Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational and Applied Mathematics 210 (2007) 159–166
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

Comparison of Krylov subspace methods on the PageRank problem☆

Gianna M. Del Corso[a,*], Antonio Gullí[a,b], Francesco Romani[a]

[a] *Dipartimento di Informatica, Università di Pisa, Italy*
[b] *Ask Jeeves/Teoma, Pisa, Italy*

Abstract

PageRank algorithm plays a very important role in search engine technology and consists in the computation of the eigenvector corresponding to the eigenvalue one of a matrix whose size is now in the billions. The problem incorporates a parameter α that determines the difficulty of the problem. In this paper, the effectiveness of stationary and nonstationary methods are compared on some portion of real web matrices for different choices of α. We see that stationary methods are very reliable and more competitive when the problem is well conditioned, that is for small values of α. However, for large values of the parameter α the problem becomes more difficult and methods such as preconditioned BiCGStab or restarted preconditioned GMRES become competitive with stationary methods in terms of Mflops count as well as in number of iterations necessary to reach convergence.

MSC: 65F15; 65Y20; 15A18

Keywords: Search engines; Krylov subspace methods; Large and sparse linear systems

1. Introduction

The PageRank algorithm is one of the key ingredients of search engine systems for assigning a rank of importance to web pages [6,14]. The computation of the PageRank vector consists in computing the dominant eigenvector of a stochastic matrix whose size is now of some billions [10]. Numerical analysts are faced with the big challenge of computing the PageRank vector with methods which are lightweight in terms of the space required and fast enough to guarantee that the ranking score of every page can be re-computed frequently.

The standard way to compute PageRank is the Power method since it converges for every choice of a nonnegative starting vector and it requires to store just two vectors. Memory space is, in fact, the most crucial resource for this problem due to the extremely big size of the web. However, despite the low requirement of space of the Power method its rate of convergence can be very slow. Many suggestions have been proposed for speeding up the convergence of the Power method such as extrapolation techniques [3,5,16], adaptive techniques [11] or permutation strategies for exploiting the block structure of the web matrix [12]. In [9] the effectiveness of algorithms based on a restarted version

☆ Partially supported by the GNCS-INDAM Project: "Problematiche Numeriche nel WEB".
* Corresponding author.
E-mail address: delcorso@di.unipi.it (G.M. Del Corso).

0377-0427/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2006.10.080

of Arnoldi process has been tested; the algorithms rely on the knowledge of the largest eigenvalue of the matrix and the sensitivity of PageRank problem is also discussed there.

Recently, in [7], the authors of this paper proved that the PageRank vector can also be computed as the solution of an equivalent sparse linear system by exploiting the effectiveness of stationary methods such as Jacobi, Gauss–Seidel and reverse Gauss–Seidel methods for the solution of the system. Moreover, in [7] many different permutation schemes have been applied to the web matrix for increasing data-locality and reducing the time necessary for computing the PageRank vector. For some iterative methods the reordering of the rows and columns of the matrix leads to an iteration matrix with a lower spectral radius than the one corresponding to the nonpermuted one and hence to faster methods.[1] The experiments performed in [7] show that these strategies allow a gain of up to 90% of the computational time needed to compute PageRank. Among the reordering techniques proposed in [7], BFS strategies of visit of the web graph have been considered. The matrices permuted according to these strategies of visit have a block triangular structure. In particular, in the permuted matrix it is possible to identify a large irreducible connected component, containing the so-called "core" and a tail formed of smaller and reducible blocks. The block triangular structure of the reordered web matrix suggests us to exploit the reducibility and moreover, that we can use the most convenient technique for computing the PageRank of pages belonging to different connected components of the web. The size of the larger connected component is so big that it is not convenient to employ on that block nonstationary methods because they need to store many vectors per iteration. On the contrary, the remaining diagonal part of the matrix, which we call "tail" is composed of many smaller components on which Krylov methods seem promising. The idea is to use faster methods—which have however more storage requirements—on matrices whose size is the maximum possible to solve the problem in main memory. In order to test the behavior of some of the most popular iterative methods on web matrices, we consider 1120 matrices obtained by taking different portions of the tail of a 24 million matrix resulting from a web crawl. On these web matrices we apply many stationary and nonstationary methods reporting statistical results on the robustness and effectiveness of these methods for this kind of problems.

From our experimental results it turns out that Gauss–Seidel is a very good method also when compared with Krylov subspace methods combined with an ILU(0) preconditioning technique. In fact, nonstationary methods are in general not very robust if used without a preconditioning technique and only preconditioned BiCGStab turns out to be competitive with Gauss–Seidel method.

The paper is organized as follows: in Section 2 we describe the PageRank algorithm for assigning a rank of importance to web pages and we describe how the PageRank vector can be obtained as the solution of a sparse linear system. Section 3 contains a description of the experimental setting and of the stationary and nonstationary methods considered in this paper, while experimental results are reported in Section 4. Section 5 contains conclusion and further works.

2. The PageRank model: a linear system approach

Link-based ranking techniques view the Web as a directed graph (the Web Graph) $G = (V, E)$, where each of the N pages is a node and each hyperlink is an arc. The problem of ranking web pages consists in assigning a rank of importance to every page based only on the link structure of the Web and not on the actual contents of the page. The intuition behind the PageRank algorithm is that a page is "important" if it is pointed by other pages which are in turn "important". This definition suggests an iterative fixed-point computation for assigning a ranking score to each page in the Web. Formally, in the original model [14], the computation of the PageRank vector is equivalent to the computation of $\mathbf{z}^{\mathrm{T}} = \mathbf{z}^{\mathrm{T}} P$, where P is the adjacency matrix of the Web graph G normalized so that each row sums to 1. This model has unfortunately two problems: the first is the presence of dangling nodes, that is pages without outlinks, the second is the reducibility of the matrix that does not guarantee the uniqueness of a unitary norm eigenvector corresponding to the eigenvalue 1. These problems can be solved introducing a parameter α and considering the matrix $\widehat{P}(\alpha)$ instead of P, defined as $\widehat{P} = \widehat{P}(\alpha) = \alpha(P + \mathbf{d}\mathbf{v}^{\mathrm{T}}) + (1 - \alpha)\ \mathbf{e}\mathbf{v}^{\mathrm{T}}$, where $\mathbf{e}$ is the vector with all entries equal to 1, and $\mathbf{v}$ is a *personalization vector* which records a generic user's preference for each page in V. α is a constant, $0 < \alpha < 1$, which is related to the probability that a random web surfer visiting a page follows a link in that page rather then jumping to any other page in the web. In our experiments we chose a uniform personalization vector, which means that $\mathbf{v} = 1/N\ \mathbf{e}$. See [13] for a deaper treatment on the characteristics of the model.

[1] This is not the case for the Jacobi method whose spectral radius is not affected by reordering techniques. However the spectral radius may change for the Gauss–Seidel or reverse Gauss–Seidel methods.