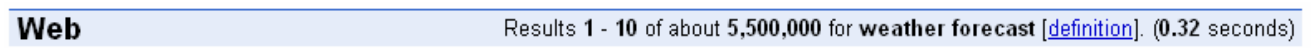
Computing the Google PageRank

The World's Largest Eigenvalue Problem

What is PageRank

A search with Google's search engine usually returns a very large number of pages. E.g., a search on 'weather forecast' returns 5.5 million pages.


Web Results 1 - 10 of about 5,500,000 for weather forecast [definition]. (0.32 seconds)

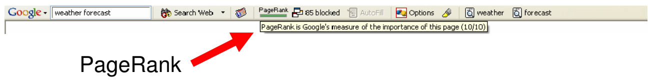
Although the search returns several million pages, the most relevant pages are usually found within the top ten or twenty pages in the list of results.

How does the search engine know which pages are the most important?

Google assigns a number to each individual page, expressing its importance. This number is known as the PageRank and is computed via the eigenvalue problem

$$Pw=\lambda w$$

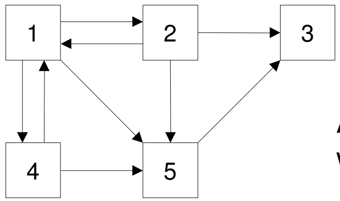
where P is based on the link structure of the Internet.



PageRank

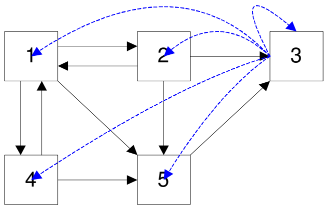
The key problem is to formulate the link structure, i.e., the matrix P, in a proper way.

The Link Structure Matrix *P*



A tiny internet with 5 pages.

The model forming the basis of the PageRank algorithm is a random walk through all the pages of the Internet. Let $p_t(x)$ denote the possibility of being on page x at time t. The PageRank of page x is expressed as $\lim(p_t(x))$ for $t \rightarrow \infty$. To make sure the random walk process does not get stuck, pages with no out-links (here: page 3) are assigned artificial links or "teleporters" to all other pages.



$$P = \begin{pmatrix} 0 & 1/3 & 1/5 & 1/2 & 0 \\ 1/3 & 0 & 1/5 & 0 & 0 \\ 0 & 1/3 & 1/5 & 0 & 1 \\ 1/3 & 0 & 1/5 & 0 & 0 \\ 1/3 & 1/3 & 1/5 & 1/2 & 0 \end{pmatrix}$$

The matrix P is irreducible and stochastic and therefore the random walk can be expressed as a Markov chain, and the PageRank of all pages can be computed as the principal eigenvector of P.

The PageRank Algorithm

The Google matrix P is currently of size 4.2×10^9 and therefore the eigenvalue computation is not trivial. To find an approximation of the principal eigenvector the *power method* is used:

```
w_0 = initial guess
For k = 1 to 50
      w_k = P*w_{k-1}
End
Return w_50
```

The special properties of the matrix P ensures that the largest eigenvalue is $\lambda = 1$, rendering normalisation in the power method unnecessary. Fast convergence of the power method makes 50 iterations adequate.

Because the computation involves an extremely large matrix, the matrix-vector multiplications must be implemented in parallel on multi-processor systems.

DTU

Informatics and Mathematical Modelling