Google PageRank

Prof. Beat Signer
<bsigner@vub.ac.be>

Department of Computer Science
Vrije Universiteit Brussel

http://www.beatsigner.com

Overview

- History of PageRank
- PageRank algorithm
- Examples
- Implications for website development

History of PageRank

- Developed as part of an academic project at Stanford University
  - research platform to *aid understanding of large-scale web data* and enable researches to easily experiment with new search technologies
  - *Larry Page* and *Sergey Brin* worked on the project about a *new kind of search engine* (1995-1998) which finally led to a functional prototype called *Google*

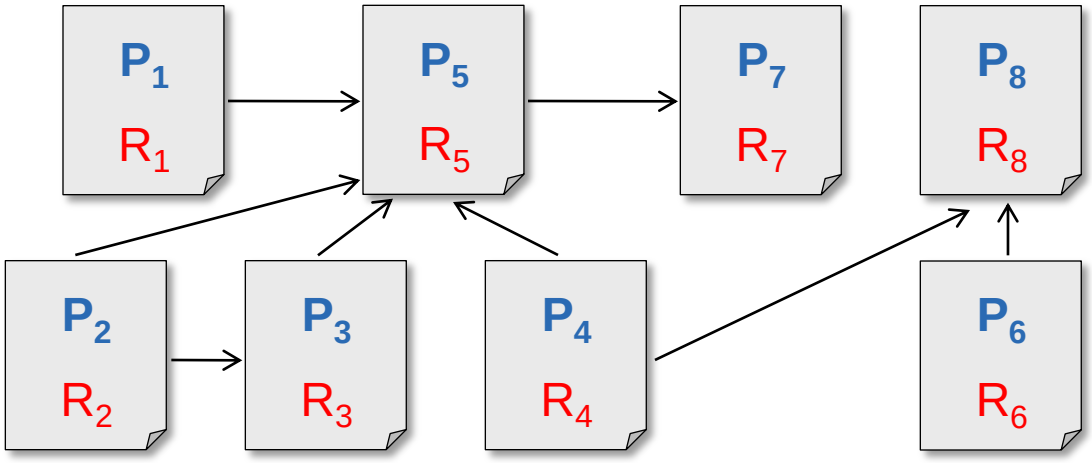

Larry Page


Sergey Brin

Web Search Until 1998

- Find all documents using a query term
  - use information retrieval (IR) solutions
  - ranking based on "on-page factors"
    → *problem: poor quality of search results (order)*
- Page and Brin proposed to compute the *absolute qualtity* of a page (*PageRank*)
  - based on the *number and quality* of pages linking to a page (votes)

PageRank

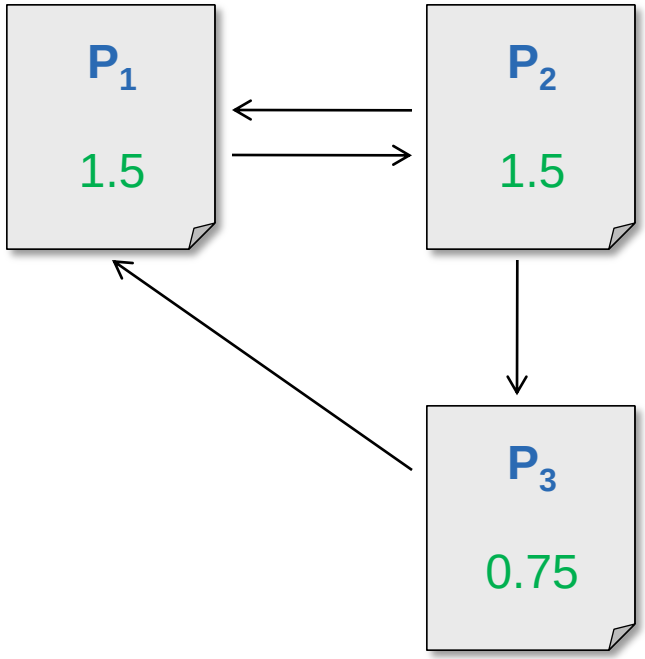


- A page has a high PageRank R if
  - there are many pages linking to it
  - or, if there are some pages with a high PageRank linking to it
- Total score = IR score *x* PageRank

PageRank Algorithm

$$R(P_i) = \sum_{P_j \in B_i} \frac{R(P_j)}{L_j}$$

- where
  - B_i is the set of pages that link to page P_i
  - L_j is the number of outgoing links for page P_j



Matrix Representation

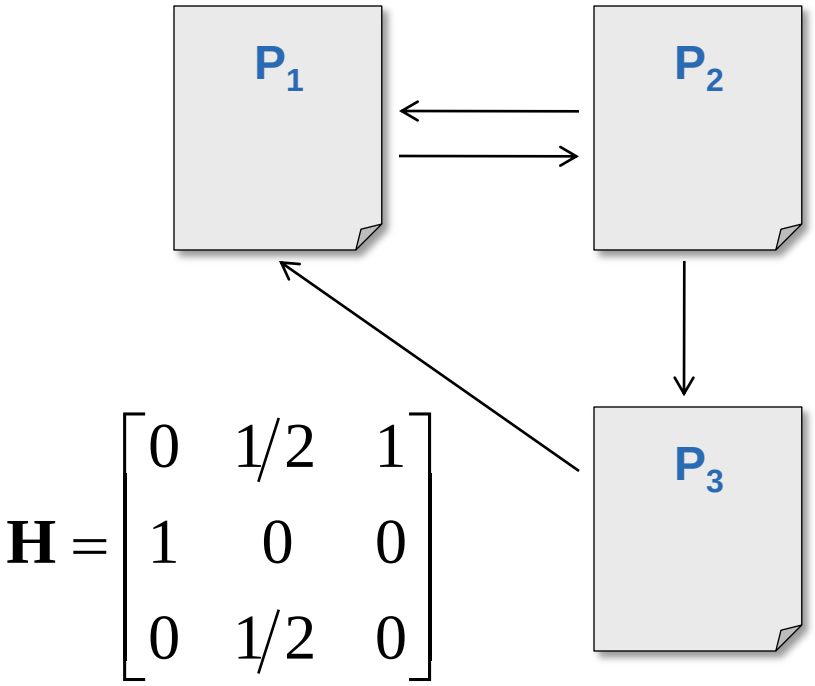
- Let us define a hyperlink matrix $\mathbf{H}$

$$\mathbf{H}_{ij} = \begin{cases} 1/L_j & \text{if } P_j \in B_i \\ 0 & \text{otherwise} \end{cases}$$

and $\mathbf{R} = \left[R(P_i)\right]$

$$\rightarrow \mathbf{R} = \mathbf{H}\mathbf{R}$$

$\mathbf{R}$ is an eigenvector of $\mathbf{H}$ with eigenvalue 1



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

Matrix Representation ...

- We can use the *power method* to find $\mathbf{R}$

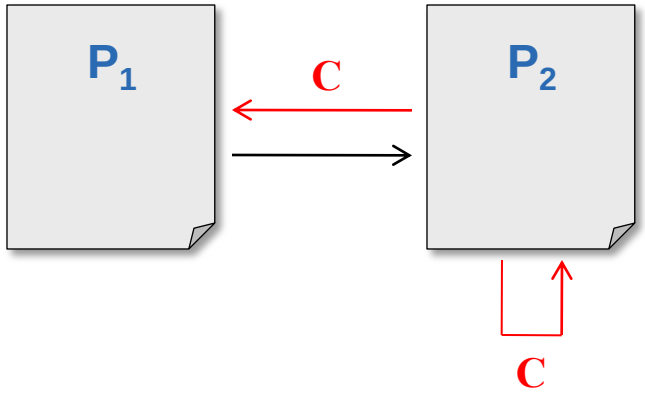
$$\mathbf{R}^{t+1} = \mathbf{H}\mathbf{R}^{t}$$

For our example $\mathbf{H} = \begin{bmatrix} 0 & 1/2 & 1 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix}$

this results in $\mathbf{R} = \begin{bmatrix} 2 & 2 & 1 \end{bmatrix}$ or $\begin{bmatrix} 0.4 & 0.4 & 0.2 \end{bmatrix}$

Dangling Pages

- Problem with pages that have no outbound links (P_2)



$$\mathbf{H} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \text{ and } \mathbf{R} = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

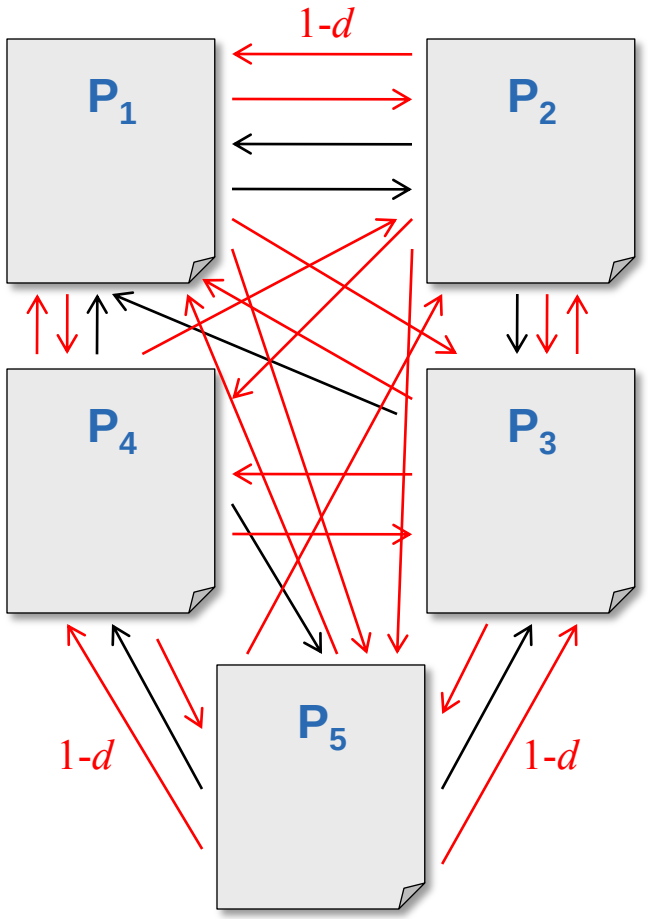
$$\mathbf{C} = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/2 \end{bmatrix} \text{ and } \mathbf{S} = \mathbf{H} + \mathbf{C} = \begin{bmatrix} 0 & 1/2 \\ 1 & 1/2 \end{bmatrix}$$

Strongly Connected Pages (Graph)

- Add new transition probabilities between all pages
  - with *probability d* we *follow* the *hyperlink structure S*
  - with *probability 1-d* we *choose a random page*

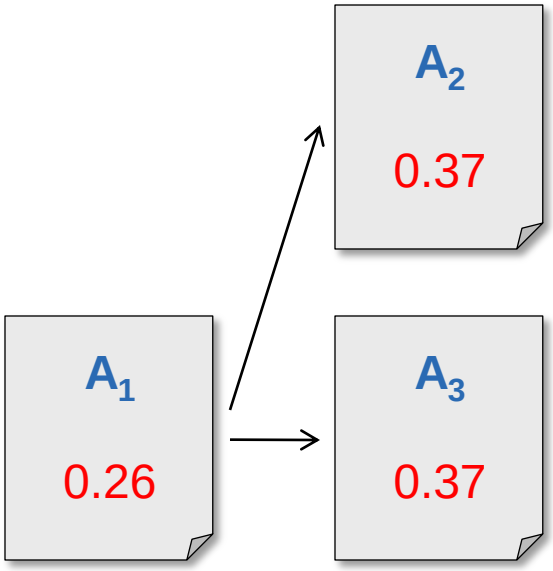
$$\mathbf{G} = \left(1-d\right)\frac{1}{n}\mathbf{1} + d\mathbf{S}$$

$$\mathbf{R} = \mathbf{G}\mathbf{R}$$



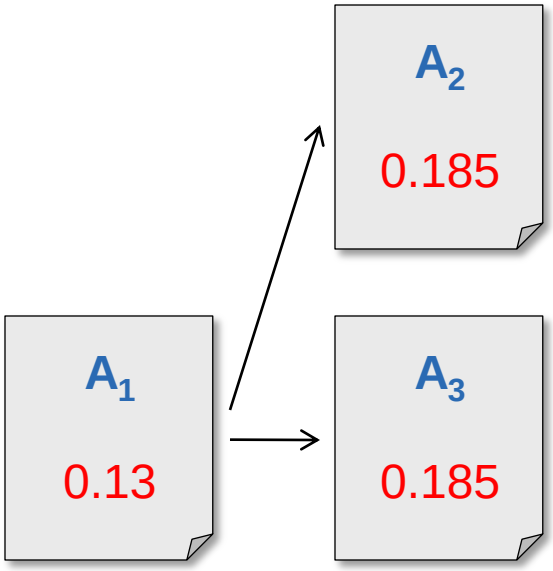
Examples

$$\mathbf{G} = (1-d)\frac{1}{n}\mathbf{1} + d\mathbf{S}$$

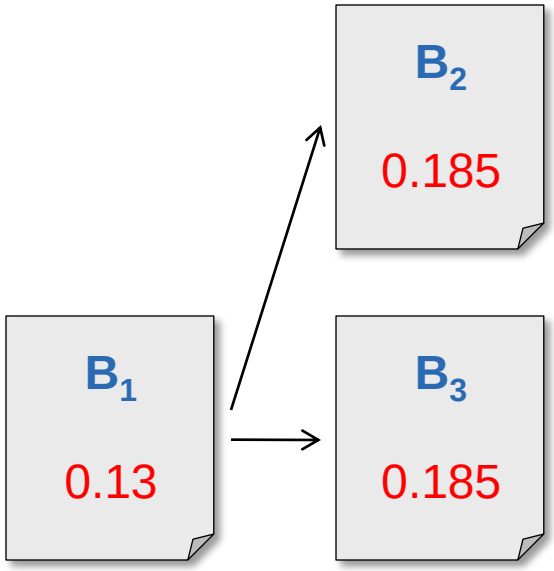


Examples ...

$$\mathbf{G} = (1-d)\frac{1}{n}\mathbf{1} + d\mathbf{S}$$



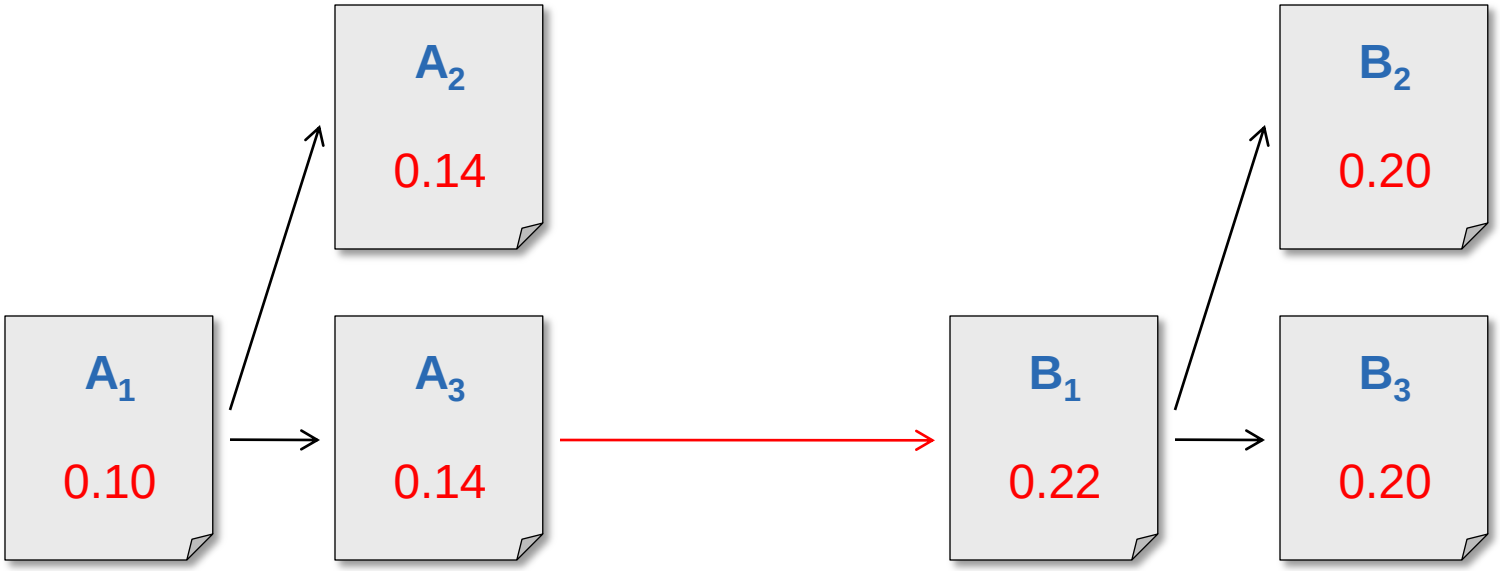
$$P(A) = 0.5$$



$$P(B) = 0.5$$

Examples ...

$$\mathbf{G} = (1-d)\frac{1}{n}\mathbf{1} + d\mathbf{S}$$

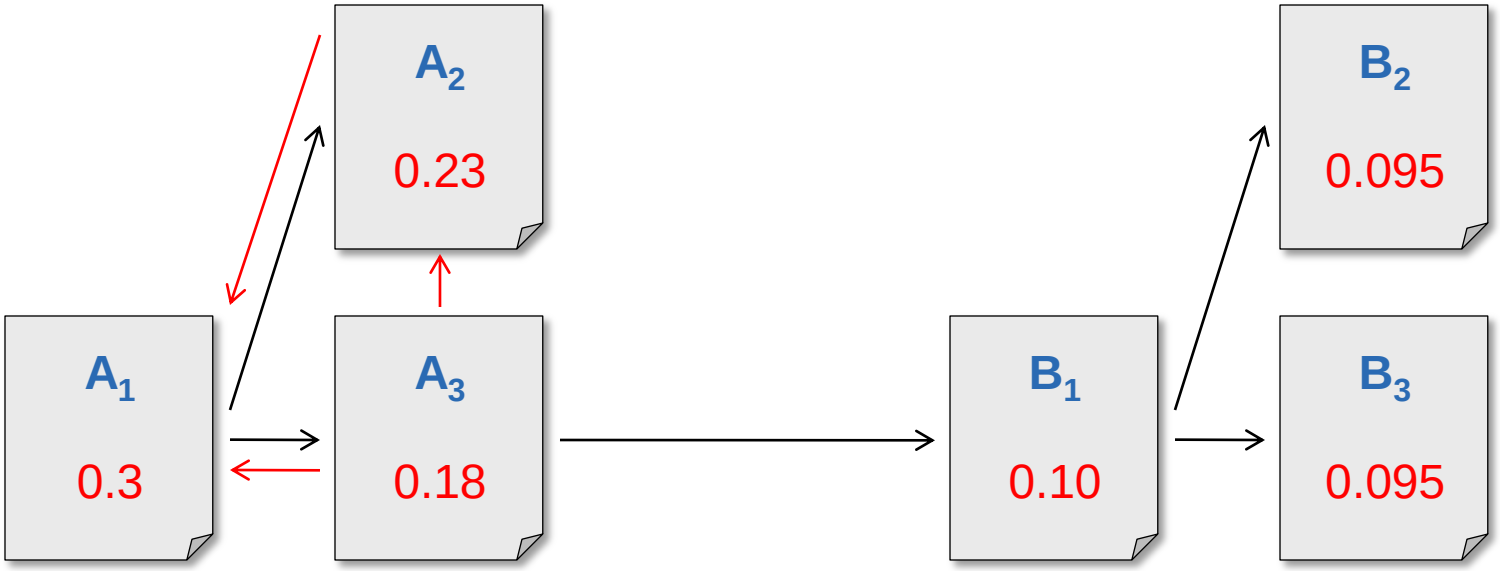


$$P(A) = 0.38$$

$$P(B) = 0.62$$

Examples ...

$$\mathbf{G} = (1-d)\frac{1}{n}\mathbf{1} + d\mathbf{S}$$

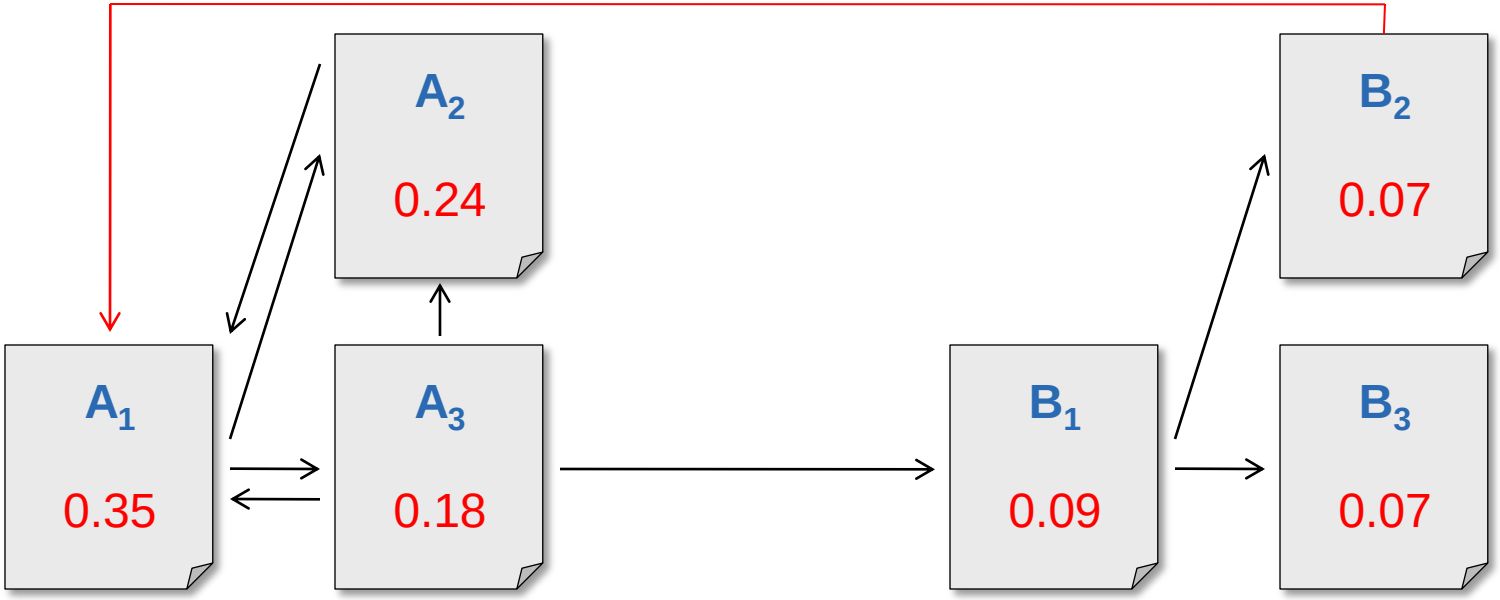


$$P(A) = 0.71$$

$$P(B) = 0.29$$

Examples ...

$$\mathbf{G} = (1-d)\frac{1}{n}\mathbf{1} + d\mathbf{S}$$

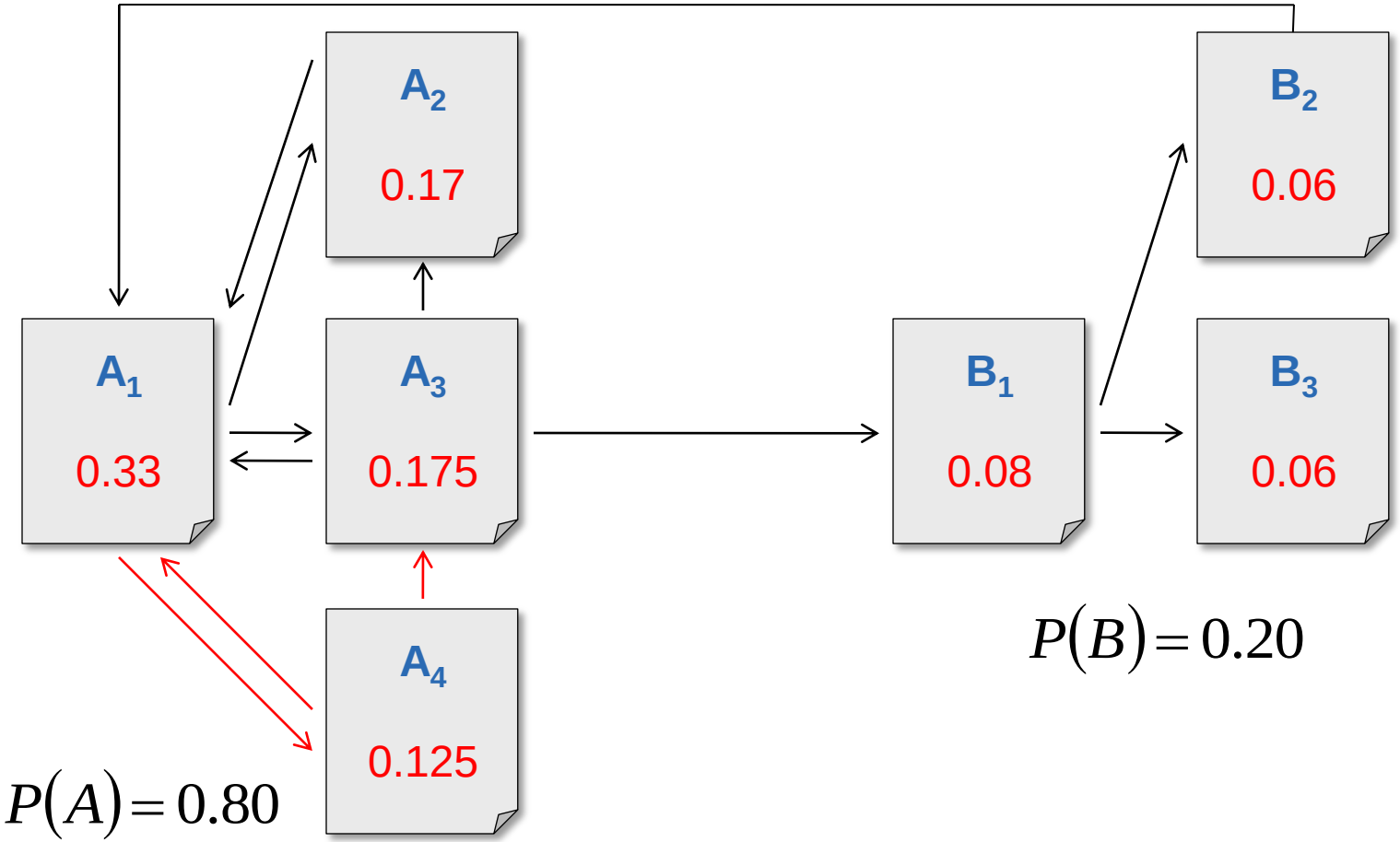


$$P(A) = 0.77$$

$$P(B) = 0.23$$

Examples ...

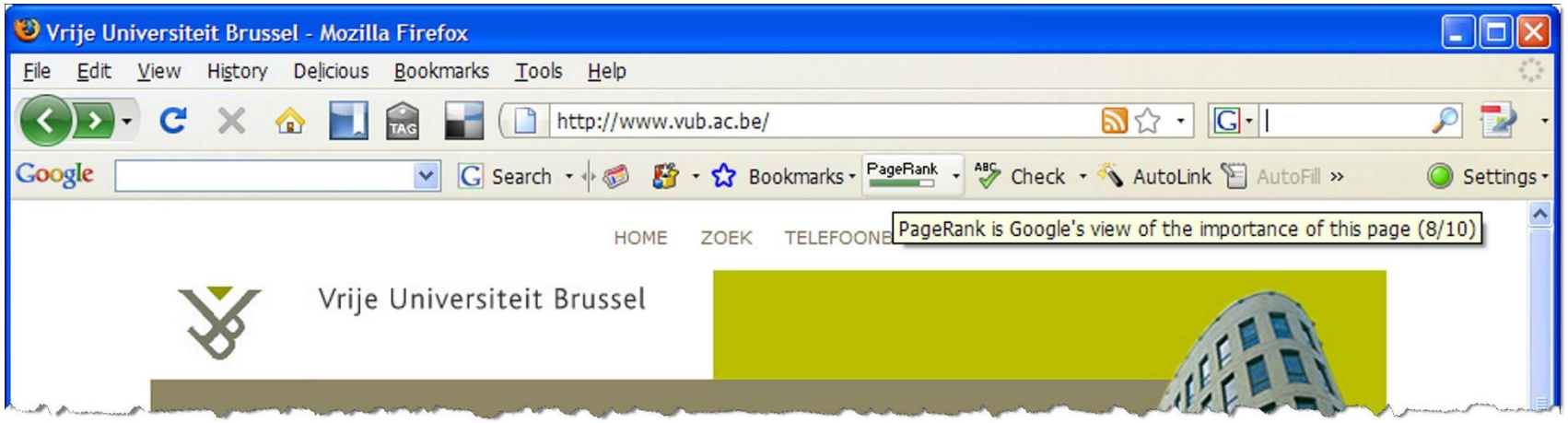
$$\mathbf{G} = (1-d)\frac{1}{n}\mathbf{1} + d\mathbf{S}$$

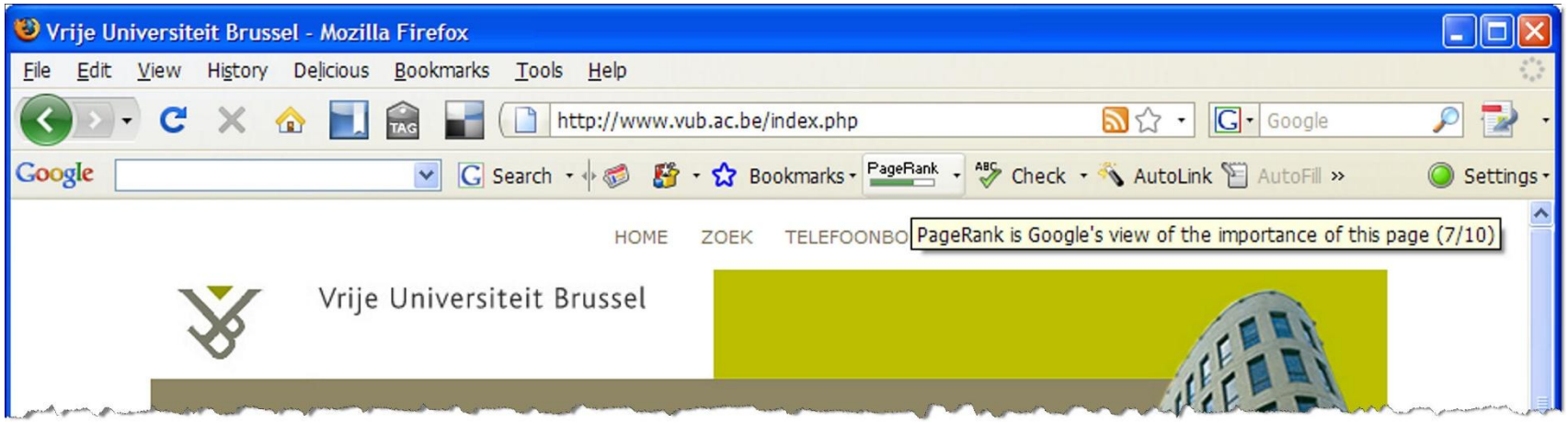


Implications for Website Development

- First make sure that your page gets indexed
  - on-page factors
- Think about your site's internal link structure
  - create many internal links for important pages
  - be "careful" about where to put outgoing links
- Increase the number of pages
- Ensure that webpages are addressed consistently
  - http://www.vub.ac.be ≠ http://www.vub.ac.be/index.php
- Make sure that you get links from good websites

Consistent Addressing of Webpages





Search Engine Optimisations (SEO)

- Internet marketing has become a big business
  - *white hat* and *black hat* optimisations
- Bad ranking or removal from index can cost a company a lot of money
  - e.g. supplemental index ("Google hell")

Black Hat Optimisations (Don'ts)

- *Link farms*
- *Spamdexing* in guestbooks, Wikipedia etc.
  - "solution": **…**
- *Doorway pages (cloaking)*
  - e.g. BMW Germany and Ricoh Germany banned in February 2006
- *Selling/buying links*
- …

On-Page Factors (Speculative)

- It is assumed that there are over 200 on-page and off-page factors
- Positive factors
  - keyword in title tag
  - keyword in URL
  - keyword in domain name
  - quality of HTML code
  - page freshness (occasional changes)
  - …

On-Page Factors (Speculative) …

- Negative factors
  - links to "bad neighbourhood"
  - over optimisation penalty (keyword stuffing)
  - text with same colour as background (hidden content)
  - automatic redirects via the **`refresh`** meta tag
  - any copyright violations
  - …

Off-Page Factors (Speculative)

- Positive factors
  - high PageRank
  - anchor text of inbound links
  - links from authority sites (Hilltop algorithm)
  - listed in DMOZ (ODP) and Yahoo directories
  - site age (stability)
  - domain expiration date
  - …

Off-Page Factors (Speculative) …

- Negative factors
  - link buying (fast increasing number of inbound links)
  - link farms
  - cloaking (different pages for spider and user)
  - limited (temporal) availability of site
  - links from bad neighbourhood?
  - competitor attack (e.g. duplicate content)?
  - …

Tools

- Google toolbar
  - PageRank information not frequently updated
- Google webmaster tools
  - meta description issues
  - title tag issues
  - non-indexable content issues
  - number and URLs of indexed pages
  - number and URLs of inbound/outbound links
  - ...

Questions

- Is PageRank fair?
- What about Google's power and influence?

Conclusions

- PageRank algorithm
  - absolute quality of a page based on incoming links
  - random surfer model
  - computed as eigenvector of Google matrix G
- Implications for website development and SEO
- PageRank is *just one* (important) *factor*

References

- *The PageRank Citation Ranking: Bringing Order to the Web,* L. Page, S. Brin, R. Motwani and T. Winograd, January 1998
- *The Anatomy of a Large-Scale Hypertextual Web Search Engine,* S. Brin and L. Page, Computer Networks and ISDN Systems, 30(1-7), April 1998

References …

- *PageRank Uncovered*, C. Ridings and M. Shishigin, September 2002
- *PageRank Calculator*, `http://www.webworkshop.net/pagerank_calculator.php`