Markov Chain Analysis of the PageRank Problem

Nelly Litvak
University of Twente, Faculty of EEMCS
n.litvak@math.utwente.nl

The PageRank is a notion used by search engines to reflect a popularity and importance of a page based on its citation ranking. Such ranking was first introduced in 1998 by Google search engine [4]. The PageRank of a page i reflects the importance of this page basing on: 1) how many pages link to i, and 2) how important are the pages that link to i. Since the web changes very fast, the PageRank has to be regularly updated. Such update is an intricate task due to the huge size of the World Wide Web. Consequently, the analysis of the PageRank has become a hot topic with vast literature ranging from the original paper by Brin and Page [4], to the latest preprints by specialists in Markov chains, linear algebra, numerical methods, information retrieval, operations research, and other fields [11].

In the proposed PhD project, we shall concentrate on the Markov chain formulation of the PageRank problem. Specifically, we suggest to analyze the effectiveness of aggregation-disaggregation methods [15, 5, 14] in PageRank computation. Such methods exploit the block structure of the web and seem to be very promising [7]. The analysis will be based on the theory of discrete-time Markov chains, Perron-Frobenius theory, perturbation theory [8], and the theory of quasi-stationary distributions [6, 9]. The project will also involve extensive numerical studies.

A student may start with the following problem. Consider two completely disconnected communities (blocks of pages) and assume that they tailor several links to each other. Such strategy is called reprocicating and is widely used by web-administrators in hope to increase their ranking [12]. The question is whether the trick really works. In [2], we studied a completely decomposable web and we analyzed the situation when one of communities gives a link to another without receiving a link back. The results insinuate that in case of reprocicating, only one of the communities wins in ranking, whereas the other one looses. This issue however requires a rigorous analysis. The results will be interesting from the practical point of view, as they will either confirm or ruin the common reprocicating myth. Besides, it will be a useful and insightful first step in the analysis of aggregation-disaggregation methods in PageRank computation.

After the first problem is been (partly) solved, the direction of research might deviate from the original plan, depending on the used methods, obtained results, and the interests of the student. Possible directions could be, for instance, the analysis of on-line algorithms [1] or Monte Carlo methods [3, 10].

References

[1] Abiteboul, S., Preda, M. and Cobena, G. (2002) Adaptive on-line page importance computation. In *The Twelfth International World Wide Web Conference WWW2003*.

[2] Avrachenkov, K.E. and Litvak, N. Decomposition of the Google PageRank and optimal linking strategy. University of Twente Technical Report 1712.

[3] Breyer, L.A. (2002) Markovian page ranking distributions: some theory and simulations. Technical report.

[4] Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *7th International World Wide Web Conference*, Brisbane, Australia.

[5] Courtois, P.J. (1977) *Decomposability: Queueing and Computer System Applications.* Academic Press, New York.

[6] Darroch, J. N. and Seneta, E. (1965) On quasi-stationary distributions in absorbing discrete-time finite Markov chains. *J. Appl. Probab.* **2**, 88–100.

[7] Kamvar, S.D., Haveliwala, T.H., Manning, C.D. and Golub, G.H. (2003) Exploiting the Block Structure of the Web for Computing PageRank. Stanford University Technical Report.

[8] Kato, T. (1982) *A Short Introduction to Perturbation Theory for Linear Operators.* Springer-Verlag, New York-Berlin.

[9] Kijima, M. *Markov Processes for Stochastic Modeling.* Stochastic Modeling Series. Chapman & Hall, London, 1997.

[10] Litvak, N. Monte Carlo methods of PageRank computation.

[11] Langville, A.N., and Meyer, C.D. (2003) Deeper Inside PageRank. Preprint, North Carolina State University.

[12] Laycock, J. To Reciprocate or not to Reciprocate.
http://websearch.about.com/cs/resources/a/linkdebate.htm

[13] Litvak, N. Monte Carlo methods of PageRank computation.

[14] Schweitzer, P.J. and Kindle, K.W. (1986) An iterative aggregation-disaggregation algorithm for solving linear equations. *Appl. Math. Comput.* **18**(4), 313–353.

[15] Stewart, W.J. (1994) *An Introduction to the Numerical Solution of Markov Chains.* Princeton University Press.