PageRank project

Summer Institute 2007, Carnegie Mellon University

Jian Wang (U. Tennessee Knoxville)

June 8, 2007

Updated copy of this document and all other relevant files are going to be downloadable from `http://www.math.cmu.edu/~masha/REU/Voronoi/`.

1 Goals

Understand the math behind Google's million-dollar worth search engine. Study the relationship between the structure of the network and convergence of the numerical algorithms as well as their well-posedness. Investigate possible ways to compute dominant eigenvalues and their advantages and drawbacks. Taking as an example some model large matrix representation of the internet, develop a set of routines for computing page rank using existing methods such as power method and Gaussian elimination, as well as their possible modifications. Compare the results of the test and try to answer some of the open questions.

2 Long-term plan

5 weeks time frame

1. Get familiar with the software
2. Learn basic numerical analysis techniques using power method and Gaussian elimination(GE) as an example
3. Experiment with power method and GE applied to various stochastic matrices
4. Study the principles behind the PageRank search engine and the properties of the "internet matrix"
5. Implement several version of PageRank algorithm for a particular large-scale matrix representation of the World Wide Web
6. Try to accelerate basic routines developed at the previous stage, taking existing extrapolation and breadth-first search methods as a starting point
7. Explore analytical and numerical properties of both algorithms and propose possible modifications to make them more efficient

3 Short-term plan

3.1 Introduction to software, basic theory of PageRank - week 1

- Get familiar with Matlab environment: Matlab tutorial can be found here `http://www.amath.washington.edu/~calhoun/464/handouts/matlab_intro.pdf`
- learn how to create m-files, plot and manipulate data, work with matrices etc
- Understand how the network can be represented by means of a matrix and the meaning of the PageRank index.
- Look at a demo for PageRank: `pagerankdemo.m`. Can you write down the matrix for this network?
- Test your skills on the following example: Implement the Maple worksheet provided in `google.mws` in Matlab, provide a best possible visualization
- Refresh the theory of eigenvalues and eigenvectors; do exercises from [1]

3.2 Finding dominant eigenvector efficiently

This will be your next step.

References

[1] "The 25,000,000,000 eigenvector: the linear algebra behind Google", by Kurt Bryan and Tanya Leise

[2] " How Google Finds Your Needle in the Web's Haystack" by David Austin (available online at `http://www.ams.org/featurecolumn/archive/pagerank.html`)