Applications of Matrix Computations to Search Engines

With no doubt, Google is currently the most widely used search engine on the Web. Behind its success, there is a very effective, ingenious and at the same time simple algorithm that arranges Web Pages in order of certain notion of importance. The construction of such an algorithm, known as PageRank, is an elegant application of a great number of results in linear algebra and matrix computations.

The central part of the algorithm deals with computing a dominant eigenvector (called the PageRank vector) of a huge stochastic irreducible matrix (the number of columns is in the order of billions). For practical and theoretical reasons the original matrix is modified or perturbed by introducing a so-called personalization vector, which allows to manipulate not only the ranks of particular WebPages, but also the speed of convergence of the algorithm [2, 6].

The central idea is this project is to approach one or more of the unsolved or not completely answered problems involved in this process. Google uses a modification of the very simple power method, mainly because more advanced methods for computing eigenvectors are out of the scope due to the size of the matrix. However, other approaches are being investigated, including adaptive techniques [4], extrapolation methods [5], and partitions of the matrix according to groups of pages with no outlinks [7], all of them with the purpose of speeding up convergence by a certain factor. These and other new methods do not compete with one another and hence there is the possibility of combining two or more algorithms to obtain greater speeds of convergence.

A second and related problem deals with updating the PageRank vector. Since the computation of such eigenvector is very expensive, Google updates it only every few weeks, even though the Web itself is being updated continuously. In addition, the PageRank vector from a prior period is nearly useless for initializing the power method for the next period, so that new computations are restarted almost from scratch. There is currently active research on how to efficiently use previous computations to obtain a new PageRank vector. Worth mentioning are iterative aggregation techniques [3, 8], Monte Carlo methods and asymptotic analysis [1], and sequential updating [9]. Although these approaches yield some faster convergence, there is plenty of room for research in this direction. In particular, in [8] it is suggested that extrapolation techniques from [5] could possibly be combined with aggregation techniques to further accelerate the updating process.

On the other hand, there is the possibility of approaching both problems at the same time: the speeding up of the algorithm and the updating of the PageRank vector, due to their relationship to the solution of an associated linear system of equations. That is, instead of computing eigenvectors, one tries instead to solve some linear system, whose solution, should give the PageRank vector. At this stage, several numerical techniques for solving linear systems could be exploited; in particular, several reordering techniques can be applied for a more efficient algorithm and for exploiting the sparsity of the original matrix. This is a very interesting real-world application and a challenging research problem for REU participants.

Prerequisites: A basic background on linear algebra, and ideally some programming experience.

References

[1] K. Avrachenkov, N. Litvak, *The effect of new links on Google PageRank,* Stochastic Models **22,** 2 (2006), 319-331

[2] K. Bryan, T. Leise *The $25,000,000,000 eigenvector: the linear algebra behind Google,* SIAM Review **48,** 3 (2006), 569-581.

[3] I. Ipsen, S. Kirkland, Convergence analysis of a pagerank updating algorithm by Langville and Meyer, SIAM J. Matrix Analysis, 27m 4 (2006) 952-967.

[4] D. Kamvar, T. Haveliwala, G. Golub. *Adaptive methods for the computation of PageRank,* Linear Algebra Appl **386** (2004), 51-65.

[5] D. Kamvar, T. Haveliwala, C. Manning, G. Golub, *Extrapolation methods for accelerating PageRank computations*, Proceedings of the 12th International Conference on WWW (2003) 261-270.

[6] A. Langville, C. Meyer, *A survey of eigenvector methods for Web information retrieval,* SIAM Review **47,** 1 (2005), 135-161.

[7] A. Langville, C. Meyer, *A reordering for the PageRank problem,* SIAM J. Sci.Comput. **27,** 6 (2006), 2112-2120.

[8] A. Langville, C. Meyer, Updating *Markov chains with an eye on Google's PageRank,* SIAM J. Matrix. Analysis **27,** 4 (2006), 968-987.

[9] F. McSherry, *A uniform approach to accelerated PageRank computation*, Proceedings of the 14th International Conference on WWW (2005), 575-582.