

Google PageRank

Rob H. Bisseling

Mathematical Institute, Utrecht University

Course Introduction Scientific Computing
February 15, 2021

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

Weblink matrices

Google matrix

Eigensystem solution

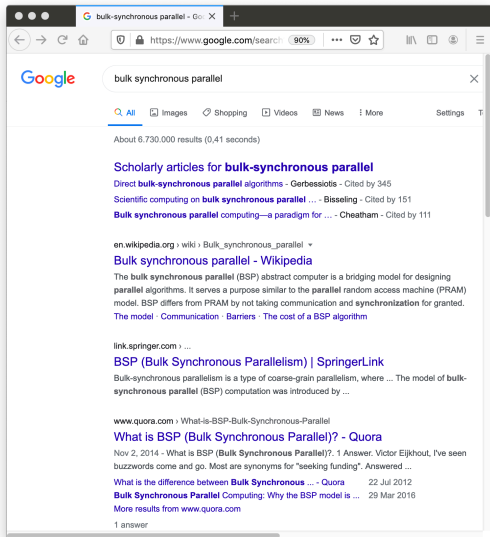
Weblink
matrices

Google matrix

Eigensystem



Web searching: which page ranks first?



A screenshot of a Google search results page for the query "bulk synchronous parallel". The browser address bar shows "https://www.google.com/search?q=bulk+synchronous+parallel". The search results are sorted by relevance, showing approximately 6,730,000 results in 0.41 seconds. The top results include:

- Scholarly articles for bulk-synchronous parallel**
 - Direct **bulk-synchronous parallel** algorithms - Gerbessiotis - Cited by 345
 - Scientific computing on **bulk synchronous parallel** ... - Bisseling - Cited by 151
 - Bulk synchronous parallel** computing—a paradigm for ... - Cheatham - Cited by 111
- en.wikipedia.org** › wiki › Bulk_synchronous_parallel
Bulk synchronous parallel - Wikipedia

The **bulk synchronous parallel** (BSP) abstract computer is a bridging model for designing **parallel** algorithms. It serves a purpose similar to the **parallel** random access machine (PRAM) model. BSP differs from PRAM by not taking communication and synchronization for granted.
The model · Communication · Barriers · The cost of a BSP algorithm
- link.springer.com** › ...
BSP (Bulk Synchronous Parallelism) | SpringerLink

Bulk-synchronous parallelism is a type of coarse-grain parallelism, where ... The model of **bulk-synchronous parallel** (BSP) computation was introduced by ...
- www.quora.com** › What-is-BSP-Bulk-Synchronous-Parallel
What is BSP (Bulk Synchronous Parallel)? - Quora

Nov 2, 2014 - What is BSP (Bulk Synchronous Parallel)? · 1 Answer. Victor Eijkhout, I've seen buzzwords come and go. Most are synonyms for "seeking funding". Answered ...

What is the difference between **Bulk Synchronous** ... - Quora 22 Jul 2012

Bulk Synchronous Parallel Computing: Why the BSP model is ... 29 Mar 2016

More results from **www.quora.com**

1 answer

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

The weblink matrix A

- ▶ Given n web pages with hyperlinks between them, we can define the sparse $n \times n$ **weblink matrix** A by

$$a_{ij} = \begin{cases} 1 & \text{if there is a hyperlink from page } j \text{ to page } i \\ 0 & \text{otherwise.} \end{cases}$$

- ▶ Let $\mathbf{e} = (1, 1, \dots, 1)^T$, represent an initial uniform importance (rank) of all web pages. Then

$$(\mathbf{A}\mathbf{e})_i = \sum_j a_{ij}e_j = \sum_j a_{ij}$$

is the **total number of hyperlinks pointing to page i** .

- ▶ The vector $\mathbf{A}\mathbf{e}$ represents the importance of the pages; $\mathbf{A}^2\mathbf{e}$ takes the importance of the pointing pages into account as well; and so on.

Weblink
matrices

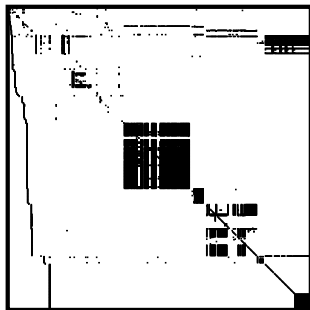
Google matrix

Eigensystem



Universiteit Utrecht

Weblink matrix bspww500



- ▶ This matrix with $n = 500$ and $nz = 13\,400$ represents 500 web pages and the hyperlinks connecting them.
- ▶ It was obtained by a **breadth-first search** in 2017 of the World Wide Web starting at <http://www.bsp-worldwide.org> and using the **web crawler** `surfer.m` by Cleve Moler.

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

The random surfer model

- ▶ A **random web surfer** chooses each of the c_j **outgoing hyperlinks** from page j with equal probability $\frac{1}{c_j}$.
- ▶ To incorporate this behaviour, we define the $n \times n$ diagonal **scaling matrix** D by

$$d_{jj} = c_j,$$

and multiply the weblink matrix A from the right by D^{-1} .

- ▶ This divides each column j of A by c_j .

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

The Google matrix

- ▶ Let α be the probability that a surfer follows an outlink of the current page. Typically $\alpha = 0.85$. The surfer jumps to a random page with probability $1 - \alpha$.
- ▶ The Google matrix is defined by

$$G = \alpha AD^{-1} + (1 - \alpha) \frac{1}{n} \mathbf{e} \mathbf{e}^T.$$

- ▶ Note that this definition is under the assumption that all $c_j > 0$.
- ▶ The PageRank of a set of web pages is obtained by repeated multiplication by G , involving sparse matrix–vector multiplication by A and some vector operations.



S. Brin and L. Page, *Computer Networks and ISDN Systems*, 30(1–7) (1998) pp. 107–117.



Universiteit Utrecht

Weblink
matrices

Google matrix

Eigensystem

Vector operation

- ▶ The vector \mathbf{e} can be viewed as an $n \times 1$ matrix of all ones and the vector \mathbf{e}^T as a $1 \times n$ matrix.
- ▶ The matrix $\mathbf{e}\mathbf{e}^T$ is an $n \times n$ matrix with all elements equal to 1.
- ▶ Multiplication of a vector \mathbf{x} by $\mathbf{e}\mathbf{e}^T$ is cheap:

$$\mathbf{e}\mathbf{e}^T\mathbf{x} = \mathbf{e}(\mathbf{e}^T\mathbf{x}) = \left(\sum_{i=1}^n x_i\right)\mathbf{e}.$$

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

Escaping from dangling nodes

- ▶ If $c_j = 0$, column j must be empty and page j a **dangling node**; it could be a PDF file or an image file.
- ▶ To avoid division by zero, the diagonal element is then **redefined** as $d_{jj} = 1$.
- ▶ To make the Google matrix stochastic (with all column sums equal to 1), we must add to G an **extra term**

$$\alpha \frac{1}{n} \mathbf{e} \hat{\mathbf{e}}^T.$$

- ▶ Here, the vector $\hat{\mathbf{e}}$ is defined by

$$\hat{e}_j = \begin{cases} 1 & \text{if } c_j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

Power method

- ▶ Let \mathbf{x} be the vector of page ranks: component x_i represents the relative **importance** (rank) of page i , with $0 \leq x_i \leq 1$ and $\sum_i x_i = 1$.
- ▶ The **power method** computes $A\mathbf{x}, A^2\mathbf{x}, A^3\mathbf{x}, \dots$, until convergence.
- ▶ The final component x_i represents the rank of page i in the **steady-state** situation, where $A\mathbf{x} = \mathbf{x}$.
- ▶ Main operation: multiplication of sparse matrix A and dense vector \mathbf{x} .

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

Alternative ranking method: HITS algorithm

- ▶ The Hyperlink-Induced Topic Search (HITS) algorithm by Jon Kleinberg repeatedly computes $A^T A \mathbf{x}$ instead of $A \mathbf{x}$, where A is a weblink matrix, to obtain two page-ranking values:
 - the **authority value** y_i which says how authoritative page i is as a source of information,
 - the **hub value** x_i which says in how far its hyperlinks point to authorities.
- ▶ An authority vector \mathbf{y} can be computed from a hub vector \mathbf{x} by $\mathbf{y} := A \mathbf{x}$, and vice versa by $\mathbf{x} := A^T \mathbf{y}$.
- ▶ Here, the power method is applied to solve the eigensystem $A^T A \mathbf{x} = \lambda \mathbf{x}$.

 J. M. Kleinberg, *Journal of the ACM*, **46**(5) (1999) pp. 604–632.

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

Expansion in eigenvectors

- ▶ Let \mathbf{x} be an initial vector of page ranks. Let $\lambda_1 \geq \dots \geq \lambda_n$ be the **eigenvalues** of the $n \times n$ matrix G and $\mathbf{v}_1, \dots, \mathbf{v}_n$ the corresponding **eigenvectors**.
- ▶ We can write \mathbf{x} as a **linear combination** of the (unknown) eigenvectors,

$$\mathbf{x} = \sum_{k=1}^n c_k \mathbf{v}_k.$$

- ▶ Applying t times G then gives

$$\begin{aligned} G^t \mathbf{x} &= G^t \sum_{k=1}^n c_k \mathbf{v}_k \\ &= \sum_{k=1}^n c_k G^t \mathbf{v}_k \\ &= \sum_{k=1}^n c_k \lambda_k^t \mathbf{v}_k. \end{aligned}$$

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

(1)

Convergence of the power method

- ▶ For the special matrix G , which is **stochastic** ($0 \leq g_{ij} \leq 1$, $\sum_{i=1}^n g_{ij} = 1$), the largest eigenvalue is $\lambda_1 = 1$ and the other eigenvalues are smaller: $0 \leq \lambda_i < 1$ for $i \geq 2$. Thus

$$G^t \mathbf{x} \approx c_1 \mathbf{v}_1 + c_2 \lambda_2^t \mathbf{v}_2.$$

- ▶ We have the nice property

$$\lambda_2 \leq \alpha = 0.85,$$

so the error decreases as $\mathcal{O}(0.85^t)$.

- ▶ After 50 iterations, the error is about 0.03%.



Amy N. Langville and Carl D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2006.

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht

Summary

- ▶ An $n \times n$ **weblink matrix** A is defined by

$$a_{ij} = \begin{cases} 1 & \text{if there is a hyperlink from page } j \text{ to page } i \\ 0 & \text{otherwise.} \end{cases}$$

- ▶ The **Google** matrix is defined by

$$G = \alpha AD^{-1} + (1 - \alpha) \frac{1}{n} \mathbf{e} \mathbf{e}^T + \alpha \frac{1}{n} \mathbf{e} \hat{\mathbf{e}}^T,$$

where $\mathbf{e} = (1, \dots, 1)^T$ and $\hat{e}_j = 1$ if page j has no outlinks.

- ▶ The page rank of a set of webpages can be computed by the power method in 50 iterations. This is the Google dance.
- ▶ There are other search engines, some respecting your privacy better: Bing, DuckDuckGo, StartPage.

Weblink
matrices

Google matrix

Eigensystem



Universiteit Utrecht