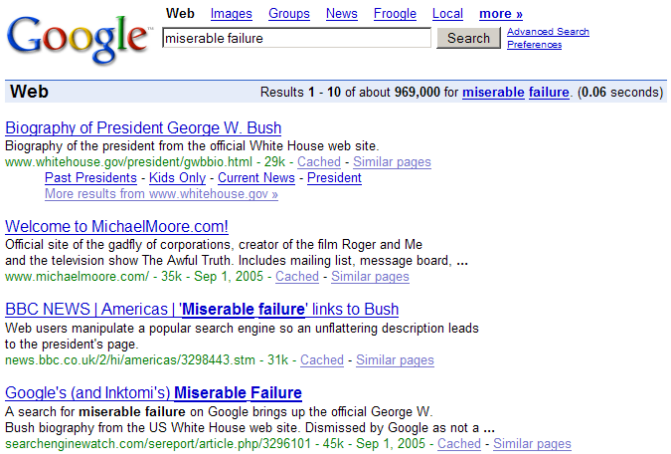


- precision: $\frac{|\{relevant\}|\cap|\{retrieved\}|}{|\{retrieved\}|}$
- recall: $\frac{|\{relevant\}|\cap|\{retrieved\}|}{|\{relevant\}|}$

Outline

Features of the Web Search Environment

- web is *large*: estimate > 22.9 billion pages
- web is *hyperlinked*: e.g., click here
- web is *self-organized*: no central authority; multiple document formats, languages
- web is *dangerous*: spammers, *malicious hackers*, *gaming*
- web is *dynamic*: large percentage of pages change content and links daily



How to Rank Web Pages?

Some (content) ideas:

- query words appear on the page
- number of times query word appears
- distance on page between multiple query words

Some (other) ideas:

- location of query words on page (e.g., <title>, <h1>, etc.)
- web page reputation
- update history of page

Some (structure) ideas:

- forward links
- backlinks

Resulting ranking formula: some function of the above

The Web is Like High School: the importance of being popular

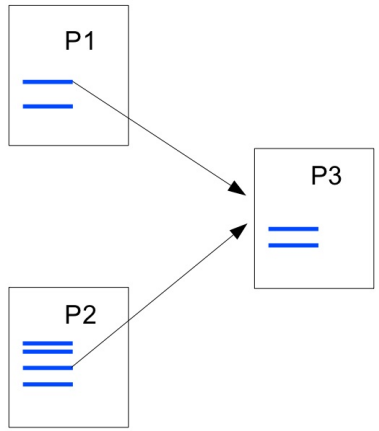


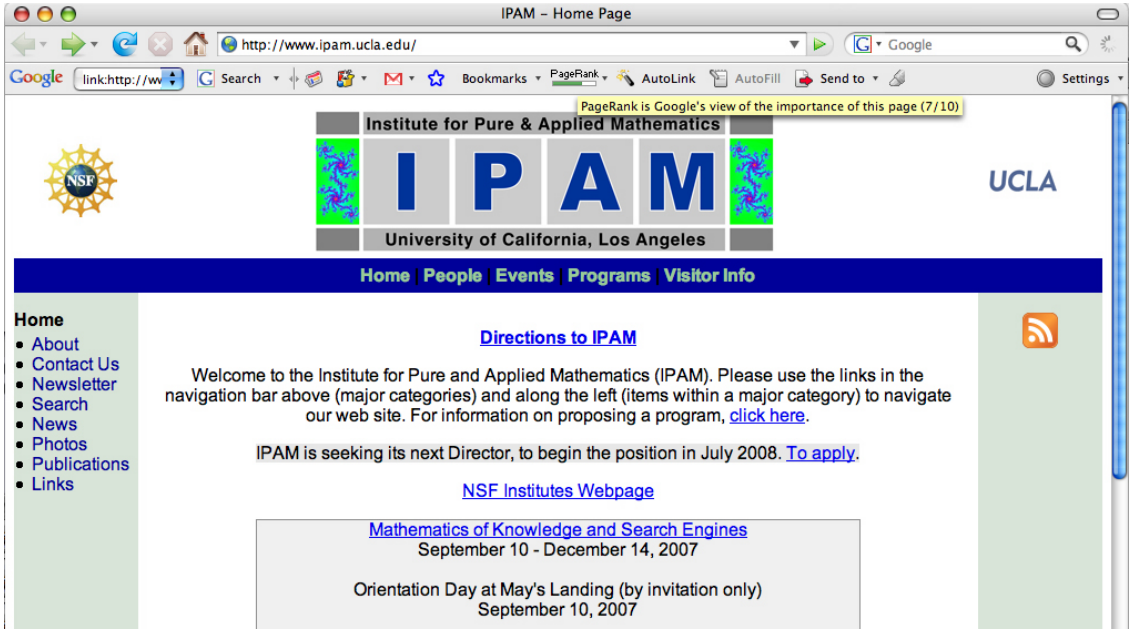
figure adapted from Sergey and Brin

- page score = f(content score, structure score, other score)
- link from P1 to P3 means that P1 gives some measure of relevance to P3
- a page with more backlinks is somehow more important or popular than a page with less
- status of recommender is also important

Outline

Basic Intuition of PageRank

a webpage is important if it is pointed to by other important pages



- www.kendallgiles.com: 2/10
- www.google.com: 10/10
- spammers or new pages: PR0 or not enough information

a page is ranked high if the sum of the ranks of its backlinks is high

Formalizing Our Intuitive Notion of Webpage Importance.

- Let u be a webpage
- Let F_u be the set of pages u points to (forward links)
- Let B_u be the set of pages that point to u (backlinks)
- Let $N_u = |F_u|$ be the number of links from u
- Let r be a simple ranking function

$$r(u) = \sum_{v \in B_u} \frac{r(v)}{N_v}$$

Computing $r(u)$ Iteratively.

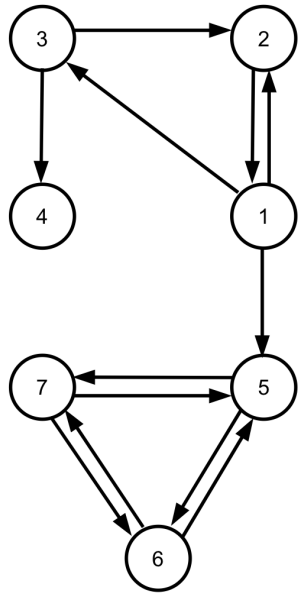
If n is the number of webpages, and we let

$$r_0(u) = 1/n$$

then

$$r_{k+1}(u) = \sum_{v \in B_u} \frac{r_k(v)}{N_v}$$

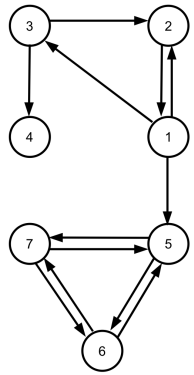
An Example of the Basic Ranking Function



example adapted from Langville and Meyer

initial	iteration$_1$	iteration$_2$	rank
$r_0(1) = .143$	$r_1(1) = .143$	$r_2(1) = .119$	4
$r_0(2) = .143$	$r_1(2) = .119$	$r_2(2) = .063$	5
$r_0(3) = .143$	$r_1(3) = .048$	$r_2(3) = .040$	6
$r_0(4) = .143$	$r_1(4) = .024$	$r_2(4) = .019$	7
$r_0(5) = .143$	$r_1(5) = .190$	$r_2(5) = .212$	1
$r_0(6) = .143$	$r_1(6) = .167$	$r_2(6) = .195$	3
$r_0(7) = .143$	$r_1(7) = .179$	$r_2(7) = .204$	2

A Matrix Representation of the Rank Formula.



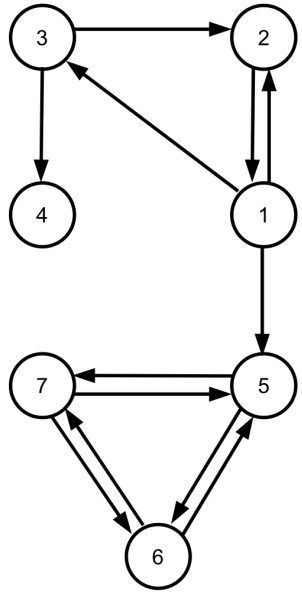
H: $n \times n$ hyperlink matrix
π^T: $1 \times n$ vector of rank values

$$\mathbf{H} = \begin{pmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Basic rank formula:
$\pi^{(k+1)T} = \pi^{(k)T} H$

Outline

Some Problems



- the subgraph of u_5, u_6, u_7 forms a rank sink.
- u_4 is a dangling node. Dangling nodes result in $\mathbf{0}^T$ rows in $\mathbf{H}$.
- will this process in general converge? converge to a unique ranking? does convergence depend on the starting vector $\pi^{(0)T}$?

Fix for Dangling Nodes

Motivation: *random surfer model*

1.) Replace $\mathbf{0}^T$ rows of $\mathbf{H}$ with $1/n\ \mathbf{e}^T$—>$\mathbf{H}$ is now stochastic.

$$\mathbf{S} = \mathbf{H} + 1/n\ \mathbf{a}\ \mathbf{e}^T$$

$$\mathbf{S} = \begin{pmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Fix for Rank Sinks

2.) Define $\alpha \in (0, 1)$ as a teleportation parameter.

This gives us the Google matrix:

$$\mathbf{G} = \alpha \mathbf{S} + (1 - \alpha) 1/n \, \mathbf{e} \, \mathbf{e}^T$$

Note: Google initially chose $\alpha = .85$.

Properties of Google matrix.

$$\mathbf{G} = \alpha \mathbf{S} + (1 - \alpha) 1/n \, \mathbf{e} \, \mathbf{e}^T$$

G is:

- stochastic
- irreducible
- aperiodic
- primitive
- dense

$$\begin{aligned} \mathbf{G} &= \alpha \mathbf{S} + (1 - \alpha) 1/n \, \mathbf{e} \, \mathbf{e}^T && (1) \\ &= \alpha (\mathbf{H} + 1/n \, \mathbf{a} \, \mathbf{e}^T) + (1 - \alpha) 1/n \, \mathbf{e} \, \mathbf{e}^T && (2) \\ &= \alpha \mathbf{H} + (\alpha \mathbf{a} + (1 - \alpha) \, \mathbf{e}) 1/n \, \mathbf{e}^T && (3) \end{aligned}$$

The PageRank Equation

$$\pi^{(k+1)T} = \pi^{(k)T}\mathbf{G}$$

$$\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)1/n\,\mathbf{e}\,\mathbf{e}^T \tag{4}$$

$$= \alpha(\mathbf{H} + 1/n\mathbf{a}\mathbf{e}^T) + (1-\alpha)1/n\,\mathbf{e}\,\mathbf{e}^T \tag{5}$$

$$= \alpha\mathbf{H} + (\alpha\mathbf{a} + (1-\alpha)\,\mathbf{e})1/n\,\mathbf{e}^T \tag{6}$$

- **H**: sparse hyperlink matrix
- **S**: sparse stochastic matrix
- **G**: dense stochastic, primitive matrix
- **E**: dense teleportation matrix
- n: number of pages
- α: scaling parameter
- π^T: stationary PageRank vector
- $\mathbf{a}^T$: binary dangling node vector

Example PageRank Calculation

$$\mathbf{G} = \begin{pmatrix} 0.021 & 0.305 & 0.305 & 0.021 & 0.305 & 0.021 & 0.021 \\ 0.871 & 0.021 & 0.021 & 0.021 & 0.021 & 0.021 & 0.021 \\ 0.021 & 0.446 & 0.021 & 0.446 & 0.021 & 0.021 & 0.021 \\ 0.143 & 0.143 & 0.143 & 0.143 & 0.143 & 0.143 & 0.143 \\ 0.021 & 0.021 & 0.021 & 0.021 & 0.021 & 0.446 & 0.446 \\ 0.021 & 0.021 & 0.021 & 0.021 & 0.446 & 0.021 & 0.446 \\ 0.021 & 0.021 & 0.021 & 0.021 & 0.446 & 0.446 & 0.021 \end{pmatrix}$$

$$\pi^T = (0.093, 0.077, 0.054, 0.050, 0.254, 0.236, 0.236)$$

ranks: $(u_5 u_6 u_7 u_1 u_2 u_3 u_4)$

Exact Form of the Current PageRank Equation?

Google is not saying.

Outline

α

$0 < \alpha < 1$

$$\mathbf{G} = \alpha\mathbf{S} + (1 - \alpha)1/n\,\mathbf{e}\,\mathbf{e}^T$$

- hyperlink structure vs teleportation matrix **E**
- speed of eigenvector convergence: $-\tau/log_{10}\alpha$ iterations
- $\alpha \to 1$: more sensitive to structure changes; slower convergence
- $\alpha \to 0$: more insensitive to structure changes; faster convergence.

H

Random surfer model $\longrightarrow$ *Intelligent surfer model*

$$\mathbf{H} = \begin{pmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

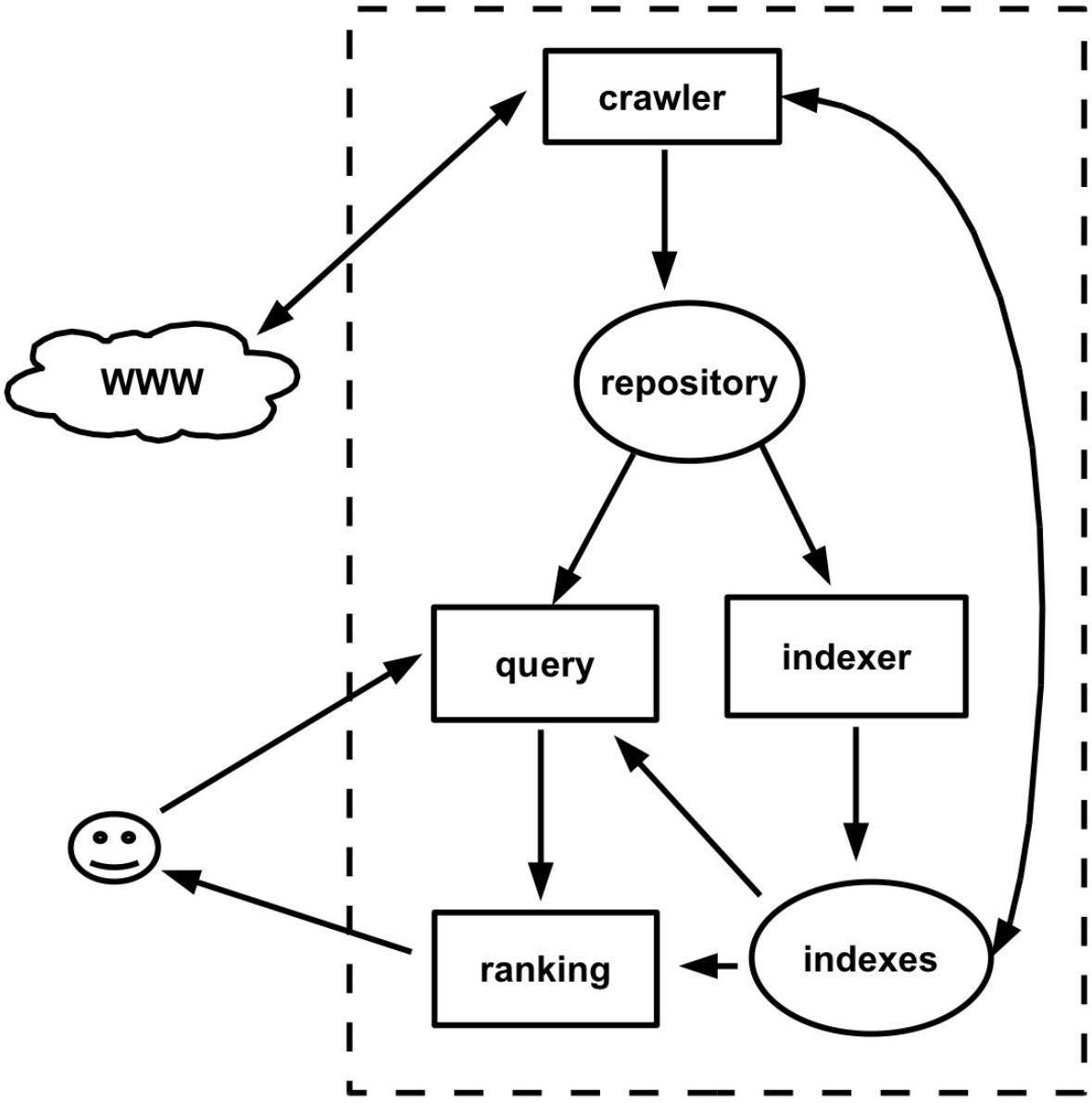
$$\mathbf{H} = \begin{pmatrix} 0 & 1/8 & 3/8 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

E

- democratic: $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)1/n \;\; \mathbf{e}\,\mathbf{e}^T$
- personal: $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\,\mathbf{e}\,\mathbf{v}^T$

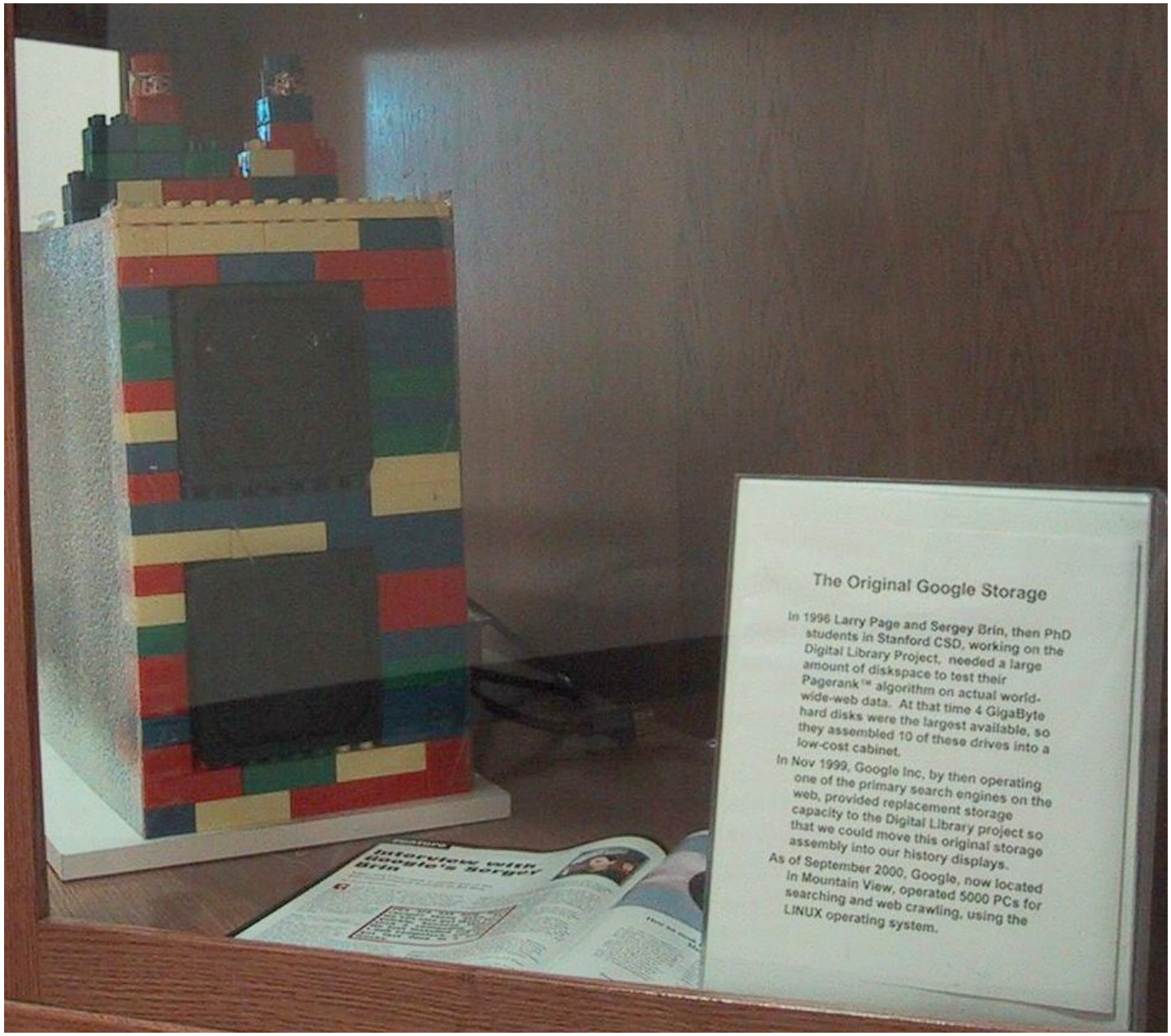
Outline

Web Search Components



Outline

Original Google Servers.



Storage Issues

Entity	Description	Storage
$\mathbf{H}$	sparse hyperlink matrix	# nonzeroes in $\mathbf{H}$ double
$\mathbf{a}$	sparse binary dangling node vector	|D| integers
$\mathbf{v}^T$	dense personalization vector	n doubles
$\pi^{(k)T}$	dense iterate of PageRank power method	n doubles

table: Langville and Meyer

- $\mathbf{H}$ is the largest component to be stored
- if $\mathbf{H}$ cannot be stored in main memory:
  - compress $\mathbf{H}$ so it fits in main memory and adjust algorithms
  - develop I/O-efficient algorithms
- Other graph compression methods available

Convergence Issues

Power Method stopping criteria:
$||\boldsymbol{\pi}^{(k+1)T} - \boldsymbol{\pi}^{(k)T}||_1 < \tau$

- Haveliwala noted that correct ordering is more important than correct π values
- stop when the ordering converges—sooner than PageRank vector value convergence
- many versions of this idea

Accuracy

- π^T is a probability vector
- follows Zipf distribution—long tail probabilities must be distinguishable
- not aware of Google convergence tests
  - could sacrifice accuracy for speed
  - eigen structure of iteration matrix could have large eigengap

Thank you.

An Introduction to Google's PageRank:
Search Engine Rankings

Kendall Giles

kgiles@cs.jhu.edu

www.kendallgiles.com

JOHNS HOPKINS UNIVERSITY

Institute for Pure & Applied Mathematics
IPAM
University of California, Los Angeles

VCU

For Further Reading:

A. Langville and C. Meyer.
Google's PageRank and Beyond.
Princeton University Press, 2006.

S. Brin and L. Page.
The Anatomy of a Large-Scale Hypertextual Web Search Engine.
Computer Networks and ISDN Systems, 33:107–117, 1998.

L. Page, S. Brin, R. Motwani, and T. Winograd.
The PageRank Citation Ranking: Bringing Order to the Web.
Technical Report, Stanford University. 1998.