Mathematical Analysis of Google PageRank

1. Google PageRank Essentials

Typically a search engine finds hundreds or thousands relevant answers to a user query.

An important problem:
How a search engine should sort the retrieved answers?

Possible solutions: (a) use the frequency of the searched terms in the Web page, (b) analyse the log files,... These solutions might be not objective.

An original idea of Google is based on two observations:

1.The more pages point to a Web page, the more important the page is.

2.If more important Web pages point to the page, the page is even more important.

In fact, the above two observations can be translated in the strict mathematical language of Markov chains.

Consider the Web as a directed graph where an edge $i \to j$ is a hyperlink from page i to page j.

Then, construct a **hyperlink matrix** with the elements:

$$P_{ij} = \begin{cases} 1/k, & \text{if } i \text{ has } k > 0 \text{ outgoing links and } j \text{ is one of the links,} \\ 0, & \text{otherwise.} \end{cases}$$

We note that the hyperlink matrix is substochastic.

If it were stochastic and represented an ergodic Markov chain, one could use its stationary distribution as a measure of the page popularity.

There are at least two reasons why in reality the hyperlink matrix is not stochastic and the Markov chain is not ergodic:

1. There are many pages that do not point to any page;
2. The Web consists of several disconnected components.

To rectify these problems, Google considers the following **Singularly Perturbed Markov Chain (SPMC)**

$$\tilde{P} = cP + (1-c)(1/n)E, \tag{1}$$

where n is the total number of Web pages ($n \approx 5\ 000\ 000\ 000$).

Furthermore, if a Web page does not have any outgoing link, we can assume that it either points to all pages or only to itself. In fact, these two formulations turn out to be equivalent.

Then, $\tilde{P}$ defines an ergodic Markov chain, and **PageRank** π is defined as a unique stationary distribution of $\tilde{P}$. Namely,

$$\pi\tilde{P} = \pi, \quad \pi\underline{1} = 1. \tag{2}$$

Interpretation of the perturbation:
A random walker surfs the Web using hyperlinks. With probability c he/she follows one of the outgoing link with the uniform distribution, and with probability $(1-c)$ he/she gets bored and jumps to a completely random page.

Even though this is a well kept secret, it seems that Google still uses the **simple power iteration method** for the PageRank computation.

$$\pi^{(k+1)} = c\pi^{(k)}P + (1-c)\frac{1}{n}\underline{1}^T$$

It can be easily estimated that using the constant $c = 0.85$ Google achieves the tolerance level (measured by the residual $\pi^{(k+1)} - \pi^{(k)}$) of $10^{-3} - 10^{-5}$ for only 50-100 iterations.

But even this number of iterations takes Google about a week to update the PageRank...

2. Some questions that we can answer:

1. How to organize a web site so that the important pages get higher PageRank?
2. Is it possible to link to another page and not to lose your own PageRank?
3. Is there an optimal linking strategy?
4. If one page of a Web site makes an inappropriate link by how much the other pages of this site suffer?
5. Does the reciprocal linking always benefit both pages?

And there are still many open questions...