The PageRank Citation Ranking: Bringing Order to the Web

by

Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd

presented by

Martin Klein, Santosh Vuppala
{mklein, svuppala}@cs.odu.edu

ODU, Norfolk, 01/31/2007

Outline

- Background
- PageRank
- Implementation
- PageRank's Convergence
- Searching and other Applications
- Discussion

Background - Authors

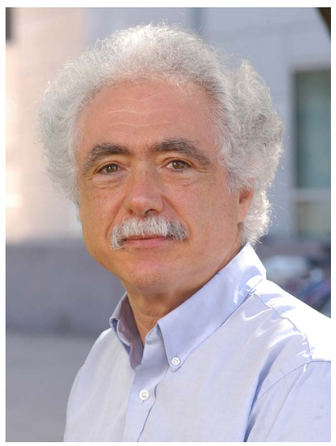
- Larry Page (~Rank)
  - BS in CE from UMich, MS from Stanford

- Sergey Brin
  - BS in Math&CS from UMD, MS from Stanford
- Google Inc. in 09/98 (google.com - 09/97)

figures from:http://www.google.com/corporate/execs.html

Background - Authors

- Rajeev Motwani
  - Ph.D 1988, CS, UC Berkeley
  - Professor at Stanford U



- Terry Winograd
  - Ph.D. 1970, M.I.T, Applied Mathematics
  - Professor at Stanford U

figures from: http://theory.stanford.edu/~rajeev/ and http://hci.stanford.edu/winograd/

Background - Paper

- Stanford WebBase project (1996 - 1999)
  http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/
  http://dbpubs.stanford.edu:8091/diglib/

- funded by NSF through DLI1
  http://www.dli2.nsf.gov/dlione/

"The Initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways." quote from the DLI1 website

Background - Paper

- it is a technical report! (working paper)
  (Stanford Digital Libraries SIDL-WP-1999-0120)
- from the paper: web size = 150M web pages
- 2005: Google claims to index more than 8B pages
  (http://blog.searchenginewatch.com/blog/041111-084221)
- 11.5B overall (http://www.cs.uiowa.edu/~asignori/web-size/)

PageRank - Motivation

"The average web page quality experienced by a user is higher than the quality of the average web page. This is because the simplicity of creating and publishing web pages results in a large fraction of low quality web pages that users are unlikely to read."

ex #1

ex #2

- Differentiate Pages
- Relative Importance
- Ranking/Search

quote taken from the paper

PageRank - Basics

- based on link structure of the web
- pages = nodes && links = edges
- forward links = outedges
- backlinks = inedges
- A and B are Backlinks of C

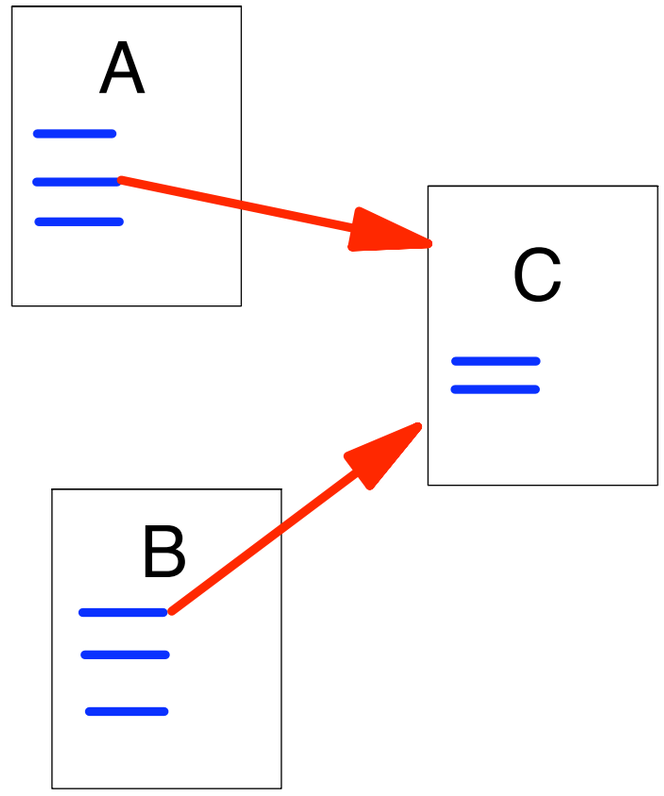


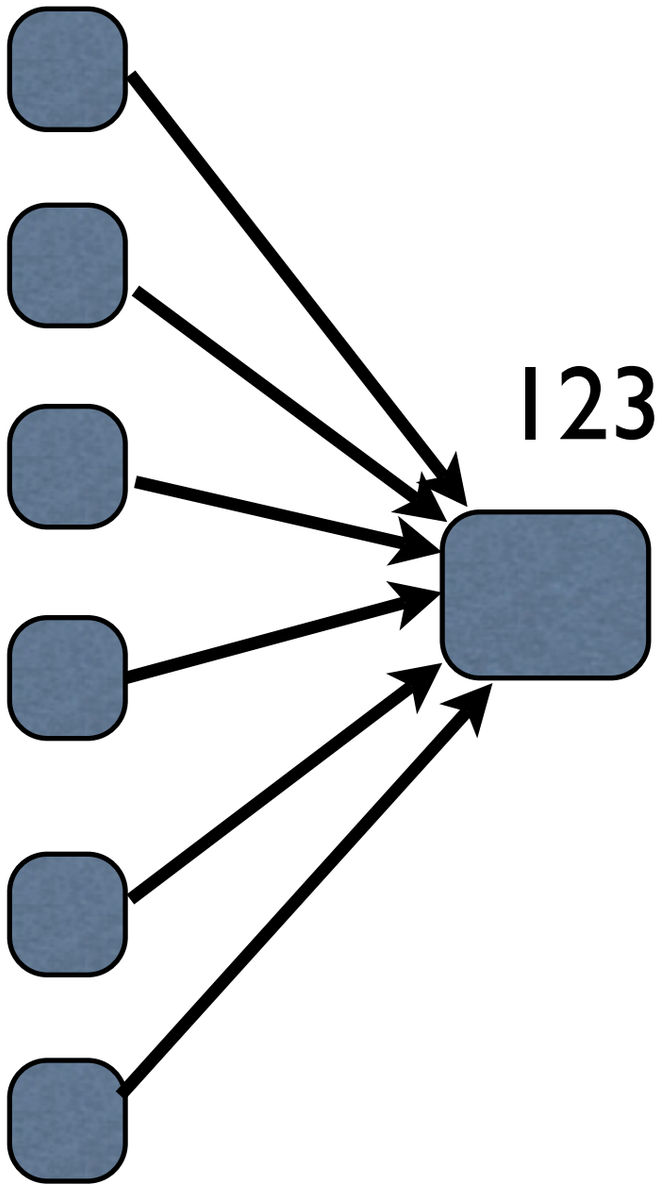
figure taken from the paper

PageRank - Assumptions

- a link from page A to page B is a vote from A to B
- highly linked pages are more “important” than pages with few links
- backlinks from high PR-pages count more than links from low PR-pages
- combination of PR and text-matching techniques result in highly relevant search results

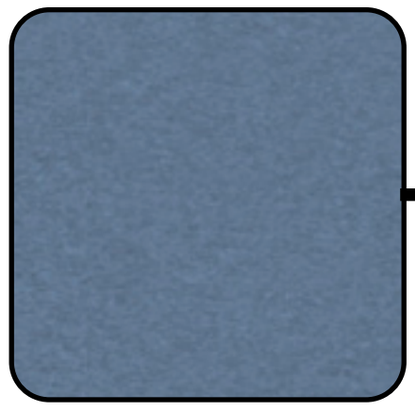
PageRank - Assumptions

p1-p6.info



123.info

cnn.com



abc.com

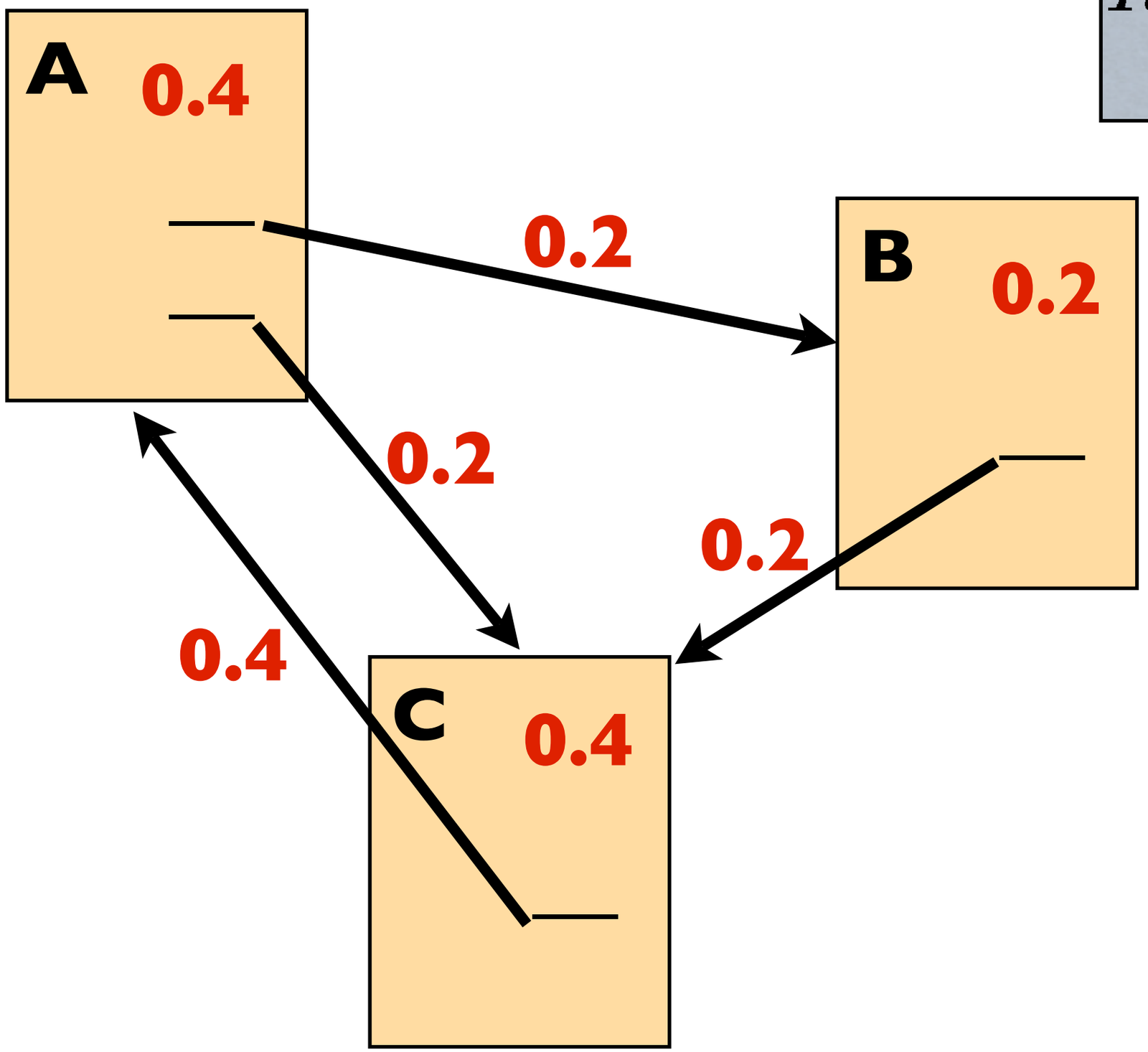
PageRank - Definition

- *u* is a web page
- F_u = set of pages *u* points to
- B_u = set of pages pointing to *u*
- c = normalization factor
- N_u = |F_u|

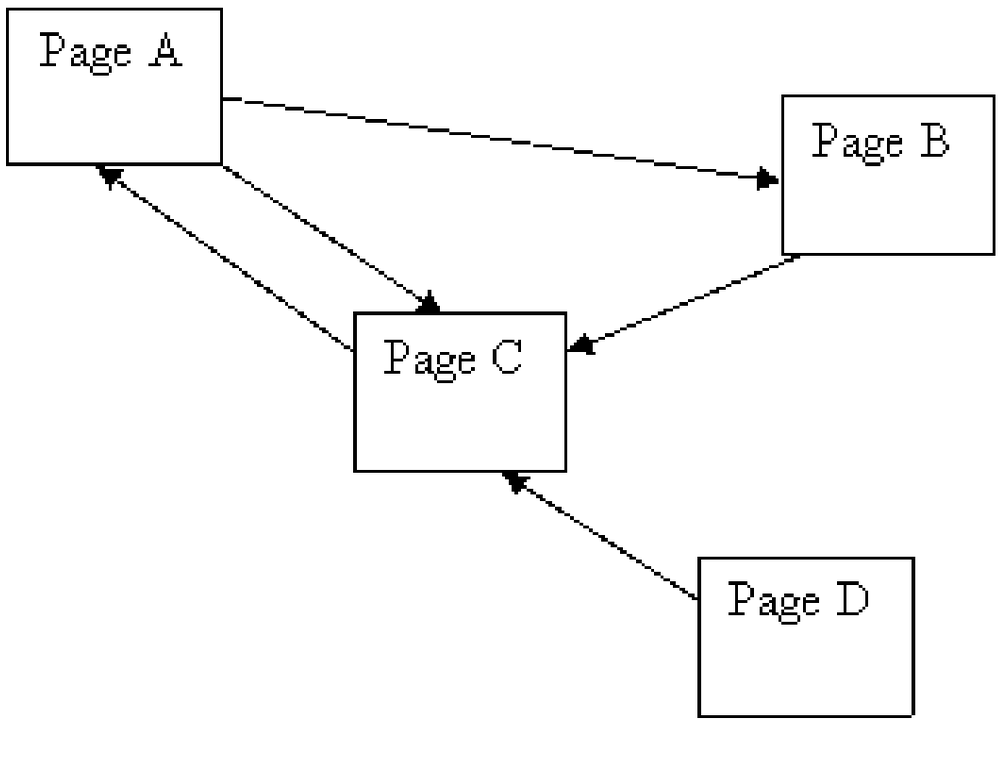
$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

PageRank - Example

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



PageRank - Iteration Example



d=0.85

Iteration 1
PR = 1 for all nodes

$$PR(A) = 1 - d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right)$$

Iteration 2	Iteration 3	Iteration 4		Iteration 10
PR(A)=1.85	PR(A)=1.8653	PR(A)=1.568		PR(A)=1.024
PR(B)=1.7225	PR(B)=1.735	PR(B)=1.4828	...	PR(B)=1.0204
PR(C)=4.036	PR(C)=3.3377	PR(C)=2.8706		PR(C)=2.057
PR(D)=0.15	PR(D)=0.15	PR(D)=0.15		PR(D)=0.15

figures from:http://www.iprcom.com/papers/pagerank/ and http://en.wikipedia.org/wiki/Pagerank

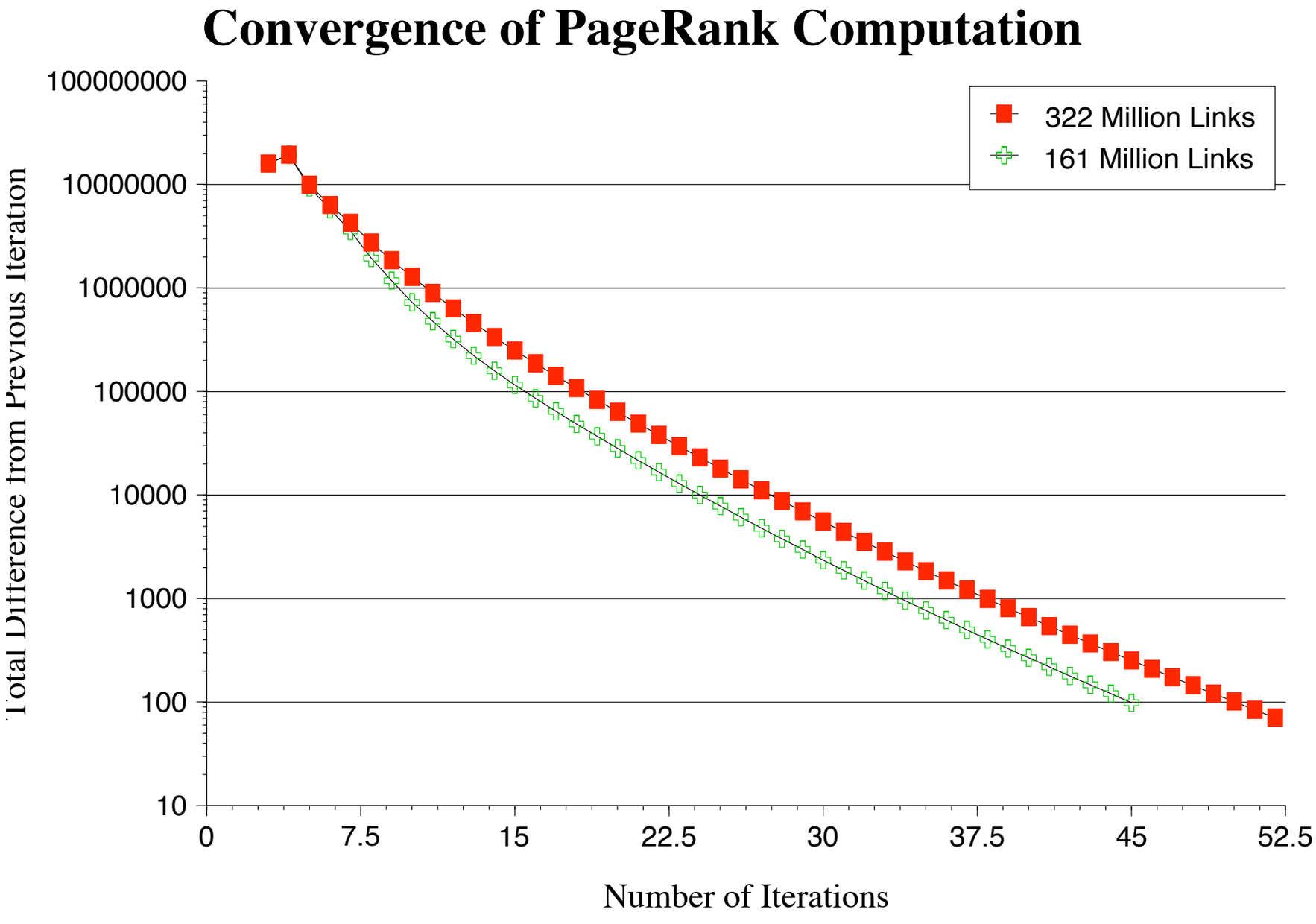
PageRank - Definition

What if two pages only link to each other and some page points to one of them?

- this loop/trap is called rank sink
- based on random surfer model
  - *E* - probability that a user visits a page

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

Convergence



- PR computation converges very quickly
- scales very well

Implementation

- built a crawling and indexing system
- repository size: 24M web pages (over 75M unique URLs)
- web crawler keeps index of links
- computing PR of entire repository takes ~5h
- issues: volume(!!!), incorrect HTML, dynamics of the web, page exclusion (robots.txt)

Search - Background

- title search and full text search (Google)
- ex.: title search
- 16M pages
- returns pages where title contains all query words

Title Search

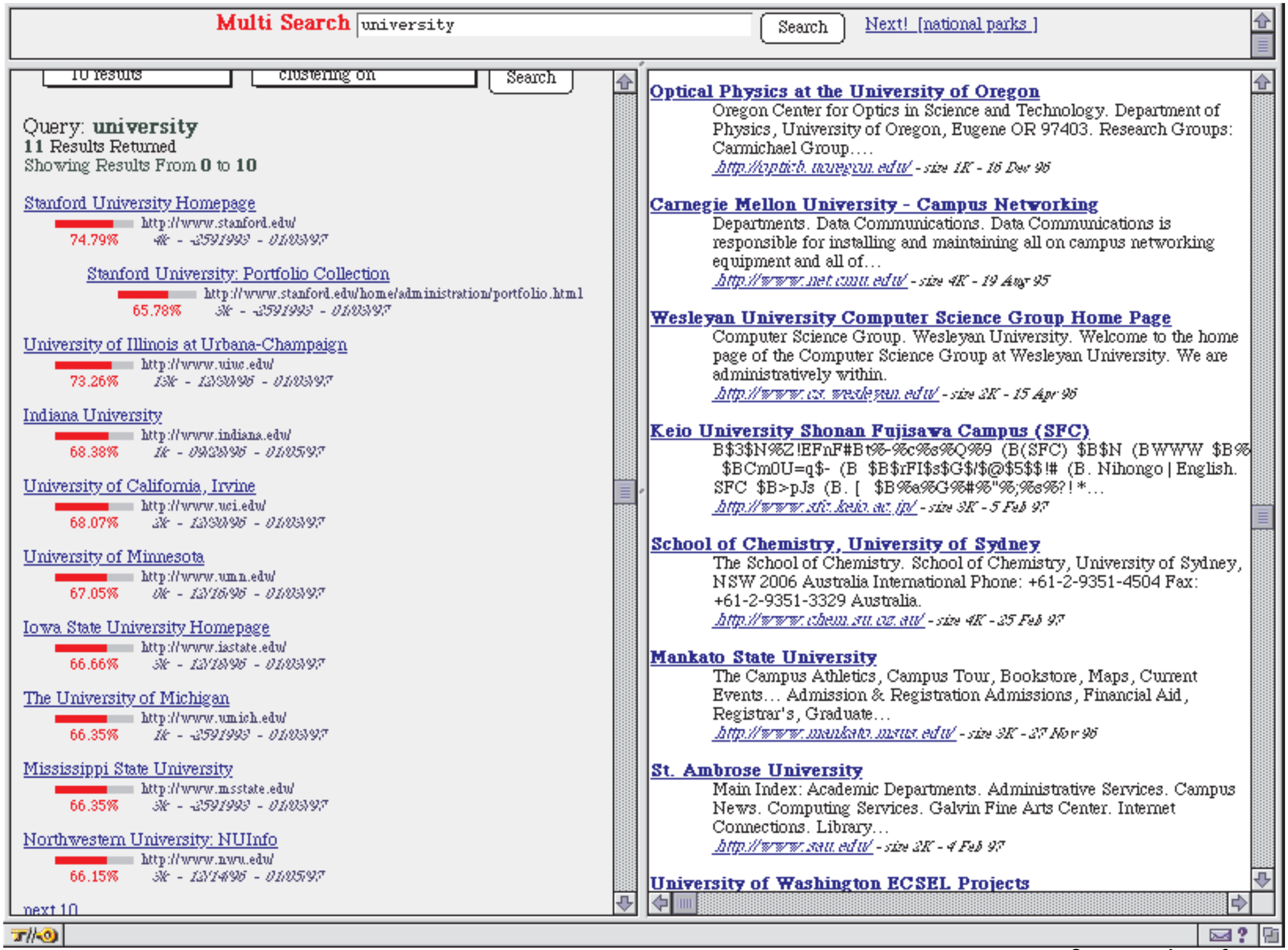


figure taken from the paper

Search - The Common Case

- page with high usage
- PR handles CC queries well
- CC for “wolverine” - U Michigan software system
- else: wiki page, imdb, etc

“It is important to note that the goal of finding a site that contains a great deal of information about wolverines is a very different task than finding the common case wolverine site.”

quote taken from the paper

Personalized PageRank

- E vector - distribution of web pages a random surfer jumps to
- usually E is uniform over all web pages (democratic)
- apply E just for one web page results in high PR value for relevant pages regarding the applied page
  - e.g. apply E for web page of faculty from cs@odu results in high PR for CS related pages

Other Uses of PageRank

- estimating web traffic - compare web page access from proxy vs PR
- PR as backlink predictor
  - efficient web crawling - better docs first
    - PR outperforms citation counts b/c number of citation count is not known in advance
- the PR proxy - annotate links with PR value
- PR is applied to the binary directed network model which is one of the methods used to model the co-authorship networks in relevance to digital libraries

Unwanted Uses of PageRank

- bmw.de banned from google in early 2006 due to its doorway page
  ~ is a page stuffed full of keywords that the site feels a need to be optimized for
  blog: http://blog.outer-court.com/archive/2006-02-04-n60.html

- "If an SEO creates deceptive or misleading content on your behalf, such as doorway pages or 'throwaway' domains, your site could be removed entirely from Google's index." *unknown at Google*

- google's webmaster helpcenter:
  http://www.google.com/support/webmasters/bin/answer.py?answer=35291

Unwanted Uses of PageRank

- "Google Bomb"
  http://searchengineland.com/070125-230048.php

  - create lots of links to one certain destination
  - label all of them with the same remarkable terms
  - query Google for those terms and you will get the linked page

Miserable Failure

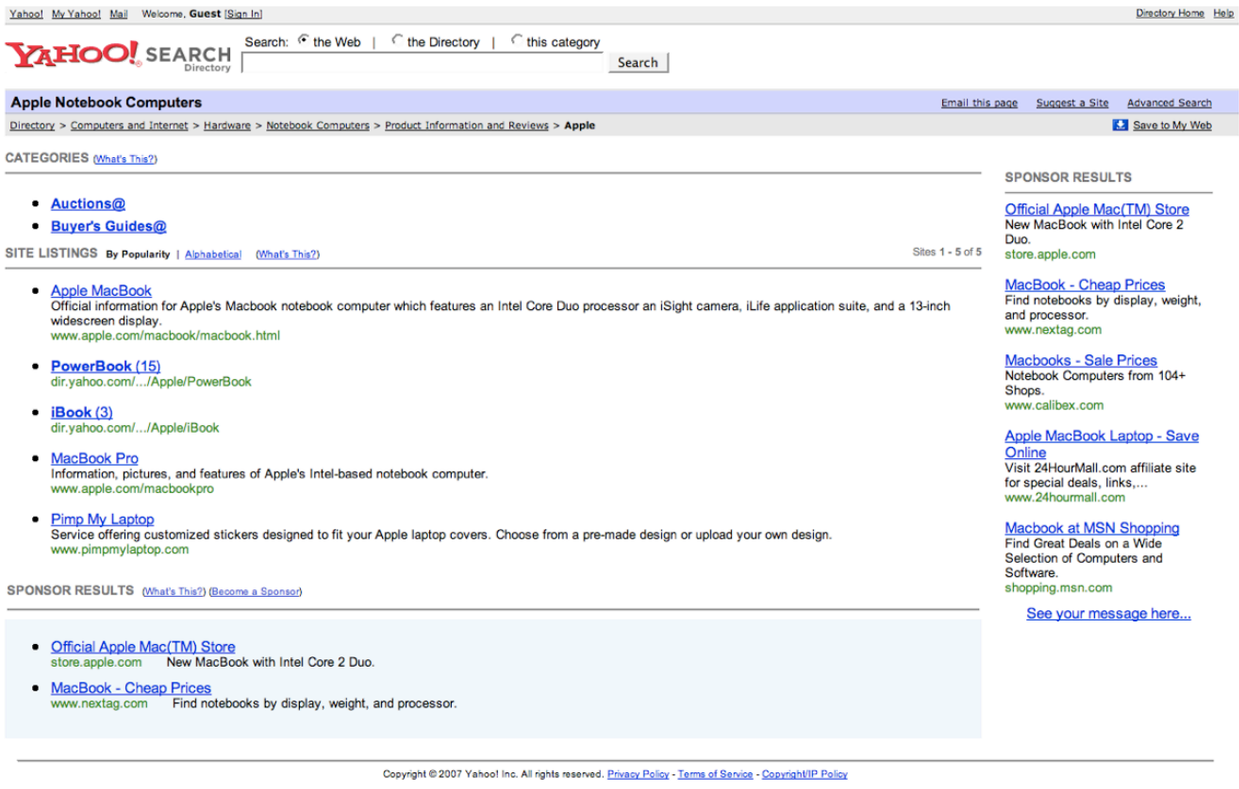
Discussion

Question 1:
PageRank is not optimal! How can it be improved? What can be changed?

Question 2:
Do you think, not publishing the PR value (Google Toolbar) would make it difference in the quest for obtaining a high PR value?

Question 3:
Considering the responsibility Google as a Search Engine has (as a prime source of information), should PageRank plus Google's additional "Ranking-VooDoo" not be more transparent to the public?


Store
iPhone
iPod + iTunes
.Mac
QuickTime
Support
Mac OS X
Hot News
Get a Mac
Hardware
Software
Made4Mac
Education
Pro
Mac@work
Developer
Where to Buy
Put some color on. iPod shuffle
Hot News Headlines
Mac OS X Tip of the Week: Search by Colors
Introducing iPhone
Macworld Keynote
Presenting tv
New Mac Ads
Visit an Apple Store
Search Apple.com
1-800-MY-APPLE | Find a Reseller | Site Map | RSS Feeds | Job Opportunities | Media Info | Terms of Use | Privacy Policy | Contact Us
Redeem iTunes Gift Cards (iTunes 7 required)
iBook G4 and PowerBook G4 Battery Recall
Copyright © 2007 Apple Inc. All rights reserved.
Apple.com Worldwide



http://www.yahoo.com

http://dir.yahoo.com/Computers_and_Internet/Hardware/Notebook_Computers/Product_Information_and_Reviews/Apple/

References

websites:
http://www.google.com/corporate/execs.html
http://www.google.com/corporate/index.html
http://www.iprcom.com/papers/pagerank/
http://www.webworkshop.net/pagerank.html
http://en.wikipedia.org/wiki/PageRank

and many more papers....

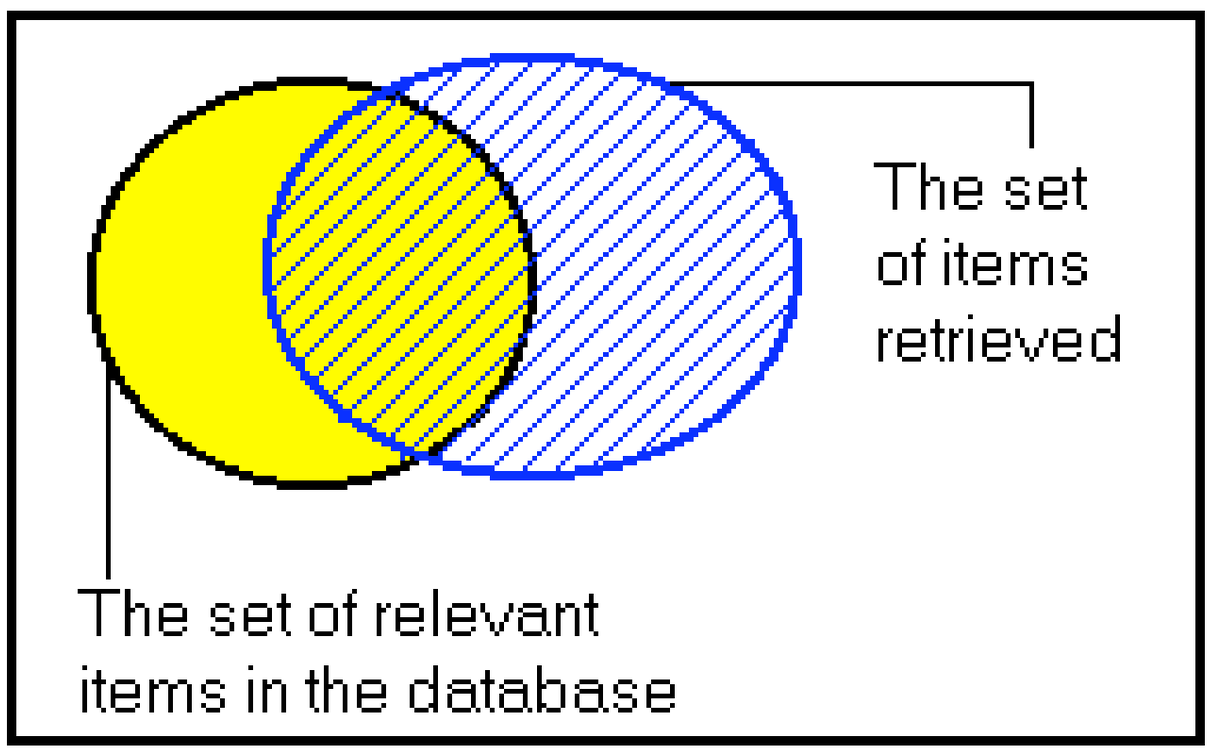
PR Computation

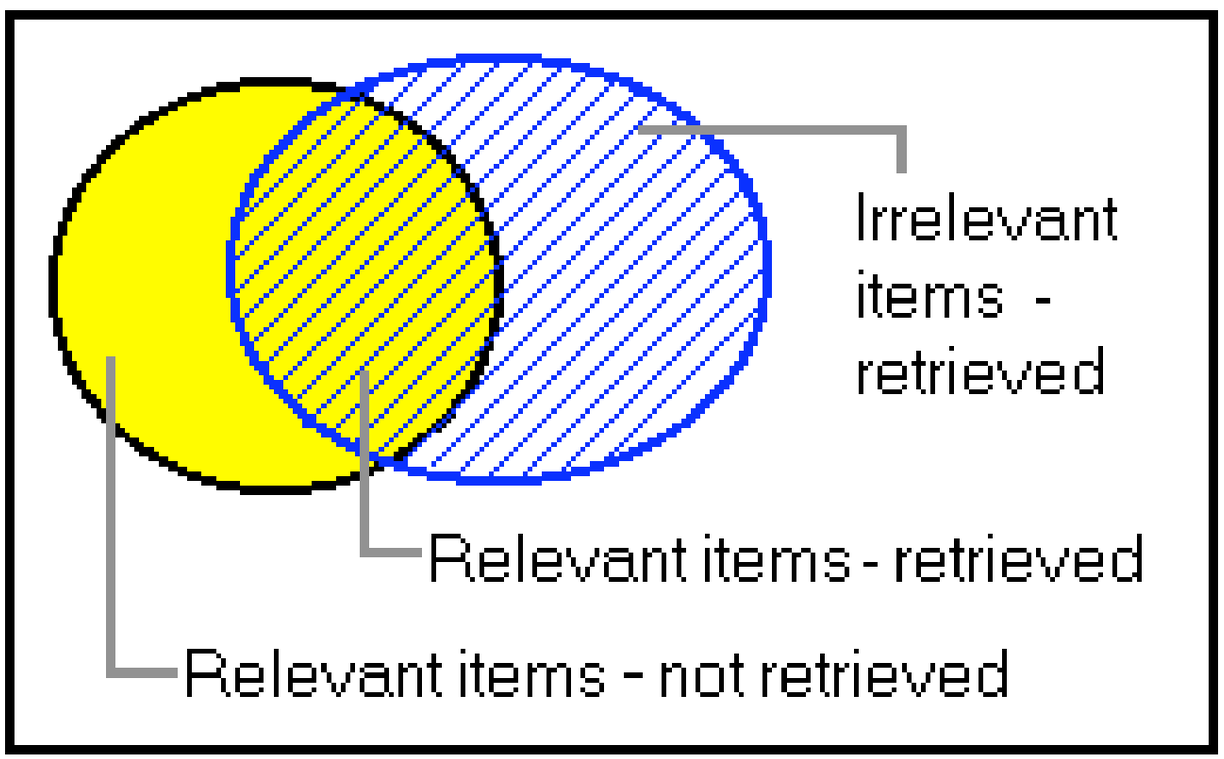
$$PR(A) = 1 - d + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right)$$

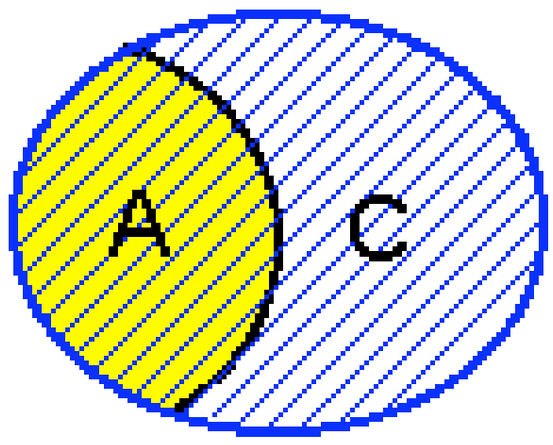
$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\right).$$

where N = number of documents in the collection

Precision and Recall



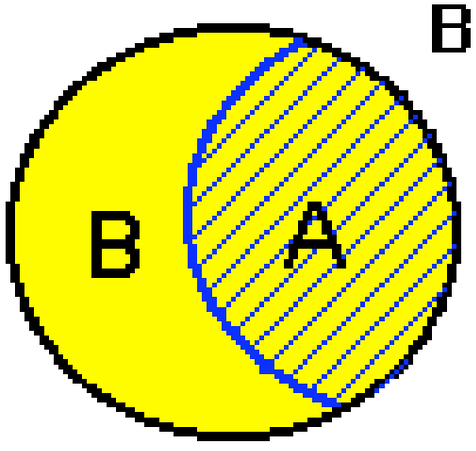




C: No. of irrelevant records retrieved.

A: No. of relevant records retrieved.

$$\text{PRECISION: } \frac{A}{A+C} \times 100\%$$



B: Number of relevant records not retrieved.

A: Number of relevant records retrieved.

$$\text{RECALL: } \frac{A}{A+B} \times 100\%$$

http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html