Some numerical analysis problems behind web search

C. Brezinski M. Redivo Zaglia

Abstract

An important problem in web search is to classify the pages according to their importance. From the mathematical point of view, Google treats this problem by finding the left principal eigenvector (the PageRank vector) of a certain matrix. Properties of this vector will be given. It could be computed by the power method whose iterates will be characterized. Then several approximations of the PageRank vector, and procedures for accelerating the power method will be discussed.

Introduction

A query to a web search engine often produces a very long list of answers because of the enormous number of pages (over 8 billions in Google's database). To help the surfer, these pages have to be listed starting from the most relevant ones. Google uses several metrics and strategies for solving this *ranking* problem.

The importance of a page is called its *PageRank* and one of the main ingredients of Google's link analysis is the *PageRank algorithm* [3, 6]. A page is considered to be important if many other important pages are pointing to it. So, the importance of a page is determined by the importance of the other pages. This means that the row vector $\mathbf{r}^T$ of all PageRanks is only defined implicitly as the solution of a fixed–point problem, as we will see.

1 The PageRank problem

Let $\deg(i) \geq 1$ be the outdegree (that is, the number of pages it points to) of the page i. Let $P = (p_{ij})$ be the matrix which describes the transition from the page i to the page $j \neq i$ with $p_{ij} = 1/\deg(i)$, and $p_{ii} = 0$.

The PageRank vector $\mathbf{r}^T$ satisfies $\mathbf{r}^T = \mathbf{r}^T P$, that is, $\mathbf{r} = P^T \mathbf{r}$, and it can be computed recursively by the standard power method

$$\mathbf{r}^{(n+1)} = P^T \mathbf{r}^{(n)}, \quad n = 0, 1, \ldots,$$

assuming that all the eigenvectors of P^T are present in the spectral decomposition of $\mathbf{r}^{(0)}$. Unfortunately, this iterative procedure has convergence problems.

For avoiding these drawbacks, the original PageRank algorithm was modified. First, since some pages have no outlink, P is not stochastic. So, P is replaced by another matrix $\widetilde{P}$. Let $\mathbf{w} = (w_1, \ldots, w_p)^T \in \mathbb{R}^p$ be a probability vector, that is such that $\mathbf{w} \geq 0$ and

$\mathbf{e}^T\mathbf{w} = 1$ with $\mathbf{e} = (1, \ldots, 1)^T$, and p the total number of pages. Let $\mathbf{d} = (d_i) \in \mathbb{R}^p$ be the vector with $d_i = 1$ if $\deg(i) = 0$, and 0 otherwise. We set

$$\widetilde{P} = P + \mathbf{d}\mathbf{w}^T.$$

The effect of the additional matrix $\mathbf{d}\mathbf{w}^T$ is to modify the probabilities so that a surfer visiting a page without outlinks jumps to another page with the probability distribution defined by $\mathbf{w}$. Thus, $\widetilde{P}$ is stochastic, and has 1 as a dominant eigenvalue with $\mathbf{e}$ as its corresponding right eigenvector.

Another problem arises since $\widetilde{P}$ is reducible. In that case, $\widetilde{P}$ can have several eigenvalues on the unit circle, thus causing convergence problems to the power method. Moreover, $\widetilde{P}$ can have several left eigenvectors corresponding to its dominant eigenvalue 1.

Then, $\widetilde{P}$ itself is finally replaced by the matrix

$$P_c = c\widetilde{P} + (1-c)E, \qquad E = \mathbf{e}\mathbf{v}^T,$$

with $c \in [0, 1]$, and $\mathbf{v}$ a probability vector. It corresponds to adding to all pages a new set of outgoing transitions with small probabilities. The probability distribution given by the vector $\mathbf{v}$ can differ from a uniformly distributed vector, and the resultant PageRank can be biased to give preference to certain kinds of pages. The matrix P_c is stochastic and irreducible since $\mathbf{v}$ is a positive vector. P_c has an eigenvalue equal to 1 with $\mathbf{e}$ as its corresponding right eigenvector. Indeed

$$P_c\mathbf{e} = c\widetilde{P}\mathbf{e} + (1-c)\mathbf{e}\mathbf{v}^T\mathbf{e} = c\mathbf{e} + (1-c)\mathbf{e} = \mathbf{e}.$$

The power iterations for the matrix P_c^T now converge to a unique vector $\mathbf{r}_c$ (obviously, depending on c) which is chosen as the PageRank vector.

2 The power method

Thus, we are faced to the following mathematical problem. For consistency to prior works, we set $A_c = P_c^T$.

The $p \times p$ matrix A_c has eigenvalues $|\lambda_p| \leq \cdots \leq |\lambda_2| < \lambda_1 = 1$, and we have to compute $\mathbf{r}_c$, its unique right eigenvector corresponding to the eigenvalue $\lambda_1 = 1$. For that purpose, we use the power method which consists in the iterations

$$\mathbf{r}_c^{(n+1)} = A_c\mathbf{r}_c^{(n)}, \qquad n = 0, 1, \ldots$$

with $\mathbf{r}_c^{(0)}$ given.

The sequence $(\mathbf{r}_c^{(n)})$ always converges to $\mathbf{r}_c$ but, if $c \simeq 1$, the convergence is slow since the power method converges as c^n. So, a balance has to be found between a small value of c, which insures a fast convergence of $(\mathbf{r}_c^{(n)})$, but to a vector $\mathbf{r}_c$ which is not close to the real PageRank vector $\widetilde{\mathbf{r}} = \lim_{c\to 1} \mathbf{r}_c$, and a value of c close to 1, which leads to a better approximation $\mathbf{r}_c$ of $\widetilde{\mathbf{r}}$, but with a slow convergence. Google usually chooses $c = 0.85$, which insures a good rate of convergence.

3 Approximation and acceleration

Since computing a PageRank vector can take several days, convergence acceleration or approximations of the PageRank vector are essential, in particular, for providing continuous updates to ranking. Moreover, some recent approaches require the computation of several PageRank vectors corresponding to different personalization vectors. Recently, several methods for accelerating the computation of the PageRank vector by the power method were proposed [5, 4]. The aim of this work is to give a theoretical justification to the methods of [5], and to put them on a firm theoretical basis. We will interpret them in a different way, and simplify, unify, and generalize them. In particular, we will explain their connection with the method of moments of Vorobyev. Other possible acceleration procedures will also be discussed.

Another problem related to PageRank computations is that, as c approaches 1, the matrix A_c becomes more and more ill conditioned since its condition number behaves as $(1-c)^{-1}$, the conditioning of the eigenproblem becomes poor, and $\mathbf{r}_c$ cannot be computed accurately. So, $\mathbf{r}_c$ can be computed for several values of c far away from 1 by any procedure, and then these vectors can be extrapolated at the point $c = 1$ (or at any other point). In order for an extrapolation procedure to work well, the extrapolating function has to mimic as closely as possible the behaviour of $\mathbf{r}_c$ with respect to the parameter c.

Since P_c is stochastic and irreducible, $\mathbf{r}_c$ is the unique right eigenvector of $A_c = P_c^T$ corresponding to the eigenvalue 1, that is, $A_c\mathbf{r}_c = \mathbf{r}_c$. We have $\mathbf{r}_c \geq 0$, and we normalize it such that it is a probability vector, that is $\mathbf{e}^T\mathbf{r}_c = 1$.

So, we will study the properties of this vector, and, in particular, we will give implicit and explicit expressions for it. Then, we will discuss its computation by the power method. The iterates given by the power method are, in fact, the partial sum of a power series with vector coefficients [1]. This discussion will lead us to various procedures for accelerating the convergence of the power method, and to processes for the approximation of the PageRank vector. In particular, the iterates of the power method, which are in fact the partial sums of a vector formal power series, will be used for constructing Padé style approximations of $\mathbf{r}_c$. The convergence of the power method itself will be accelerated by using various vector sequences transformations, in particular, the ε-algorithms and Aitken's Δ^2 process. The acceleration processes proposed in [5] will be put on a firm theoretical basis and explained in the framework of the method of moments. They will also be generalized.

All these results are explained in details in [2].

References

[1] P. Boldi, M. Santini, S. Vigna, PageRank as a function of the damping factor, Poster Proceedings of the 14th International World Wide Web Conference, May 10-14, 2005, Chiba, Japan.

[2] C. Brezinski, M. Redivo–Zaglia, The PageRank vector: properties, computation, approximation, and acceleration, submitted.

[3] S. Brin, L. Page, The anatomy of a large–scale hypertextual web search engine, Comput. Networks ISDN Syst., 30 (1998) 107–117.

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

[5] S.D. Kamvar, T.H. Haveliwala, C.D. Manning, G.H. Golub, Extrapolations methods for accelerating PageRank computations, in *Proceedings of the Twelfth International World Wide Web Conference*, ACM Press, 2003, pp. 261-270.

[6] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: bringing order to the Web, Stanford University Technical Report, 1999, http://dbpubs.stanford.edu/pub/1999-66

Claude Brezinski
Laboratoire Paul Painlevé, UMR CNRS 8524
UFR de Mathématiques Pures et Appliquées
Université des Sciences et Technologies de Lille
59655 - Villeneuve d'Ascq cedex
France
claude.brezinski@univ-lille1.fr

Michela Redivo Zaglia
Dipartimento di Matematica Pura ed Applicata
Università degli Studi di Padova
Via G.B. Belzoni, 7
35131 - Padova
Italy
Michela.RedivoZaglia@unipd.it