Mathematical and Computer Modelling 57 (2013) 2248–2256

Contents lists available at SciVerse ScienceDirect

Mathematical and Computer Modelling

journal homepage: www.elsevier.com/locate/mcm

On some analytical properties of a general PageRank algorithm

Andrei Bourchtein *, Ludmila Bourchtein

Institute of Physics and Mathematics, Pelotas State University, Brazil

ARTICLE INFO

Article history:
Received 30 October 2010
Received in revised form 13 June 2011
Accepted 13 June 2011

Keywords:
PageRank
Stochastic matrix
Markov chain
Convergence
Stationary distribution

ABSTRACT

In this study, we present a theoretical analysis of some properties of a general algorithm for computation of the PageRank vector. It is shown that the convergence to the PageRank vector is generally nonuniform. Expressions for the limiting forms of the principal matrix of the PageRank approximation and the PageRank vector are derived in terms of the original stochastic matrix and the personalization vector. Some implications of the obtained results for ranking web pages are discussed.

1. Introduction

The methods for finding the PageRank vector, including the formation of the original stochastic matrix, the effect of the personalization vector, the reasoning behind the PageRank approach, and the application of different numerical algorithms, are well documented in a number of sources (see, for example [1–4] and references therein). So in this section, we just give a very brief description of the problem focusing on the points important for our study.

The original matrix A of the PageRank algorithm is defined considering the web of N pages as a directed graph whose vertices are web pages and edges are links. The entries of $A = (a_{ij})$ are defined as follows:

$$a_{ij} = \begin{cases} 1/d_i, & \text{if } d_i \geq 1 \text{ and there is outlink from } i \text{ to } j; \\ 0, & \text{otherwise,} \end{cases}$$

where d_i denotes the number of outlinks from page i. The rank of each of the N pages is defined as the principal (left-hand nonnegative normalized) eigenvector of A corresponding to the eigenvalue $\lambda = 1$.

There are some problems with such a definition of the rank related to the structure of the original matrix A. The first is the existence of dangling nodes ($d_i = 0$ for ith page), because there are pages with no outlinks [5,1,2]. This can lead to the non-existence of the eigenvalue $\lambda = 1$, the non-uniqueness of the principal eigenvector and artificially increased ranks for the dangling nodes. To remedy this problem, the original row-substochastic matrix A is transformed to the row-stochastic matrix $S = (s_{ij})$ by substituting zero rows corresponding to the dangling nodes by a stochastic (row) vector $w^T = (w_1, \ldots, w_n)$, that is:

$$s_{ij} = \begin{cases} 1/d_i, & \text{if } d_i \geq 1 \text{ and there is outlink from } i \text{ to } j; \\ w_j, & \text{if } d_i = 0; \\ 0, & \text{otherwise.} \end{cases}$$

* Corresponding address: Rua Anchieta 4715 bloco K, ap.304 Pelotas - RS, 96015-420, Brazil. Tel.: +55 53 32757343; fax: +55 53 32757343.
E-mail address: bourchtein@gmail.com (A. Bourchtein).

0895-7177/$ – see front matter
doi:10.1016/j.mcm.2011.06.033

In practice, the vector $w = \frac{1}{N}e$, $e = (1, \ldots, 1)^T$ is frequently used [1,2]. The modified N-order matrix S can be considered as the transition matrix of a finite discrete-time homogeneous Markov chain with state space consisting of web sites and matrix entries s_{ij} interpreted as the probability to move from site i to site j. Then the problem of ranking web sites is considered as the problem of finding the PageRank vector π, that is, the left-hand nonnegative normalized ($||\pi||_1 = 1$) eigenvector of the matrix S corresponding to the eigenvalue $\lambda = 1$ (hereinafter we use the 1-norm [6]). In the theory of Markov chains, such a vector is also called the stationary distribution vector.

Another problem is the existence of cyclic paths: some pages can form subchains resulting in a non-unique stationary distribution vector [5,2,7]. To circumvent this difficulty, the computational matrix $G(\alpha)$ is defined as $G(\alpha) = \alpha S + (1-\alpha)V$, where V is a rank-one row-stochastic matrix defined by a stochastic vector (called personalization vector) v in the form $V = ev^T$ and $\alpha \in [0, 1]$ is a scalar called teleportation parameter or damping factor [5,2,7]. The algorithms of the PageRank computation are essentially based on finding the Perron vectors (the left-hand positive normalized eigenvectors for the eigenvalue $\lambda = 1$) for matrices $G(\alpha)$ with different α and considering the PageRank vector as the limit of such vectors as α approaches 1. We will refer to this procedure as the general PageRank approximation or the PageRank algorithm. In practice, different approximations to that limiting vector are considered as the PageRank vector. For example, Google reports to use the Perron vector of the matrix $G(\alpha)$ corresponding to the specific value of $\alpha = 0.85$ as the PageRank vector [2,7,4].

Since the computational matrix $G(\alpha)$ is positive row-stochastic for any $\alpha \in [0, 1)$, it is a primitive matrix and, therefore, iterative algorithms (for example, the power method) converge to the only Perron vector $\pi(\alpha)$ of $G(\alpha)$. Furthermore, the numerical experiments show that a sequence of computed $\pi(\alpha)$ converges to some (nonnegative, normalized) vector as α approaches 1. These results were obtained in a number of studies focused on the development of efficient iterative solvers for the PageRank problem [8–12], and also confirmed in important theoretical contributions, which contain results about the structure of the Google matrix and convergence of the PageRank algorithms [5,13,2,7,14]. In particular, in [5] an explicit expression for the stationary distribution vector is obtained in the case $v = \frac{1}{N}e$ by applying singular perturbation theory, and an overestimation of the isolated subnets (the so-called "dead-ends") is revealed in the case of the large values of the parameter α.

In this study, using a simple approach involving only basic concepts of matrix theory and Markov chains, we derive expressions for the limiting transition matrix and the distribution vector of the PageRank approximation in the case of arbitrary personalization vector v, and discuss some implications of the obtained results. The text is structured as follows. First, in Section 2, we show that the uniform convergence is not generally observed in the above PageRank approximation involving two parameters: n (the number of iterations for specific $G(\alpha)$ with $\alpha < 1$) and α. It means that some nice consequences of uniform convergence (such as the possibility to interchange the order of limits) cannot be applied. In Section 3, we provide a simple proof of the convergence of the PageRank approximation by direct calculation, and establish a relation between the limiting matrix and the matrix S. Finally, in Section 4, we derive an expression for the limiting form of the PageRank vector, and comment on some properties of the PageRank algorithm, which follow from this form.

For the sake of simplicity, from now on we will use the term stochastic to refer to the row-stochastic matrices. We will also omit an indication of variable α in the computational matrices and their characteristics, except for Section 2 where this indication is essential. Throughout the text, we use standard terminology and different results from matrix analysis and theory of Markov chains. All used classical definitions and statements can be found in [15,6].

2. Absence of uniform convergence

Let us show the absence of uniform convergence for a simple example of the second order matrix

$$S = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

(Hereinafter I is the identity matrix of the appropriate order.) In this case, the expression for $G^n(\alpha)$ is the following:

$$G^n(\alpha) = \alpha^n I + (1-\alpha)V(1 + \alpha + \cdots + \alpha^{n-1})I = V + \alpha^n(I - V). \tag{1}$$

We will show that for such a sequence $G^n(\alpha)$ the Cauchy criterion for uniform convergence is not satisfied. In fact, evaluating the difference

$$G^{n+p}(\alpha) - G^n(\alpha) = \alpha^n(\alpha^p - 1)(I - V),$$

choosing $\alpha_n = \frac{n}{n+1}$ (in the left-hand side, the subscript n is the index of the elements of the chosen sequence), $p = n$, and noting that $V - I$ is a non-zero matrix, we obtain

$$G^{2n}(\alpha_n) - G^n(\alpha_n) = \frac{1}{(1+1/n)^n}\left(\frac{1}{(1+1/n)^n} - 1\right)(I - V). \tag{2}$$

Passing to the limit in the last equation, we have

$$\lim_{n\to\infty}(G^{2n}(\alpha_n) - G^n(\alpha_n)) = \frac{e-1}{e^2}(V - I) \neq 0, \tag{3}$$

where e is the base of the natural logarithm (unlike other sections where e denotes the vector of ones). The final result means that the sequence $G^n(\alpha)$ is not uniformly convergent.

It is easy to see that the absence of uniform convergence for $G^n(\alpha)$ is not related to the limiting properties of the matrix S itself, that is, to the information if $\lim_{n\to\infty} S^n$ exists or not. In fact, if we choose the non-convergent stochastic matrix

$$S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

then we can show that $G^n(\alpha)$ is not uniformly convergent by applying similar reasoning. First, noting that $S^{2n} = I$ and $S^{2n+1} = S$, we obtain

$$G^{2n}(\alpha) = \alpha^{2n} I + \frac{1-\alpha^{2n}}{1+\alpha} V(I+\alpha S) \tag{4}$$

and

$$G^{2n+2p}(\alpha) - G^{2n}(\alpha) = \alpha^{2n} \frac{1-\alpha^{2p}}{1+\alpha} [V(I+\alpha S) - (1+\alpha)I].$$

Then choosing $\alpha_n = \frac{n}{n+1}$ and $p = n$, we obtain

$$\lim_{n\to\infty} (G^{4n}(\alpha_n) - G^{2n}(\alpha_n)) = \frac{e^2-1}{2e^4} [V(I+S) - 2I]. \tag{5}$$

Finally, checking that $V(I+S) - 2I = S - I \neq 0$, we conclude that $G^n(\alpha)$ is not uniformly convergent for the chosen non-convergent S.

It can be shown that these considerations, which prove the absence of uniform convergence, can be generalized to the case of matrices S of arbitrary order. In fact, if one chooses the N-order identity matrix $S = I$, then all formulas, which lead to (3), are still valid. On the other hand, if one chooses an arbitrary symmetric permutation matrix S, then the relevant property $S^2 = I$ used for deriving (4) is satisfied, and therefore, all the transformations resulting in (5) are valid. Noting that the matrix $V(I+S)$ contains non-zero entries outside the main diagonal, we conclude that $V(I+S) - 2I \neq 0$ and uniform convergence does not take place.

3. Convergence of the PageRank approximation

It follows from matrix theory that a positive stochastic matrix G has the unique Perron vector π associated with the only eigenvalue $\lambda = 1$ located on the boundary of the spectral disk. In this case, the sequence G^n converges and

$$\lim_{n\to\infty} G^n = e\pi^T. \tag{6}$$

According to the theory of iterative methods, the power method, which provides the largest eigenpair (the largest in absolute value eigenvalue and the associated eigenvector), converges for such a matrix, and for any initial guess p (initial distribution vector) the limiting vector can be expressed as

$$\lim_{n\to\infty} p_n = \lim_{n\to\infty} p^T G^n = p^T \lim_{n\to\infty} G^n = p^T e\pi^T = \pi^T. \tag{7}$$

It means that for such G, convergence of the power method is equivalent to convergence of the sequence of powers of G. (Different implementations of the power method are the most used algorithms for PageRank computations, so we refer to this method in order to specify considerations.)

Let us make the direct calculation of $\lim_{n\to\infty} G^n$ when $\alpha < 1$. Using definition of G

$$G = \alpha S + (1-\alpha)V, \tag{8}$$

we readily obtain

$$\begin{aligned} G^n &= (\alpha S)^n + (1-\alpha)V(I-(\alpha S)^n)(I-\alpha S)^{-1} \\ &= (\alpha S)^n + V(I-(\alpha S)^n)C_\alpha, \qquad C_\alpha = (1-\alpha)(I-\alpha S)^{-1}, \end{aligned} \tag{9}$$

where I is the identity matrix (of course, $(I-\alpha S)^{-1}$ exists because all eigenvalues of αS are inside the unit disk). Taking the limit as $n \to \infty$ in the last formula, one obtains

$$B_\alpha = \lim_{n\to\infty} G^n = (1-\alpha)V(I-\alpha S)^{-1} = VC_\alpha = e\pi^T. \tag{10}$$

This is the expression for the limiting transition matrix (of the Markov chain with the transition matrix G) given in terms of the primitive parameters of the PageRank problem: matrices S and V, and parameter α. In order to show that the $\lim_{\alpha\to 1} B_\alpha$ (that is, the PageRank vector) exists, we should establish the existence of the $\lim_{\alpha\to 1} C_\alpha$. To do this, we consider two cases.

Case 1. Let us suppose that the matrix S is irreducible. To specify the approach, consider first the case when S is semisimple (that is, diagonalizable). Then, assuming $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is the matrix of the eigenvalues placed in the decreasing order of the absolute value, and R is the matrix whose rows are the respective left-hand eigenvectors of S, we obtain

$$\begin{aligned} C_\alpha &= (1-\alpha)(I-\alpha(R^{-1}\Lambda R))^{-1} = (1-\alpha)(R^{-1}(I-\alpha\Lambda)R)^{-1} \\ &= (1-\alpha)R^{-1}(I-\alpha\Lambda)^{-1}R = (1-\alpha)R^{-1}\mathrm{diag}\left(\frac{1}{1-\alpha}, \frac{1}{1-\alpha\lambda_2}, \ldots, \frac{1}{1-\alpha\lambda_N}\right)R \\ &= R^{-1}\mathrm{diag}\left(1, \frac{1-\alpha}{1-\alpha\lambda_2}, \ldots, \frac{1-\alpha}{1-\alpha\lambda_N}\right)R. \end{aligned} \tag{11}$$

Then

$$\lim_{\alpha\to 1} C_\alpha = R^{-1}\lim_{\alpha\to 1}\left(\mathrm{diag}\left(1, \frac{1-\alpha}{1-\alpha\lambda_2}, \ldots, \frac{1-\alpha}{1-\alpha\lambda_N}\right)\right)R = R^{-1}\mathrm{diag}(1, 0, \ldots, 0)R, \tag{12}$$

that is, the limit of C_α as α approaches 1 exists.

Now consider a general irreducible matrix S. Instead of the spectral decomposition of S we can use the Jordan canonical form J with the same order of the eigenvalues, and the respective eigenvectors and generalized eigenvectors placed in the rows of the matrix R:

$$\begin{aligned} C_\alpha &= (1-\alpha)(I-\alpha(R^{-1}JR))^{-1} = (1-\alpha)(R^{-1}(I-\alpha J)R)^{-1} \\ &= (1-\alpha)R^{-1}(I-\alpha J)^{-1}R, \end{aligned} \tag{13}$$

where $I-\alpha J$ is the block-diagonal matrix

$$I-\alpha J = \mathrm{diag}(I-\alpha J_1, I-\alpha J_2, \ldots, I-\alpha J_m) \tag{14}$$

and I is the identity matrix of the respective order. The inverse to $I-\alpha J$ has the form

$$(I-\alpha J)^{-1} = \mathrm{diag}((I-\alpha J_1)^{-1}, (I-\alpha J_2)^{-1}, \ldots, (I-\alpha J_m)^{-1}). \tag{15}$$

According to the assumption, the first block is of order 1:

$$I-\alpha J_1 = (1-\alpha).$$

Therefore,

$$(I-\alpha J_1)^{-1} = \frac{1}{1-\alpha}. \tag{16}$$

For any other separate block J_l of order 1 (corresponding to semisimple eigenvalue) we have:

$$(I-\alpha J_l)^{-1} = \frac{1}{1-\alpha\lambda_l}. \tag{17}$$

For any separate m-order Jordan block J_l, it is easy to check that the inverse to $I-\alpha J_l$ has the form

$$\begin{aligned} (I-\alpha J_l)^{-1} &= \frac{1}{(1-\alpha\lambda_l)^m} \\ &\times \begin{pmatrix} (1-\alpha\lambda_l)^{m-1} & -(1-\alpha\lambda_l)^{m-2} & \cdots & (-1)^{m-2}(1-\alpha\lambda_l) & (-1)^{m-1} \\ 0 & (1-\alpha\lambda_l)^{m-1} & \cdots & (-1)^{m-3}(1-\alpha\lambda_l)^2 & (-1)^{m-2}(1-\alpha\lambda_l) \\ & & \cdots & & \\ 0 & 0 & \cdots & (1-\alpha\lambda_l)^{m-1} & -(1-\alpha\lambda_l)^{m-2} \\ 0 & 0 & \cdots & 0 & (1-\alpha\lambda_l)^{m-1} \end{pmatrix}. \end{aligned} \tag{18}$$

Multiplying each of relations (16)–(18) by $1-\alpha$ and passing to the limit as α approaches 1, we obtain

$$\lim_{\alpha\to 1}(1-\alpha)(I-\alpha J_1)^{-1} = 1 \tag{19}$$

for the first block and

$$\lim_{\alpha\to 1}(1-\alpha)(I-\alpha J_l)^{-1} = 0 \tag{20}$$

for all other blocks. Finally, substituting the obtained limits for the individual Jordan blocks in (15), we obtain:

$$\lim_{\alpha\to 1} C_\alpha = R^{-1}\lim_{\alpha\to 1}(1-\alpha)(I-\alpha J)^{-1}R = R^{-1}\mathrm{diag}(1, 0, \ldots, 0)R \tag{21}$$

just like in (12).

Case 2. Let us consider a reducible matrix S. In this case the eigenvalue $\lambda_1 = 1$ is semisimple, which means that the Jordan canonical form of S has k Jordan blocks of order one related to $\lambda_1 = 1$. Nevertheless, general consideration keeps the same form as in the previous case. Indeed, the inverse to each of the first k blocks has the form (16) and the inverses to other blocks are found in the form (17) or (18). So, using the results (16)–(18), we obtain

$$\begin{aligned} C_\alpha &= (1-\alpha)(I-\alpha(R^{-1}JR))^{-1} = (1-\alpha)R^{-1}(I-\alpha J)^{-1}R \\ &= R^{-1}\mathrm{diag}(1,\ldots,1,(1-\alpha)(I-\alpha J_{k+1})^{-1},\ldots,(1-\alpha)(I-\alpha J_m)^{-1})R. \end{aligned} \tag{22}$$

Applying the results (19)–(20), we get

$$\lim_{\alpha\to 1} C_\alpha = R^{-1}\lim_{\alpha\to 1}(1-\alpha)(I-\alpha J)^{-1}R = R^{-1}\mathrm{diag}(1,\ldots,1,0,\ldots,0)R, \tag{23}$$

where the number of ones in the diagonal matrix is k.

Therefore, for an arbitrary stochastic matrix S, the limit $\lim_{\alpha\to 1} C_\alpha$ exists in the form:

$$\lim_{\alpha\to 1} C_\alpha = R^{-1}\mathrm{diag}(1,\ldots,1,0,\ldots,0)R. \tag{24}$$

It implies the existence of $\lim_{\alpha\to 1} B_\alpha$, which can be expressed in the form:

$$\lim_{\alpha\to 1} B_\alpha = VR^{-1}\mathrm{diag}(1,\ldots,1,0,\ldots,0)R = \tilde{B}, \tag{25}$$

where the number of ones in the diagonal matrix is equal to the dimension of the subspace for $\lambda_1 = 1$.

It is suitable to rewrite formula (10) in the form

$$B_\alpha(I-\alpha S) = (1-\alpha)V \tag{26}$$

and, taking the limit as α approaches 1 in the last relation, we obtain

$$\tilde{B}(I-S) = 0. \tag{27}$$

Therefore, the matrix $\tilde{B}$ satisfy the equation

$$\tilde{B}S = \tilde{B}. \tag{28}$$

Irreducible matrix S has the unique eigenvalue $\lambda_1 = 1$ and, therefore, the only Perron vector π. Since the matrix equation (28) represents N vector equations of the form $b_i^T S = b_i^T$, the unique nonnegative normalized solution for all these equations is the vector π, that is, the matrix $\tilde{B}$ assumes the form

$$\tilde{B}^T = (\pi^T,\ldots,\pi^T). \tag{29}$$

As it is well known [6], the Cesaro limit for an irreducible stochastic matrix is equal to this matrix $\tilde{B}$, that is

$$\tilde{B} = \lim_{n\to\infty}\frac{I+S+\cdots+S^{n-1}}{n}. \tag{30}$$

Note that for a primitive S the matrix $\tilde{B}$ coincides also with $\lim_{n\to\infty} S^n$.

Let us recall that any reducible stochastic matrix S can be transformed by symmetric permutations (that is, by reordering the web sites) to the (stochastic) canonical form:

$$P = \begin{pmatrix} T_{11} & 0 \\ T_{21} & T_{22} \end{pmatrix}, \tag{31}$$

where T_{11} and T_{22} are square blocks with spectral radius $\rho(T_{11}) = 1$ and $\rho(T_{22}) < 1$. Representing $\tilde{B}$ in a similar block form

$$\tilde{B} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \tag{32}$$

and substituting it in the equation $\tilde{B}S = \tilde{B}$, we obtain the following relations:

$$B_{11}T_{11} + B_{12}T_{21} = B_{11}, \qquad B_{21}T_{11} + B_{22}T_{21} = B_{21}, \qquad B_{12}T_{22} = B_{12}, \qquad B_{22}T_{22} = B_{22}. \tag{33}$$

Since $\rho(T_{22}) < 1$, the last two equations are satisfied only if $B_{12} = 0, B_{22} = 0$. Therefore, the first two equations are simplified to the form

$$B_{11}T_{11} = B_{11}, \qquad B_{21}T_{11} = B_{21}. \tag{34}$$

To specify the square block B_{11} we need more detailed information on the structure of the block T_{11}:

$$T_{11} = \begin{pmatrix} P_{11} & 0 & \cdots & 0 \\ 0 & P_{22} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & P_{kk} \end{pmatrix}. \tag{35}$$

Here $P_{jj}, j = 1, \ldots, k$ are irreducible stochastic matrices. Then each of P_{jj} has the unique Perron vector π_j and the block-diagonal matrix $\Pi = \mathrm{diag}(e\pi_1^T, e\pi_2^T, \ldots, e\pi_k^T)$ is the unique matrix satisfying the first equation in (34).

Finally, in order to specify the rectangular block B_{21} we note that due to (10) the following relation

$$(I - S)B_\alpha = (1 - \alpha)(I - S)V(I - \alpha S)^{-1} = (1 - \alpha)(V - V)(I - \alpha S)^{-1} = 0 \tag{36}$$

holds for any $\alpha \in [0, 1)$. Therefore, the limiting matrix $\tilde{B}$ satisfies the same relation, which can also be rewritten in terms of the canonical matrix P:

$$(I - P)\tilde{B} = 0. \tag{37}$$

Substituting the found blocks of $\tilde{B}$ in the last equation, we find the remaining block:

$$T_{21}\Pi + T_{22}C_{21} = C_{21}, \tag{38}$$

that is,

$$C_{21} = (I - T_{22})^{-1}T_{21}\Pi. \tag{39}$$

Thus, the matrix $\tilde{B}$ has the form

$$\tilde{B} = \begin{pmatrix} \Pi & 0 \\ (I - T_{22})^{-1}T_{21}\Pi & 0 \end{pmatrix} \tag{40}$$

with the square block Π specified above. Comparing this form with the form of the limit of the Cesaro sums for the stochastic matrix P (see, for example, [6]) we conclude that these two forms coincide.

Thus, for any stochastic matrix S the limiting transition matrix in the described PageRank approximation is equal to the matrix of the Cesaro limit for S.

4. Expression for the PageRank vector

Let us analyze formula (25) rewritten for the PageRank vector g computed starting from an initial guess p:

$$g = \lim_{\alpha \to 1}(\lim_{n \to \infty} p^T G^n) = \lim_{\alpha \to 1} p^T B_\alpha = p^T \tilde{B} = v^T R^{-1}\mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)R. \tag{41}$$

For an irreducible matrix S, this formula simplifies to

$$g = v^T R^{-1}\mathrm{diag}(1, 0, \ldots, 0)R = v^T(\bar{r}, 0, \ldots, 0)R, \tag{42}$$

where $\bar{r}$ is the first-column vector in the matrix R^{-1}. Recalling that (according to the used ordering of eigenvalues) the first-line vector in the matrix R is the only Perron vector π^T of the matrix S, we obtain

$$g = t\pi^T, \quad t = v^T\bar{r}. \tag{43}$$

Noting that (due to (10)) all vectors $p^T B_\alpha = p^T e\pi^T(\alpha) = \pi^T(\alpha)$ are located on the unit sphere, we finally conclude that the limiting vector g is on the unit sphere also, that is,

$$g = \pi^T. \tag{44}$$

Thus, for an irreducible matrix S (either primitive or not), the PageRank vector is the unique Perron vector π^T of S.

For a reducible stochastic matrix, recall again that it can be transformed to (stochastic) canonical form:

$$P = \begin{pmatrix} P_{11} & \cdots & 0 & 0 & \cdots & 0 \\ & & \cdots & & & \\ 0 & \cdots & P_{k,k} & 0 & \cdots & 0 \\ P_{k+1,1} & \cdots & P_{k+1,k} & P_{k+1,k+1} & \cdots & 0 \\ & & \cdots & & & \\ P_{m,1} & \cdots & P_{m,k} & P_{m,k+1} & \cdots & P_{m,m} \end{pmatrix}, \tag{45}$$

where the square blocks $P_{jj}, j = 1, \ldots, k$ are stochastic irreducible and the square blocks $P_{jj}, j = k+1, \ldots, m$ are substochastic irreducible. It means that each $P_{jj}, j = 1, \ldots, k$ has the unique right-hand and left-hand Perron vector (we use here a multiple of the right-hand Perron vector):

$$P_{jj}e_j = e_j, \qquad \pi_j^T P_{jj} = \pi_j^T, \quad j = 1, \ldots, k. \tag{46}$$

Therefore, the matrix P has exactly k right-hand and left-hand eigenvectors corresponding to the eigenvalue $\lambda_1 = 1$ given as follows:

$$\tilde{e}_1^T = (e_1^T, 0, \ldots, 0), \tilde{e}_2^T = (0, e_2^T, 0, \ldots, 0), \ldots, \tilde{e}_k^T = (0, \ldots, 0, e_k^T, 0, \ldots, 0); \tag{47}$$

$$\tilde{\pi}_1^T = (\pi_1^T, 0, \ldots, 0), \tilde{\pi}_2^T = (0, \pi_2^T, 0, \ldots, 0), \ldots, \tilde{\pi}_k^T = (0, \ldots, 0, \pi_k^T, 0, \ldots, 0). \tag{48}$$

Evidently,

$$\tilde{e}_j^T \tilde{\pi}_i = \delta_{ij}, \quad i, j = 1, \ldots, k, \tag{49}$$

where δ_{ij} is the Kronecker delta. Formula (41) was deduced under supposition that in the Jordan canonical form the order-one blocks corresponding to $\lambda_1 = 1$ are written first. Let us specify that these blocks are ordered according to the numbering of the eigenvectors $\tilde{\pi}_j, j = 1, \ldots, k$. Then, due to formula (49), the first k columns of the matrix R^{-1} are the vectors $\tilde{e}_j, j = 1, \ldots, k$ given in this exact order.

Applying this specification to (41), we obtain

$$\begin{aligned} g &= v^T R^{-1} \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0) R = v^T (\tilde{e}_1, \ldots, \tilde{e}_k, 0, \ldots, 0) R \\ &= (v^T \tilde{e}_1, \ldots, v^T \tilde{e}_k, 0, \ldots, 0)(\tilde{\pi}_1^T, \ldots, \tilde{\pi}_k^T, 0, \ldots, 0)^T \end{aligned} \tag{50}$$

or

$$g = t_1 \tilde{\pi}_1^T + \cdots + t_k \tilde{\pi}_k^T, \tag{51}$$

where each

$$t_j = v^T \tilde{e}_j, \quad j = 1, \ldots, k \tag{52}$$

is the sum of the components of v that correspond to the non-zero entries of the vector $\tilde{e}_j$ (the number of such entries is equal to the order of the block P_{jj}). As in the irreducible case, the limiting vector should lie on the unit sphere, that is,

$$\sum_{j=1}^{k} t_j = 1. \tag{53}$$

Evidently, formula (44) can be considered as a particular case of (51).

Thus, for any stochastic matrix S, personalization vector v and initial guess p the PageRank vector is given by formulas (51)–(53).

The following properties of the PageRank algorithm are immediate consequences of the obtained results.

Property 1. *Any transient class is ranked with* 0 *weight.*

It seems to be a bit strange, because adding a reference (may be erroneous) from any of sites (even the least important) of an ergodic class (isolated subnet) to a transient node will turn the whole ergodic class into the transient one, ranked with 0 . On the other hand, isolated subnets can have a strong connection with some pages in transient classes, however if they do not provide appropriate links it will result in a complete underestimation of the transient classes and accumulation of higher ranks by isolated subnets. In general, for the large values of α it leads to overrating the ergodic classes and underrating the transient classes, as it was indicated in [5].

Property 2. *A difference in ranking between ergodic classes is determined by the size (the number of sites) of each class and the weights of the personalization vector.*

Supposing that personalization vector is "neutral", that is, $v = \frac{1}{N}e$, the only mechanism to differentiate one ergodic class from another one is their sizes. It can lead to undesirable consequences. For example, if one creates his own closed (ergodic) class of sites, then this class can receive a reasonable rank based exclusively on the fact that this class has no outlinks, even if this subnet has no importance and no other class refers to it. In fact, let us consider an artificial closed (ergodic) subnet of two sites, say corresponding to the block

$$P_{11} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

with the left-hand Perron vector

$$\pi_1^T = \left(\frac{1}{2},\ \frac{1}{2}\right).$$

If $v = \frac{1}{N}e$, then the two nodes of this class will receive the rank

$$g_1^T = t_1\pi_1^T = \frac{1}{N}(1,\ 1).$$

If the left-hand Perron vector of a "valuable" (real) ergodic class, say corresponding to the block P_{22} of size n_2, is uniform

$$\pi_2^T = \left(\frac{1}{n_2},\ \ldots,\ \frac{1}{n_2}\right),$$

then all the sites of the last class will receive the same ranking:

$$g_2^T = t_2\pi_2^T = \frac{1}{N}\,(1,\ \ldots,\ 1)\,.$$

Therefore, the two artificial pages will be rated equally with the real "valuable" sites. If the vector π_2^T is not uniform, then the artificial nodes still will be ranked as an "average" site of a real ergodic class. It happens because a great weight received by a large ergodic class is divided (proportionally to its Perron vector) among all its sites.

The artificial nodes can gain even higher rank by creating a subnet oriented for these specific nodes. To specify considerations let us consider an example where the block P_{11} in (45) corresponds to such artificially created ergodic class of size n_1. Let us suppose that the first node in this group of sites is chosen to gain an artificially high rank. Then the subnet with the following link matrix solves this task:

$$P_{11} = \begin{pmatrix} 0 & 1/q & \cdots & 1/q \\ 1 & 0 & & 0 \\ & \cdots & & \\ 1 & 0 & \cdots & 0 \end{pmatrix},$$

where $q = n_1 - 1$. Indeed, the left-hand Perron vector of this matrix is

$$\pi_1^T = \frac{1}{2}\left(1,\ \frac{1}{q},\ \ldots,\ \frac{1}{q}\right)$$

and (for $v = \frac{1}{N}e$) the first n_1 components of the PageRank vector g are:

$$g_1^T = t_1\pi_1^T = \frac{n_1}{N}\left(\frac{1}{2},\ \frac{1}{2q},\ \ldots,\ \frac{1}{2q}\right).$$

By increasing the number of sites n_1 in such a subnet, one can achieve high ranking for the first node, albeit all this subnet is an artificial creation without any real significance.

Property 3. *The personalization vector influences strongly on the relative weights of different ergodic classes, but it does not influence on the relative weights of the nodes within each ergodic class.*

Within each ergodic class the relative ranking of the sites is defined only by the limiting probability distribution for this separate class. It seems to be a desirable property showing independence from subjective parameters like the personalization vector. In particular, an attempt to increase artificially the rank of some nodes within an ergodic class by using artificial sites created exclusively to point to the chosen nodes (as in the above example) will not be successful. In order to change the distribution of ranks within a chosen ergodic class, these artificial nodes should receive links from the "valuable" sites (sites of the ergodic part) of this class. Of course, some artificial manipulations with distribution of the rank within an ergodic class are still possible, but their existence is not caused by the use of the personalization vector.

Acknowledgements

We are grateful to Dr. Maxim Naumov of NVIDIA Corporation for stimulating discussions, and to two anonymous reviewers for useful comments on an earlier version of this paper.

References

[1] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1998) 107–117.
[2] N. Eiron, K.S. McGurley, J.A. Tomlin, Ranking the web frontier, in: Proc. Thirteenth International World Wide Web Conference, ACM Press, 2003, pp. 261–270.
[3] A.N. Langville, C.D. Meyer, A survey of eigenvector methods of web information retrieval, SIAM Rev. 47 (2005) 135–161.
[4] A.N. Langville, C.D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, Princeton, 2006.
[5] K. Avrachenkov, N. Litvak, K.S. Pham, A singular perturbation approach for choosing the PageRank damping factor, Internet Math. 5 (2008) 47–69.
[6] C.D. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, Philadelphia, 2000.
[7] T.H. Haveliwala, S.D. Kamvar, The Second eigenvalue of the Google matrix, Technical Report 2003–20, Stanford University, Stanford, 2003.
[8] C. Brezinski, M. Redivo-Zaglia, Rational extrapolation for the PageRank vector, Math. Comput. 77 (2008) 1585–1598.
[9] G.M. DelCorso, A. Gulli, F. Romani, Comparison of Krylov subspace methods on the PageRank problem, J. Comput. Appl. Math. 210 (2007) 159–166.
[10] G.H. Golub, C. Greif, An Arnoldi-type algorithm for computing Page Rank, BIT Numer. Math. 46 (2006) 759–771.
[11] S.D. Kamvar, T.H. Haveliwala, G.H. Golub, Adaptive methods for the computation of PageRank, Linear Algebra Appl. 386 (2004) 51–65.
[12] C.P.-C. Lee, G.H. Golub, S.A. Zenios, A fast two-stage algorithm for computing PageRank, Technical Report SCCM03-15, Stanford University, Stanford, 2003.
[13] P. Boldi, V. Lonati, M. Santini, S. Vigna, Graph fibrations, graph isomorphism, and PageRank, Theor. Inform. Appl. 40 (2006) 227–253.
[14] S. Serra-Capizzano, Jordan canonical form of the Google matrix: a potential contribution to the PageRank computation, SIAM. J. Matrix Anal. Appl. 27 (2005) 305–312.
[15] A. Berman, R.J. Plemmons, Nonnegative Matrices in the Mathematical Sciences, SIAM, Philadelphia, 1994.