GOOGLE SEARCH ENGINE
Some of its mathematical aspects

Ivo Marek*

The Google search engine is based on a stochastic model, more precisely on a Markov chain model, interpreting the web page system as a graph of states (vertices) with referencing consisting of superlinks of the pages representing edges of this graph. In order to obtain the transition matrix as primitive and stochastic one, the model must be adapted appropriately. In the end the Google search engine matrix will have he following form

$$G = \alpha B_1 + (1 - \alpha)B_2,\ 0 < \alpha < 1,$$

where α is a suitable positive real number, B_1 is the matrix representing the web pages in the sense of the graph described above and B_2 denotes a *personalization matrix* acting in the model as perturbation guaranteeing, among other things, primitivity of matrix G. The choice of value α is a rather delicate question both as mathematics as well as practical aspects of exploiting the search engine concerns.

From viewpoint of mathematics the most involved part of the Google search engine is the Page Rank computation. This is realized in the above model as a kind of stationary probabilkity vector of some suitably constructed stochastic matrix. The stationary probability vector gives rise to forming an order of web pages that are to be offered subsequently to a random searcher as replies to his questions. True actions processed by Google result as implications of interactions of quite many parts or subsystems of the Google search engine exploiting various Computer Science models such as computer linguistics, learning systems, information theory, etc.

The lecture will be focused on description of the original Google project as well as on some of its possible improvements. In particular, in place of the personalization matrix being originally a rank-one matrix, a more complex though the same easily computable matrix is proposed. Naturally, introducing this new matrix will enlarge closer approach to personalization. Simultaneously, the power method applied for the rank-one perturbation

*Katedra matematiky, Stavební fakulta Českého vysokého učení technického, Thákurova 7, 160 00 Praha 1, Czech Republic.

personalization is to be exchanged by a variant of some of more efficient multi-level methods such as Iterative Aggregation/Disaggregation.

The size N of the Google matrix G, being equal to the number of all web pages on Internet, is huge: At present N is estimated by relation $3.10^9 \leq N \leq 4.10^9$. Obviously, this matrix is fictitious and just its actions are needed in the computations.

References:

[1] I. C.F. Ipsen, S. Kirkland *Convergence analysis of an improved pagerank algorithm.* Preprint 2004.

[2] A. N. Langville, C. D. Meyer. *Deeper Insight into PageRank.* Preprint 2004.

[3] I. Marek, P. Mayer *Iterative aggregation/disaggregation method for computation stationary probability vectors of stochastic matrices can be finitely terminating.* International Journal of Differential Equations Vol. **3**, p. 313 (2002).

Acknowledgement. Research on which this contribution is based has been partially supported by the Grant Agency of the Czech Republic under the Nr. 201/02/0595 and by grants MSMT 210000010 and MSMT 113200007. The support is sincerely acknowledged.