Parallelization techniques for accelerating PageRank computation

H. Migallón∗, V. Migallón† and J. Penadés†

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. Let \( G = [g_{ij}]_{i,j=1}^n \) be a Web graph adjacency matrix with elements \( g_{ij} = 1 \) when there is a link from page \( j \) to page \( i \), with \( i \neq j \), and zero otherwise. From this matrix we can construct a transition matrix \( P = [p_{ij}]_{i,j=1}^n \) as follows: \( p_{ij} = \frac{g_{ij}}{c_j} \) if \( c_j \neq 0 \) and 0 otherwise, where \( c_j = \sum_{i=1}^n g_{ij}, 1 \leq j \leq n \), represents the number of out-links. For pages with a nonzero number of out-links the matrix \( P \) is column stochastic. In this case, the PageRank vector can be obtained by solving \( P\pi = \pi \). The Power method is one of the oldest and simplest iterative methods for solving this eigenvector problem. When the matrix \( P \geq O \) is irreducible and stochastic, the Power method converges to the eigenvector corresponding to \( \lambda_{\text{max}} = 1 \). However, the Web contains many pages without out-links. In this case, the matrix \( P \) is non-stochastic and the Power method cannot be used. Moreover, the matrix irreducibility is not satisfied for a Web graph. In order to overcome these difficulties, Page and Brin change the transition matrix \( P \) to a column stochastic matrix \( \bar{P} = \alpha(P + vd^T) + (1 - \alpha)v^Tv^T \), where \( d \in \mathbb{R}^n \) is defined by \( d_i = 1 \) if and only if \( c_i = 0 \) and the vector \( v \in \mathbb{R}^n \) is some probability distribution over pages. Then, setting \( \alpha \) such that \( 0 < \alpha < 1 \), the Power method can be used to solve the stationary distribution of the ergodic Markov chain defined by \( \bar{P}\pi = \pi \).

Following [1], the PageRank vector can be obtained solving a linear system. Thinking about PageRank as a sparse linear system opens new research lines combining iterative techniques and parallel processing. In this work we propose iterative techniques for accelerating the parallel calculation of the PageRank vector. These iterative schemes allow us to reduce the number of global iterations by eliminating synchronization points at which a process must wait for information from other processes. The parallel implementation has been developed using a mixed MPI/OpenMP model to exploit parallelism beyond a single level. That is, these programming models have been combined into a hybrid paradigm in which MPI is used for data distribution among nodes and OpenMP to exploit loop level parallelism within each node. In order to investigate and analyze the proposed parallel algorithms, we have used several realistic large datasets. The numerical results show that the proposed algorithms can significantly speed up the convergence time with respect to the parallel Power algorithm.

Acknowledgements

This research was supported by the Spanish Ministry of Economy and Competitiveness and the European Commission (FEDER funds) under Grant Number TIN2015-66972-C5-4-R.

References
