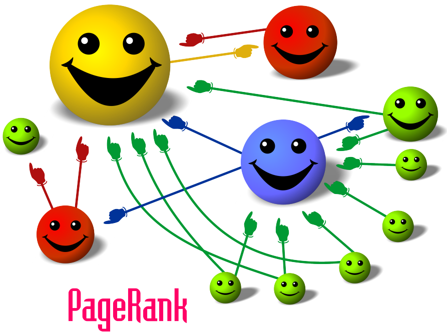
Bachelor thesis project proposal:
Modern numerical algorithms for the pagerank

Project description



Modern web search engine as Google, Yahoo and Bing Search sort the results of your query by measuring the importance of webpages containing your quest. The algorithm used is called Pagerank and it was the first algorithm used by Google.

The idea behind the Pagerank follows a simple logic: the importance of a page is roughly estimated by counting the number and quality of links that point to that page. In other words, if you image the webpages as people, your importance depends of the importance of your friends.

Although this argument seems a tautology, it has a solid mathematical foundation based on probability theory. The practical computation of the importance of the webpages is computed by solving a linear eigenvalue problem. More precisely, we enumerate all the webpages from 1 to n and we define the matrix $G \in \mathbb{R}^{n \times n}$ such that

$$G_{i,j} = \begin{cases} 1/m_i & \text{if there is a link from the page } i \text{ to the page } j \\ 0 & \text{otherwise} \end{cases}$$

where m_i is the number of links contained in the page i. Let $v \in \mathbb{R}^n$ such that $Gv = v$, one can show that the i-th component of the vector v measures the importance of the page i.

However, many users tried to "hack" the Pagerank algorithm in a way that certain pages gain more importance. Moreover the whole Internet evolved quickly in the last decade. Therefore the Pagerank algorithm has also changed in different ways.

- *Personalization*: ideally the importance of a webpage depends on the user. What it is important for one user, it may be irrelevant for another. Therefore the Pagerank should adapt to the user features. Is he/she young? Where he/she lives? What are his/her interests?
- *Update*: computing the importance of the webpages is computationally demanding. Since websites and links are created and deleted continuously, the importance of the webpages is usually re-computed every month by Google. Is there a way to recycle the old results?
- *Communities*: it is important to divide the web-pages by communities/topic: sport, politics, news, movies, etc. How is this operation performed?

The project consists of the following tasks:

- Mathematically formulate the problem and understand the probabilistic theory behind the model.
- Identify numerical methods suitable for solving the problem. It is important to detect and exploit the properties of the problem that we will solve.
- Implement and improve the methods. The Pagerank method (and its variations) can be efficiently implemented in Matlab. Moreover datasets (pieces of the Internet) can be generated and imported in Matlab. The whole algorithm can be therefore tested on real data.
- Analyze the method theoretically and practically. How fast is the algorithm theoretically? How fast is practically?

References and further reading:

- Langville, Amy N., and Carl D. Meyer. Google's Pagerank and beyond: The science of search engine rankings. Princeton University Press, 2011.
- Langville, Amy N., and Carl D. Meyer. "Updating Markov chains with an eye on Google's Pagerank." SIAM journal on matrix analysis and applications 27.4 (2006): 968-987.
- https://www.seo-guide.se/pagerank

Supervision: Giampaolo Mele
contact: gmele@kth.se
co-supervisor: Elias Jarlebring