Mining large graphs

AMSC808N/CMSC828V

Maria Cameron

December 10, 2020

References

- David F. Gleich and Michael W. Mahoney, Mining large graphs, Handbook of Big Data, Handbooks of modern statistical methods, 2016
- Sergey Brin and Lawrence Page, The anatomy of a large-scale hyper textual Web search engine, Computer Networks and ISDN Systems 30 (1998) 107—117
- L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical report 1999-66, Stanford University, 1999

Sizes of LARGE graphs

Already outdated but gives some idea

- Google (2008): indexed over 10^{12} URLs
- Facebook (2012): $721*10^6$ individuals and $137*10^9$ links
- Phone companies (2013) process a few trillion calls a year
- The human brain (2011) has around $100*10^9$ and $100*10^{12}$ neuronal connections

Graph representations

- Edge list
- Adjacency list

Graph mining tasks

- Random walk steps (e.g. to extract a massive graph nearby the seed) O(1)
- Connected components O(n)
- PageRank — determine importance of nodes O(n)
- Effective diameter (mean shortest path or longest shortest path to connect 90% of possible node pairs) O(n)
- Extremal eigenvalues of graph Laplacian (the first eigenvector helps to split the graph) O(n log n)
- Triangle counting (detect interesting groups) $O(n^{3/2})$
- All-pairs problems $O(n^3)$ time and $O(n^2)$ memory

Classification of large graphs

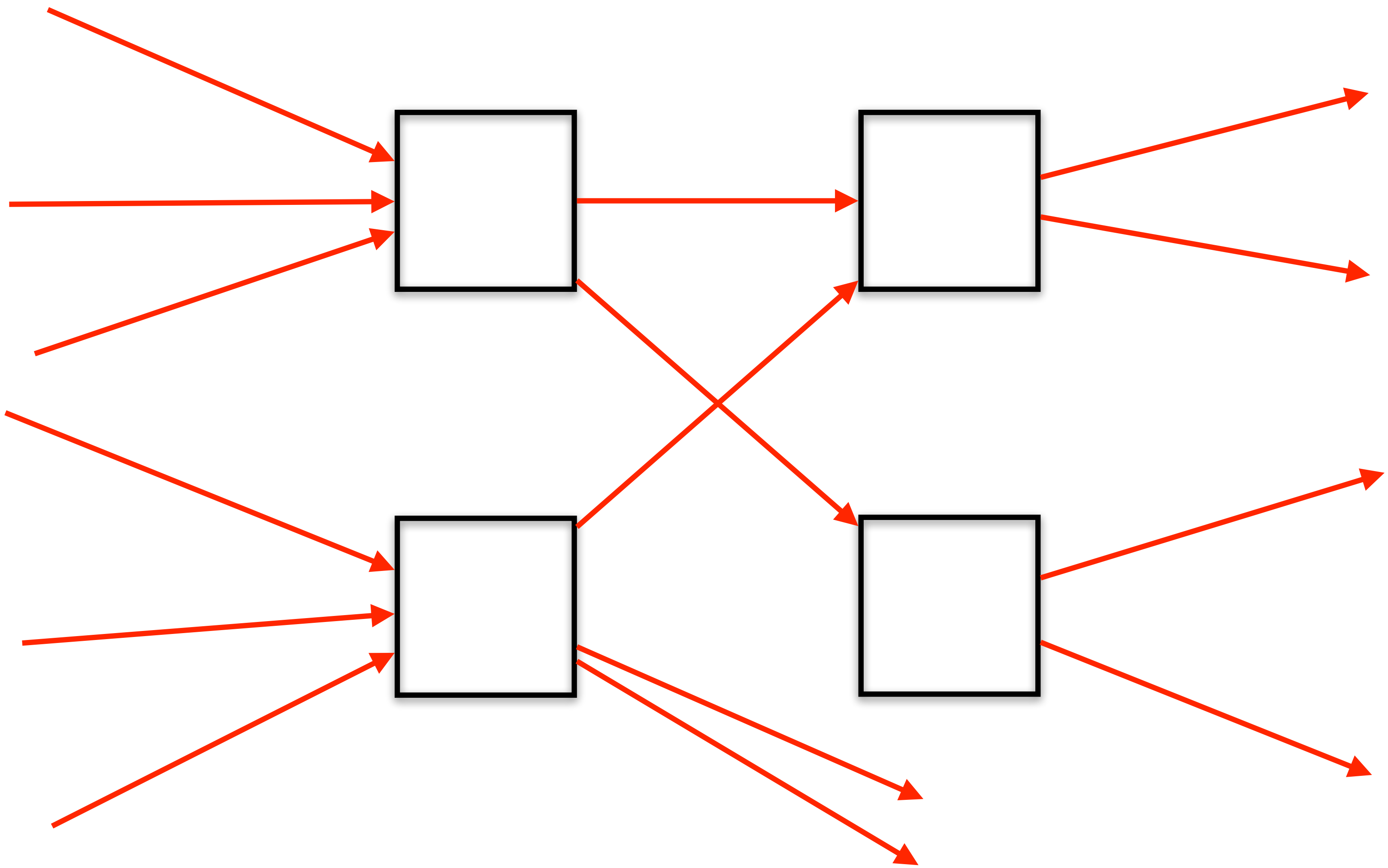
Graphs are sparse: the number of edges O(n)

- Small graphs (< 10^4 vertices) — all algorithms are feasible
- large Small graphs (10^4—10^6 vertices) — $O(n^2)$ in time is fine but $O(n^2)$ in memory may be prohibitive
- small Large graphs (10^6—10^8 vertices) — $O(n^2)$ is prohibitive without specialized computing resources
- Large graphs (10^8—10^{10} vertices)
- LARGE graphs (> 10^{10} vertices)

Sources for graph data

- https://snap.stanford.edu/data/index.html
- http://law.di.unimi.it/datasets.php
- http://www.lemurproject.org/clueweb12/webgraph.php/
- https://sparse.tamu.edu

The PageRank



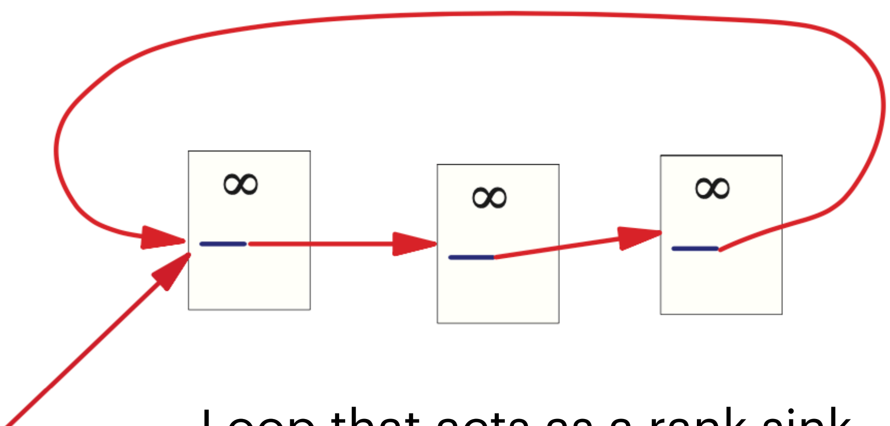
We are interested at backlinks of the page

The PageRank

Definition. Let *E*(*u*) be some vector over the Web pages that corresponds to a source of rank. Then, the PageRank of a set of Web pages is an assignment, *R′*, to the Web pages which satisfies

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

Such that c is maximized and $|| R'||_1 = 1$.



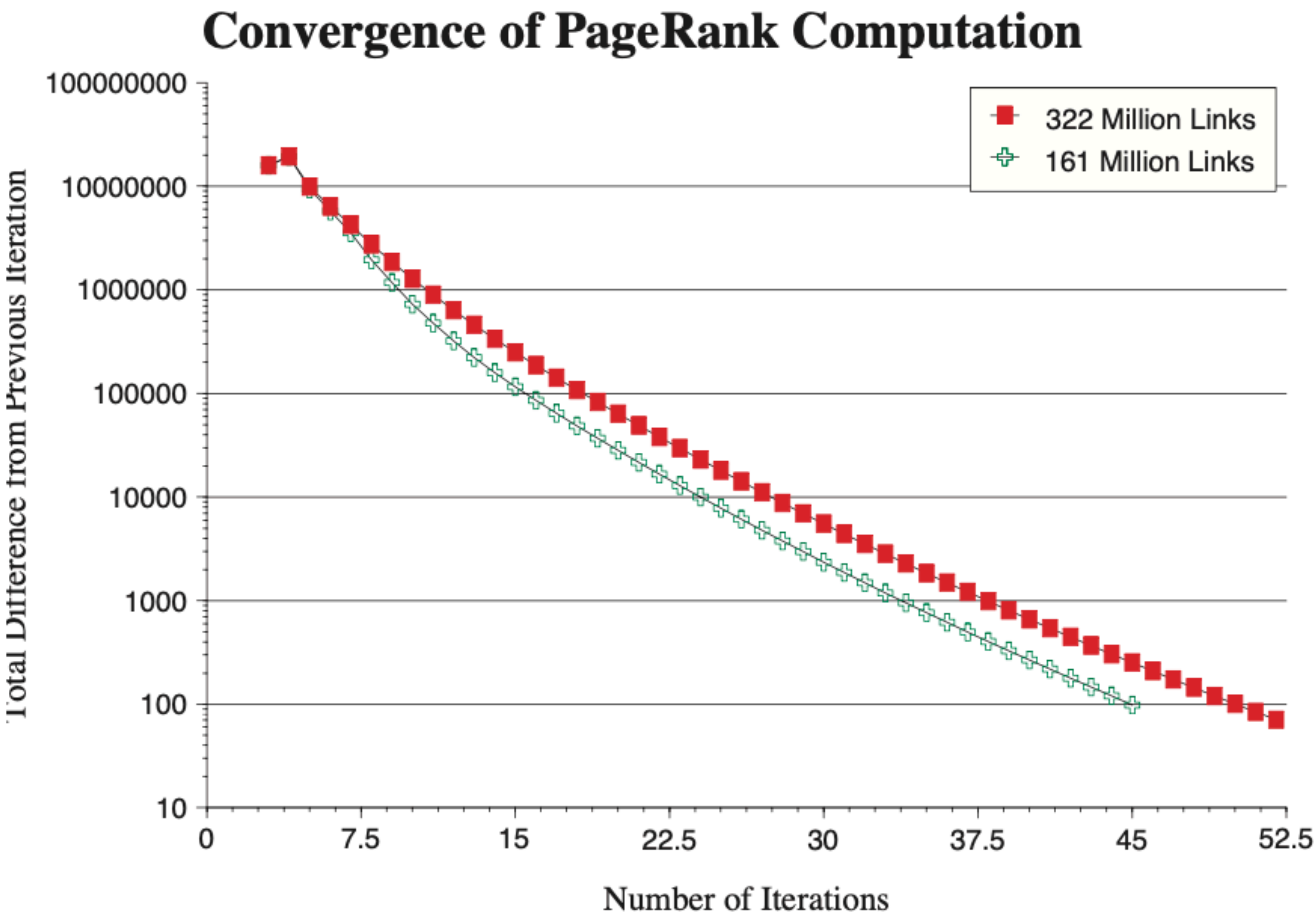
Loop that acts as a rank sink

Eigenvalue problem:

$$R' = c(A + E1^\top)R'$$

Computing PageRank

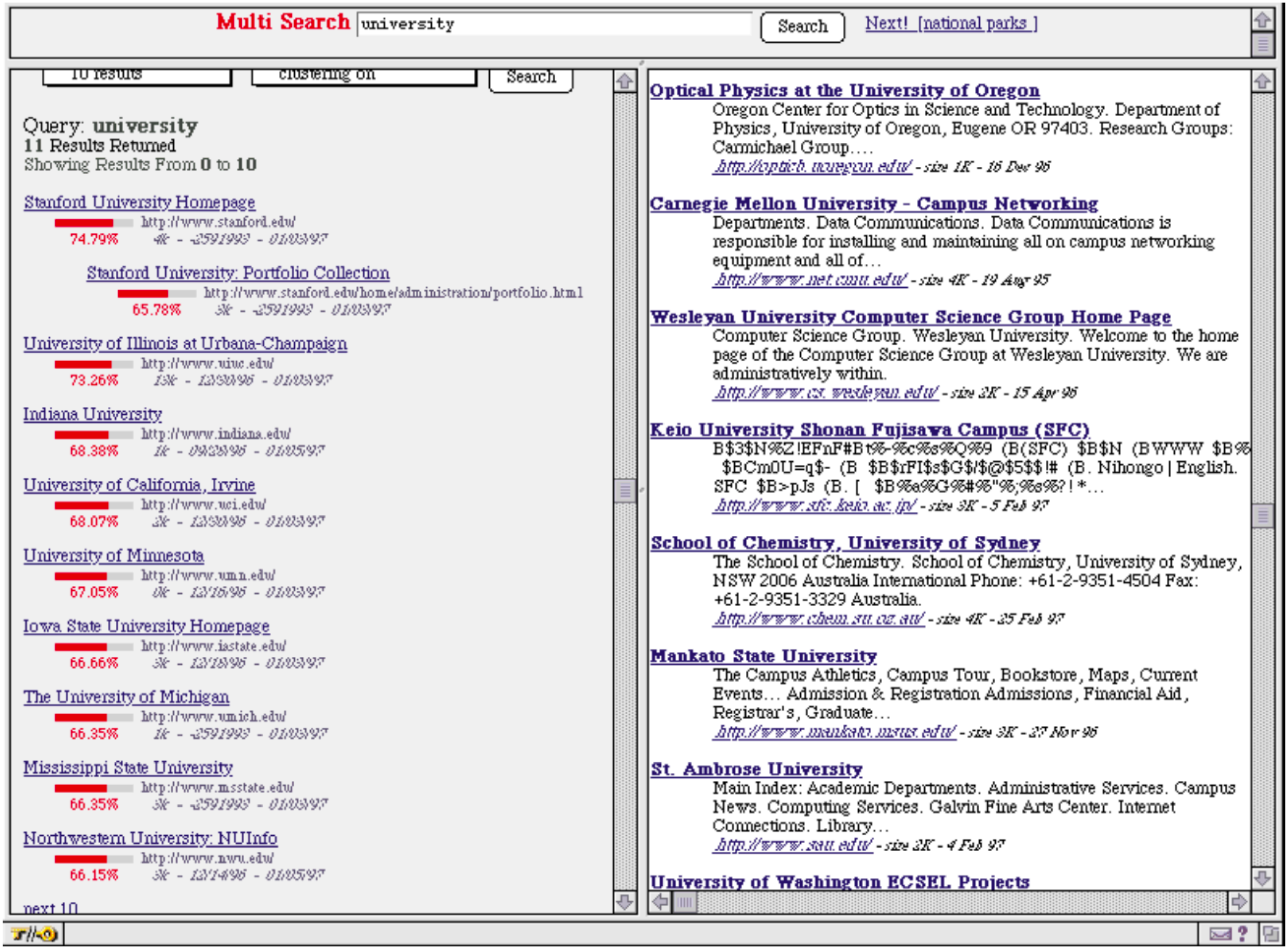
$$
\begin{aligned}
R_0 &= \text{an initial guess for the rank vector} \\
\textbf{while } \delta > \epsilon \\
R_{j+1} &= AR_j \\
d &= \|R_j\|_1 - \|R_{j+1}\|_1 \\
R_{j+1} &= R_{j+1} + dE \\
\delta &= \|R_{j+1} - R_j\|_1 \\
\textbf{end}
\end{aligned}
$$



Convergence of PageRank Computation
322 Million Links
161 Million Links
100000000
10000000
1000000
100000
10000
1000
100
10
0
7.5
15
22.5
30
37.5
45
52.5
Total Difference from Previous Iteration
Number of Iterations

Choosing source vector *E*

- E is uniform. Problem: sites like highly interlinked mailing lists archives receive overly high ranking.
- E is personalized: E(u) > 0 only for one page, e.g. user's personal web page.
- A compromise: E consists of all root-level pages of all web servers.


Multi Search
university
Search
Next! [national parks]
10 results
clustering on
Search
Query: university
11 Results Returned
Showing Results From 0 to 10
Stanford University Homepage
http://www.stanford.edu/
74.79%
Stanford University: Portfolio Collection
http://www.stanford.edu/home/administration/portfolio.html
65.78%
University of Illinois at Urbana-Champaign
http://www.uiuc.edu/
73.26%
Indiana University
http://www.indiana.edu/
68.38%
University of California, Irvine
http://www.uci.edu/
68.07%
University of Minnesota
http://www.umn.edu/
67.05%
Iowa State University Homepage
http://www.iastate.edu/
66.66%
The University of Michigan
http://www.umich.edu/
66.35%
Mississippi State University
http://www.msstate.edu/
66.35%
Northwestern University: NUInfo
http://www.nwu.edu/
66.15%
next 10
Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
Keio University Shonan Fujisawa Campus (SFC)
Nihongo | English.
School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
University of Washington ECSEL Projects

Web Page	PageRank (average is 1.0)
Download Netscape Software	11589.00
http://www.w3.org/	10717.70
Welcome to Netscape	8673.51
Point: It's What You're Searching For	7930.92
Web-Counter Home Page	7254.97
The Blue Ribbon Campaign for Online Free Speech	7010.39
CERN Welcome	6562.49
Yahoo!	6561.80
Welcome to Netscape	6203.47
Wusage 4.1: A Usage Statistics System For Web Servers	5963.27
The World Wide Web Consortium (W3C)	5672.21
Lycos, Inc. Home Page	4683.31
Starting Point	4501.98
Welcome to Magellan!	3866.82
Oracle Corporation	3587.63

Table 1: Top 15 Page Ranks: July 1996