Bol. Soc. Esp. Mat. Apl. n°0 (0000), 1–6

Mathematical Properties and Analysis of Google's PageRank

Ilse C.F. Ipsen, Rebecca S. Wills

Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA

ipsen@ncsu.edu, rmwills@ncsu.edu

Abstract

To determine the order in which to display web pages, the search engine Google computes the *PageRank* vector, whose entries are the PageRanks of the web pages. The PageRank vector is the stationary distribution of a stochastic matrix, the *Google matrix*. The Google matrix in turn is a convex combination of two stochastic matrices: one matrix represents the link structure of the web graph and a second, rank-one matrix, mimics the random behaviour of web surfers and can also be used to combat web spamming. As a consequence, PageRank depends mainly the link structure of the web graph, but not on the contents of the web pages. We analyze the sensitivity of PageRank to changes in the Google matrix, including addition and deletion of links in the web graph.

Due to the proliferation of web pages, the dimension of the Google matrix most likely exceeds ten billion. One of the simplest and most storage-efficient methods for computing PageRank is the power method. We present error bounds for the iterates of the power method and for their residuals.

Palabras clave : *Markov matrix, stochastic matrix, stationary distribution, power method, perturbation bounds*

Clasificación por materias AMS : *15A51,65C40,65F15,65F50,65F10*

1. Introduction

How does the search engine Google determine the order in which to display web pages? The major ingredient in determining this order is the PageRank vector, which assigns a score to a every web page. Web pages with high scores are displayed first, and web pages with low scores are displayed later. The PageRank of a web page is based on the link structure of the web graph and

does not depend on the content of web pages. The importance of PageRank is emphasized in one of Google's web pages [1]:

> The heart of our software is PageRank™, a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And while we have dozens of engineers working to improve every aspect of Google on a daily basis, PageRank continues to provide the basis for all of our web search tools.

The PageRank vector is the stationary distribution of a stochastic matrix, called the Google matrix. In §2 we describe the Google matrix and define the PageRank vector. The sensitivity of PageRank to changes in the Google matrix is analyzed in §3, and the power method for computing PageRank is presented in §4.

2. PageRank and the Google Matrix

The link structure of the web graph can be represented mathematically as a matrix H [9]. Suppose web page i has $l_i > 0$ outlinks. If page i contains a link to another page $j \neq i$, then $H_{ij} = 1/l_i$, otherwise, $H_{ij} = 0$. Matrix element H_{ij} represents the likelihood that a surfer follows the link from page i to page j. If web page i has no outlinks then row i of H is zero. Such as web page, called a *dangling node*, can be a pdf file or a page whose links have not yet been crawled.

To transform H into a stochastic matrix[1] S, one can fill every row corresponding to a dangling node with a vector w^T. That is, $S \equiv H + dw^T$, where $d_i = 1$ if page i has no outlinks, and $d_i = 0$ otherwise; and w is a column vector with $w \geq 0$ and $\|w\|_1 = 1$. A popular choice is to set to $1/n$ every element in the dangling node rows, where n is the number of nodes in the web graph; in other words $w = \frac{1}{n}\mathbf{1}$, where $\mathbf{1}$ is the column vector of all ones.

The Google matrix is defined as a convex combination of S and a rank-one matrix, i.e.

$$G \equiv \alpha S + (1-\alpha)\mathbf{1}v^T, \qquad 0 \leq \alpha < 1, \qquad v \geq 0, \quad \|v\|_1 = 1.$$

The *damping factor* α, originally set to .85, models the possibility that a web surfer jumps from one web page to the next without necessarily following a link [3]. The *personalization vector* v can be used to combat link spamming [7].

The matrix G is row stochastic and, in general, reducible. However it has a distinct dominant eigenvalue. To see this, denote the eigenvalues of S by $\lambda_1(S) = 1$ and $\lambda_i(S)$, $i \geq 2$, where $|\lambda_i(S)| \leq 1$. The eigenvalues of G are 1 and $\alpha\lambda_i(S)$, $i \geq 2$ [5]. Due to the uniqueness of the dominant eigenvalue, the stationary distribution π of G is unique. Therefore the *PageRank* vector is defined as the stationary distribution π of G,

$$\pi^T G = \pi^T, \qquad \pi \geq 0, \quad \|\pi\|_1 = 1.$$

[1] A real square matrix is stochastic if all its elements lie between 0 and 1, and the elements in each row sum to 1.

The ith entry of π is the PageRank for web page i.

3. Sensitivity of PageRank

We show that the sensitivity of the PageRank vector to changes in the matrix S, in the personalization vector v and in the damping factor α is governed by the damping factor α; and that PageRank can be considered insensitive to changes in G.

Perturbation theory for stationary distributions of Markov chains is well understood, see for instance [4]. The results presented here exploit the particular structure of the Google matrix. Several of these have already appeared in the literature, but our proofs are rigorous and simple [8]. The proofs make use of the fact that the eigenvector problem $\pi^T G = \pi^T$, $\|\pi\|_1 = 1$ is mathematically equivalent to a system of linear equations whose coefficient matrix is a strictly row diagonally dominant M-matrix [2]

$$\pi^T(I - \alpha S) = (1 - \alpha)v^T,$$

as well as to a linear system whose right-hand side does not depend on α,

$$\pi^T \left(I - \alpha(S - \mathbf{1}v^T)\right) = v^T.$$

The equivalence of the eigenvector problem and the linear systems follows from $\|\pi\|_1 = \pi^T\mathbf{1} = 1$.

3.1. Changes in the Matrix S

The sensitivity of the PageRank vector π to changes in S depends on a condition number that is bounded by $\alpha/(1-\alpha)$.

Specifically, let $S + E$ be a stochastic matrix, and set

$$\tilde{G} \equiv \alpha(S + E) + (1 - \alpha)\mathbf{1}v^T.$$

The perturbed PageRank vector is $\tilde{\pi}$, where $\tilde{\pi}^T\tilde{G} = \tilde{\pi}^T$, $\tilde{\pi} \geq 0$, and $\|\tilde{\pi}\|_1 = 1$. We obtain for the absolute error in $\tilde{\pi}$,

$$\tilde{\pi}^T - \pi^T = \alpha\tilde{\pi}^T E(I - \alpha S)^{-1}, \qquad \|\tilde{\pi} - \pi\|_1 \leq \frac{\alpha}{1-\alpha}\,\|E\|_\infty.$$

For the original damping factor $\alpha = .85$ $\alpha/(1-\alpha) \approx 5.7$. Even for larger damping factors, the sensitivity is still low: If $\alpha = .99$ then $\alpha/(1-\alpha) = 99$.

3.2. Changes in the Damping Factor α

The sensitivity of the PageRank vector π to changes in the damping factor α depends on a condition number that is bounded by $2/(1-\alpha)$.

Specifically, let $0 \leq \alpha + \mu < 1$ be a perturbed damping factor, and set $\tilde{G} \equiv (\alpha+\mu)S + (1-(\alpha+\mu))\mathbf{1}v^T$. The perturbed PageRank vector is $\tilde{\pi}$, where $\tilde{\pi}^T\tilde{G} = \tilde{\pi}^T$, $\tilde{\pi} \geq 0$, and $\|\tilde{\pi}\|_1 = 1$. The error in $\tilde{\pi}$ can be bounded by

$$\|\tilde{\pi} - \pi\|_1 \leq \frac{2}{1-\alpha}\,|\mu|.$$

The condition number bound $2/(1-\alpha)$ is an increasing function in α. Comparing this bound to the bounds for condition number $\alpha/(1-\alpha)$ in §3.1 shows that π is slightly more sensitive to changes in the parameter α than to changes in the matrix S. For the original damping factor $\alpha = .85$, the condition number is $2/(1-\alpha) \approx 13.4$. For $\alpha = .99$, we get $2/(1-\alpha) = 200$.

3.3. Changes in the Personalization Vector v

The PageRank vector π is perfectly conditioned with regard to changes in the personalization vector v.

Specifically, let $v+f$ be the perturbed personalization vector with $v+f \geq 0$ and $\|v+f\|_1 = 1$; and set $\tilde{G} \equiv \alpha S + (1-\alpha)\mathbf{1}(v+f)^T$. The perturbed PageRank vector is $\tilde{\pi}$, where $\tilde{\pi}^T\tilde{G} = \tilde{\pi}^T$, $\tilde{\pi} \geq 0$, and $\|\tilde{\pi}\|_1 = 1$. The error bound for $\tilde{\pi}$ contains a condition number that is bounded by one,

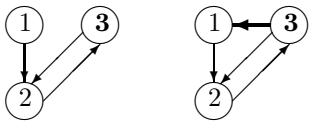
$$\|\tilde{\pi} - \pi\|_1 \leq \|f\|_1.$$

3.4. Addition of Inlinks

Adding an inlink to a web page increases its PageRank. Specifically, if a link is added from webpage j to web page $l \neq j$ and if web page j does not have a link to itself then the PageRank of page l increases, i.e. $\tilde{\pi}_l > \pi_l$.

3.5. Addition of Outlinks

Adding an outlink to a web page can decrease the PageRank. In the following example of a web graph with 3 web pages we add an outlink from page 3 to page 1:



Original Modified

The PageRanks for web page 3 before and after addition of the link are

$$\pi_3 = \frac{1+\alpha+\alpha^2}{3(1+\alpha)} > \frac{1+\alpha+\alpha^2}{3(1+\alpha+\alpha^2/2)} = \tilde{\pi}_3.$$

Hence, adding an outlink from page 3 to page 1 decreases the PageRank for page 3 from π_3 to $\tilde{\pi}_3$.

Although adding an outlink may decrease the PageRank of an individual web page, we can still bound the total change in the entire PageRank vector. If outlinks are added to and/or deleted from web page j then the new PageRank vector $\tilde{\pi}$ differs from the old one by

$$\|\tilde{\pi} - \pi\|_1 \leq \frac{2\alpha}{1-\alpha}\, \tilde{\pi}_j.$$

Thus adding and deleting outlinks does not change the entire PageRank vector significantly, provided the new PageRank of page j is not too large.

4. Computing PageRank

The definition of PageRank $\pi^T G = \pi^T$ implies that π is a left eigenvector of G associated with the dominant eigenvalue 1. The simplest way to compute π is to apply the power method to G [9].

Pick $x^{(0)} > 0$, $\|x^{(0)}\|_1 = 1$, $k = -1$
Repeat $k = k + 1$, $\quad [x^{(k+1)}]^T = [x^{(k)}]^T G$
until $\|x^{(k+1)} - x^{(k)}\| \leq \tau$

The difference of successive iterates in the stopping criterion is just the residual, $[x^{(k+1)}]^T - [x^{(k)}]^T = [x^{(k)}]^T G - [x^{(k)}]^T$. The norm can be the one-, two-, or infinity-norm. The parameter τ often lies between 10^{-8} and 10^{-4}.

Although the matrix $G = \alpha S + (1-\alpha)\mathbf{1}v^T$ is dense, matrix multiplication with G can be performed in a sparse manner by exploiting that $S = H + dw^T$, see §2. Thus matrix vector multiplication of G with a vector $x \geq 0$, $\|x\|_1 = 1$ amounts to:

$$x^T G = \alpha x^T H + (\alpha x^T d) w^T + (1-\alpha) v^T.$$

This is a sparse multiplication with H, followed by adding multiples of the vectors w^T and v^T. The term $x^T d$ is obtained by adding of all components of x corresponding to dangling nodes. The cost of matrix vector multiplication with G is proportional to the number of non-zeros in H, i.e. the number of links in the web graph.

From the expressions for the eigenvalues of G in §2 follows that the power method converges (in exact arithmetic), with an asymptotic convergence rate bounded by α. This is also reflected in the error bounds for the iterates of the power method and their residuals,

$$\|x^{(k)} - \pi\|_{1,\infty} \leq 2\alpha^k, \qquad \|[x^{(k)}]^T G - [x^{(k)}]^T\|_{1,\infty} \leq 2\alpha^k.$$

Another way to compute PageRank is as the solution to the linear system $\pi^T(I - \alpha S) = (1-\alpha)v^T$, see §3, via stationary iterative methods (such as the Jacobi method) or Krylov subspace methods (such as BiCGSTAB), see for instance [6].

References

[1] *http://www.google.com/technology/index.html*.

[2] A. ARASU, J. NOVAK, AND J. TOMKINS, A. AMD TOMLIN, *PageRank computation and the structure of the web: Experiments and algorithms*, in Proc. Eleventh International World Wide Web Conference (WWW2002), ACM Press, 2002.

[3] S. BRIN AND L. PAGE, *The anatomy of a large-scale hypertextual web search engine*, Comput. Networks and ISDN Systems, 30 (1998), pp. 107–17.

[4] G. E. CHO AND C. D. MEYER, *Comparison of perturbation bounds for the stationary distribution of a Markov chain*, Linear Algebra Appl., 335 (2001), pp. 137–150.

[5] L. ELDEN, *The eigenvalues of the Google matrix*, Tech. Rep. LiTH-MAT-R-04-01, Department of Mathematics, Linköping University, Sweden, December 2003.

[6] D. GLEICH, L. ZHUKOV, AND P. BERKHIN, *Fast parallel PageRank: A linear system approach*, tech. rep., Yahoo!, 2004.

[7] Z. GYÖNGYI, H. GARCIA-MOLINA, AND J. PEDERSEN, *Combating web spam with TrustRank*, in Proc. 30th International Conference on Very Large Databases, Morgan Kaufmann, 2004, pp. 576–587.

[8] I. C. F. IPSEN, *Numerical analysis of PageRank*. In preparation.

[9] L. PAGE, S. BRIN, R. MOTWANI, AND T. WINOGRAD, *The PageRank citation ranking: Bringing order to the web*, tech. rep., Stanford Digital Library Technologies Project, 1998.