INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Decomposition of the Google PageRank and Optimal Linking Strategy

Konstantin AVRACHENKOV — Nelly LITVAK

N° 5101

27 Janvier 2004

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--5101--FR+ENG

Decomposition of the Google PageRank and Optimal Linking Strategy

Konstantin AVRACHENKOV* , Nelly LITVAK†

Thème 1 — Réseaux et systèmes
Projet MAESTRO

Rapport de recherche n° 5101 — 27 Janvier 2004 — 15 pages

Abstract: We provide the analysis of the Google PageRank from the perspective of the Markov Chain Theory. First we study the Google PageRank for a Web that can be decomposed into several connected components which do not have any links to each other. We show that in order to determine the Google PageRank for a completely decomposable Web, it is sufficient to compute a subPageRank for each of the connected components separately. Then, we study incentives for the Web users to form connected components. In particular, we show that there exists an optimal linking strategy that benefits a user with links inside its Web community and in contrast inappropriate links penalize the Web users and their Web communities.

Key-words: Google, PageRank, Decomposition, Optimal Linking, Perturbed Markov Chains

* INRIA Sophia Antipolis, 2004, Route des Lucioles, B.P.93, 06902, France, e-mail: k.avrachenkov@sophia.inria.fr

† University of Twente, Dep. Appl. Math., Faculty of EEMCS, P.O. Box 217, 7500 AE Enschede, The Netherlands, e-mail: n.litvak@math.utwente.nl

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Décomposition de PageRank de Google et Stratégie Optimale de Création de Liens

Résumé : Nous analysons le mécanisme Google PageRank du point de vue de la théorie des chaînes de Markov. Nous étudions d'abord Google PageRank pour un Web pouvant être décomposé en plusieurs composantes connexes qui n'ont pas de liens entre elles. Nous montrons que pour déterminer le PageRank dans un Web entièrement décomposable, il suffit de calculer séparément un sous-PageRank pour chacune des composantes connexes. Puis nous étudions l'intérêt pour les utilisateurs de former des composantes connexes. En particulier, nous montrons qu'il existe une stratégie optimale d'établissement de liens qui profite aux usagers ayant des liens au sein de leur communauté, et qu'en revanche des liens inappropriés pénalisent les utilisateurs du Web et leurs communautés.

Mots-clés : Google, PageRank, Décomposition, Stratégie Optimale d'Établissement de Liens, Chaînes de Markov Perturbées

1 Introduction

Surfers on the Internet often use search engines to find pages satisfying their query. However, there are usually hundreds of relevant pages available, so listing them in a proper order is a crucial and non-trivial task. The original idea of Google presented in 1998 by Brin et al [4] is to list pages according to their PageRank which reflects popularity of a page. The PageRank is defined in the following way. Denote by n the total number of pages on the web and define the $n \times n$ hyperlink matrix P as follows. Suppose that page i has $k > 0$ outgoing links. Then $P_{ij} = 1/k$ if j is one of the outgoing links and $P_{ij} = 0$ otherwise. If a page does not have outgoing links, there are several logical solutions: either to make this state absorbing, or to introduce the effect of "back button" [5] or to spread the probability among some subset of the Web community. We shall discuss these options later in the paper.

In order to make the hyperlink graph connected, it is assumed that a random surfer goes with some probability to an arbitrary web page with the uniform distribution. Thus, the PageRank is defined as a stationary distribution of a Markov chain whose state space is the set of all web pages, and the transition matrix is

$$\tilde{P} = cP + (1-c)(1/n)E, \tag{1}$$

where E is a matrix whose all entries equal one, n is the number of web pages, and $c \in (0,1)$ is the probability of not jumping to a random page, and is chosen to be 0.85. The Google matrix $\tilde{P}$ is stochastic, aperiodic, and irreducible, so there exists a unique row vector π such that

$$\pi\tilde{P} = \pi, \quad \pi\underline{1} = 1, \tag{2}$$

where $\underline{1}$ is a column vector of ones. The vector π satisfying (2) is called a PageRank vector, or simply PageRank. If a surfer follows a hyperlink with probability c and jumps to a random page with probability $(1-c)$, then π_i can be interpreted as a stationary probability that such a surfer is at page i.

The PageRank can be regarded from two distinct points of view. From a user point of view, an essential question is if it is possible to improve the user's PageRank by an appropriate linking to the other pages. We show that an optimal linking strategy does exist.

From the Google point of view, an essential question is how to organize efficiently the computation of the Google PageRank. Since the number of web pages is huge, updating the PageRank involves very high computational costs. In order to keep up with constant modifications of the web structure, Google updates its PageRank once a month. Thus, it is very important to be able to compute the vector π at minimal possible cost. One of the ways to do it is to explore specific properties of the hyperlink matrix P. For example, by its very nature, the matrix P is reducible [11]. Google overcomes this problem using the uniform perturbation (1). We demonstrate that one can take a great advantage of the reducibility of the web. In this paper we study the case when the web can be decomposed into several connected components that do not communicate with each other. This setting creates a natural opportunity for the application of the parallel processing.

The contribution of the paper is threefold: In Section 2, we show how one can take advantage of the web reducibility. In Sections 3 and 4 we discuss incentives for the Web users to form connected components. In particular, in Section 3, we study an optimal linking strategy of an individual user; and in Section 4, we demonstrate how inappropriate links penalize the Web users and their Web communities.

2 Decomposition of the PageRank

Let us show how one can take an advantage of the Web reducibility. Towards this goal, we consider a hyperlink matrix P of the form

$$P = \begin{bmatrix} P_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & P_N \end{bmatrix}, \tag{3}$$

where the elements of diagonal blocks P_I, $I = 1, ..., N$, correspond to links inside the Ith group of pages. Denote by n_I the number of pages in group I. A group I does not communicate with the outside world but it might have itself a complex ergodic structure. In particular, each block itself might have connected components which do not communicate with each other. This can be useful, as one can consider only the number of blocks that correspond to the number of available parallel processors. Some connected components of the Web can be too tiny to be considered on their own and might be grouped in a larger block, for instance, according to their contents.

Next we consider the Google matrix corresponding to the hyperlink matrix (3) $\tilde{P} = cP + (1-c)(1/n)E$, and let vector π be the PageRank of $\tilde{P}$ such that $\pi\tilde{P} = \pi$ and $\pi\underline{1} = 1$. Furthermore, for block I, define the perturbed matrix

$$\tilde{P}_I = cP_I + (1-c)(1/n_I)E, \quad I = 1, \ldots, N, \tag{4}$$

and let vector π_I be the PageRank of $\tilde{P}_I$ such that

$$\pi_I\tilde{P}_I = \pi_I, \quad \pi_I\underline{1} = 1.$$

Then the following theorem holds.

Theorem 1 *The PageRank π is given by*

$$\pi = ((n_1/n)\pi_1, (n_2/n)\pi_2, \ldots, (n_N/n)\pi_N). \tag{5}$$

Proof. Let us verify that (5) is indeed a stationary distribution of $\tilde{P}$. Define

$$\bar{E} = \begin{bmatrix} (1/n_1)E & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (1/n_N)E \end{bmatrix}.$$

Then for π given by (5), we have

$$\begin{aligned}
\pi\tilde{P} &= \pi[cP + (1-c)\bar{E} - (1-c)\bar{E} + (1-c)(1/n)E] \\
&= \pi[cP + (1-c)\bar{E}] + \pi[(1-c)(1/n)E - (1-c)\bar{E}] \\
&= ((n_1/n)\pi_1\tilde{P}_1, \ldots, (n_N/n)\pi_N\tilde{P}_N) + (1-c)(1/n)\underline{1}^T - (1-c)(1/n)\underline{1}^T \\
&= ((n_1/n)\pi_1, \ldots, (n_N/n)\pi_N) = \bar{\pi}.
\end{aligned}$$

Now the statement of the theorem follows since there is a unique stationary vector of $\tilde{P}$. □

As we see, the proof of Theorem 3 is quite easy. In fact, it is sufficient to guess that such result should hold. We noticed this decomposition property when applying perturbation analysis [15] to the matrix $\tilde{P}$. This analysis along with other results will be presented in our forthcoming paper [1]. Note that the decomposition property of the PageRank can also be deduced from the known formula [3, 11, 12]

$$\pi = \frac{1-c}{n}\underline{1}^T[1 - cP]^{-1}. \tag{6}$$

Moreover, it is easy to see from the above formula that the Topic-Sensitive PageRank [7] can be decomposed as well.

Let us now discuss the implications of the result of Theorem 1 to the computation of the Google PageRank. If one does not know the block structure of the hyperlink matrix, then one can determine the connected components of the web using any graph-traversal method such as Depth-First Search method or Breadth-First Search method (see, e.g., [14]). The complexity of the graph-traversal algorithms is $O(n + l)$, where n is the number of web pages and l is the number of links. Note that because of the sparsity of the hyperlink matrix, the complexity is close to linear in the number of pages. Furthermore, we note that the complexity of one power iteration is comparable with the complexity of a graph-traversal algorithm. Since Google performs the number of power iterations of the order 100, it is justified to spend around 1% of the computational time to determine the connected components of the Web graph and then using the result of Theorem 1 to take an advantage of the complete decomposability of the problem. In particular, each connected component of the Web graph can be processed thoroughly in an independent manner. This way we save not only the computation time but also the memory resources, as we can store different parts of the PageRank approximation vector in different data bases. We would like to note that graph-traversal algorithms can also be efficiently implemented on parallel processors [10]. In particular, the theoretical cost-optimal number of processors can be estimated as $n/log(n)$ [10]. Since the number of pages n is huge, as many parallel processors as possible should be used for the graph-traversal algorithm. Furthermore, the graph-traversal algorithm for the detection of connected components can be combined with filtering of "link farms" [7]. Once a "link farm" component is detected, it can be altogether excluded from the PageRank computation. Finally, we would like to note that even better computational efficiency can be achieved by combining on-line the graph-traversal algorithms with the Web crawling by Google robots.

Now let us discuss what one should do with the pages without outgoing links. From Theorem 1 it is clear that if one wants to take a full advantage of the Web reducibility, then one should assign transition probabilities either using the "back button" idea [5] or spreading the probability only among the states of a block to which a page with no outgoing links belongs. The latter suggestion has a nice interpretation if connected components are grouped in blocks according to their contents. This idea represents the case when a surfer jumps not to a completely random page but to a random page within a given context. For instance, he/she might do this by using different search engines.

3 Optimal linking strategy

In this section we discuss how a new link from page i to page j affects the ranking of these two pages and why it is important to have relevant outgoing links in order to be ranked higher.

The results of this section are based on the following well-known result from the theory of Markov chains. Let π be a stationary vector of an aperiodic irreducible Markov chain $\{X_k, k \geq 0\}$ with set of states $\{1, 2, \ldots, n\}$. Then

$$\pi_i = 1/\mu_i, \quad i = 1, \ldots, n, \tag{7}$$

where μ_i is the average number of transitions between two successive visits to state i.

In Lemma 2 below we prove an intuitive statement that a new link from page i to page j would increase a PageRank of j. This lemma might be already known in the theory of Markov chains but we shall prove it here in the context of PageRank for the sake of completeness.

Lemma 2 *For all $i, j = 1, \ldots, n$; $i \neq j$, if a page i adds a new link to page j then the PageRank of j increases.*

Proof. For a Markov chain $\{X_k, k \geq 0\}$ with state space $\{1, 2, \ldots, n\}$ and transition matrix $\tilde{P}$, and for any $m > 0$, define the taboo probabilities

$$\begin{aligned}
f_{jj}^{(m)} &= \mathbb{P}(X_m = j, X_k \neq i, X_k \neq j, 1 \leq k < m \,|X_0 = j), \\
f_{ji}^{(m)} &= \mathbb{P}(X_m = i, X_k \neq i, X_k \neq j, 1 \leq k < m \,|X_0 = j), \\
g_{ij}^{(m)} &= \mathbb{P}(X_m = j, X_k \neq j, 1 \leq k < m \,|X_0 = i).
\end{aligned}$$

The value $f_{ji}^{(m)}$, for example, is the probability that starting from state j the process reaches state i for the first time after m steps without visiting state j. Further, $g_{ij}^{(m)}$ is the probability that starting from i the process will reach j for the first time after m steps.

The value μ_j corresponding to the original matrix $\tilde{P}$ (without the new link $i \to j$) can be written as a sum of three terms:

$$\mu_j \quad = \quad \sum_{m=1}^{\infty} m f_{jj}^{(m)} + \sum_{m=1}^{\infty} m f_{ji}^{(m)} + \sum_{m=1}^{\infty} m g_{ij}^{(m)}. \tag{8}$$

The first term in the right-hand side of (8) is the contribution of the paths that start from j and come back to j without passing through i. The second term sums over the paths which start from j and finish when reaching i for the first time. Finally, the third term corresponds to the paths which lead from i back to j.

Now suppose that the new link from i to j is added. With the new matrix, the first term in the right-hand side of (8) remains unaltered because this term is the contribution of the paths which do no pass through i and thus they will not be affected by the link update. The same is true for the second term. As for the third term, it will obviously decrease. Indeed, the length of the paths which are not using the new link remains the same but the probability of such paths reduces because some of the 'weight' is given to the new paths which are obviously shorter. Thus, the new link leads to reducing of μ_j. Then (7) implies immediately that the new rank of j is greater than the old one. □

Corollary 3 *For any page $i = 1, 2, \ldots, n$, it holds*

$$\min_{1 \le i \le n} \{\pi_i\} \ge (1-c)/n.$$

Proof. Indeed, a page i has the minimal rank if no other page refers to it. In this case we have

$$\pi_i = [\pi \tilde{P}]_i = (1-c)(1/n)\pi e = (1-c)(1/n),$$

which proves the statement of the corollary. □

Let us now consider how outgoing links from page i influence its PageRank. Assume that the links from i to i itself are not allowed (or rather that Google does not count them as links). Then we can easily define an optimal linking strategy for any site i. Define the average time needed to come from j to i by

$$\mu_{ji} = \sum_{m=1}^{\infty} m g_{ji}^{(m)}.$$

Consider some page $i \in \{1, \ldots, n\}$ and assume that i has links to the pages $i_1, \ldots, i_k$ where $i_l \neq i$ for all $l = 1, \ldots, k$. Then for the Google matrix $\tilde{P}$, we have

$$\mu_i = 1 + \frac{c}{k} \sum_{l=1}^{k} \mu_{i_l i} + \frac{1}{n}(1-c) \sum_{\substack{j=1 \\ j \neq i}}^{n} \mu_{ji}, \tag{9}$$

where $c = 0.85$ is the Google constant [4]. The objective now is to choose k and $i_1, \ldots, i_k$ such that μ_i becomes as small as possible. From (9) one can see that μ_i is a linear function of μ_{ji}'s. Moreover, outgoing links from i do not affect μ_{ji}'s. Thus, linking from i to j, one can only alter the coefficients in the right-hand side of (9). It means that the owner of the page i has a very little control over its PageRank. The best what he/she can do is to link only to one site j^* such that

$$\mu_{j^* i} = \min_j \{\mu_{ji}\}.$$

Note that (surprisingly) the PageRank of j^* plays no role here.

Of course, in reality, it is not practical to link only to one page if we want to have a really high ranking because incoming links are much more important than outgoing ones. Roughly speaking, when we link to other people, we offer them shorter paths back to their pages. Then it becomes more likely that they will link to us as well, and this is exactly what will make our PageRank higher. However, we should link only to relevant pages, i.e., to pages which (potentially) have a short average path back to our page. Inappropriate links would only damage our PageRank. In the next section we provide a more enlightening discussion on the effect of inappropriate links.

Interestingly, the discussion on optimal linking strategy partially explains a "practical" advise according to which, a Web site owner should view his/her site as a set of pages and maintain a good inter-link structure and to refer to his/her colleagues [13]. Indeed, it follows from our argument that such a policy will certainly increase the PageRank of all pages in a group.

The major goal of Google is to rank good and useful pages high. Naturally, one would expect that a good web site has convenient inner links and many relevant outgoing links. Now we saw that these qualities would indeed lead to a higher ranking.

4 The effect of inappropriate links

Here we study yet another incentive for the Web users to refer only to relevant pages. Namely, we study the consequences of inappropriate links. We define inappropriate links as links that point from one Web community to another and these two Web communities have different contents. One can expect that the number of such links is much less than the number of links connecting members of the same community and that often inappropriate links point from one community towards another but not vice versa. To model such a situation, we consider two Web communities, say Community 1 and Community 2 with some links going from Community 2 to Community 1 but no link going from Community 1 to Community 2. Without inappropriate links the hyperlink matrix would be completely decomposable

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}, \tag{10}$$

where P_1 and P_2 are hyperlink matrices of Community 1 and Community 2, respectively. However, with inappropriate links the hyperlink matrix has the following form

$$\hat{P} = \begin{bmatrix} P_1 & 0 \\ R & \hat{P}_2 \end{bmatrix}, \tag{11}$$

where the matrix R represents the links from Community 2 to Community 1. Note that P_1 and P_2 are stochastic matrices and $\hat{P}_2$ is a substochastic matrix. Below be study two cases: a general case, when any page of Community 2 can have inappropriate links, and a special case, when only one of the pages of Community 2 has inappropriate links.

4.1 General case: several pages with inappropriate links

Theorem 4 *Let* $[\pi_1\ \pi_2]$ *be a PageRank vector of the Google matrix* $\tilde{P} = cP + (1-c)\frac{1}{n}E$, *where* P *is given by* (10). *And let* $[\hat{\pi}_1\ \hat{\pi}_2]$ *be a PageRank vector of the Google matrix of the Web with inappropriate links, with the hyperlink matrix as in* (11). *Then,* $\hat{\pi}_1$ *and* $\hat{\pi}_2$ *can be expressed as follows:*

$$\hat{\pi}_2 = \frac{1-c}{n}\underline{1}^T[I - c\hat{P}_2]^{-1}. \tag{12}$$

$$\hat{\pi}_1 = \pi_1 + c\hat{\pi}_2 R[I - cP_1]^{-1}. \tag{13}$$

Proof. With the help of the decomposition idea, the proof follows immediately from the formula

$$[\hat{\pi}_1\ \hat{\pi}_2][I - c\hat{P}] = \frac{1-c}{n}\underline{1}^T.$$

written in the block form. □

Now let us find an expression for $\hat{\pi}_2$ in terms of π_2. Assume that m pages $i_1, i_2, \ldots, i_m$ of Community 2 have links to Community 1. Denote by U an $m \times n_2$ matrix that consists of m non-zero rows of $P_2 - \hat{P}_2$. Further, let $V^T = (e_{i_1}, e_{i_2}, \ldots, e_{i_m})$ be an $n_2 \times m$ matrix whose ν-th column e_{i_ν} equals the i_ν-th column of the identity matrix I. Then, we can write

$$P = \hat{P} + V^T U,$$

and apply the Woodbury formula [6]

$$\begin{aligned}[I - c\hat{P}_2]^{-1} &= [I - cP_2 + c(P_2 - \hat{P}_2)]^{-1} \\ &= [I - cP_2]^{-1} - [I - cP_2]^{-1}cV^T[I + cU[I - cP_2]^{-1}V^T]^{-1}U[I - cP_2]^{-1}.\end{aligned} \tag{14}$$

Now, substitution of (14) in (12) together with (6) yields

$$\hat{\pi}_2 = \pi_2\left\{I - cV^T[I + cU[I - cP_2]^{-1}V^T]^{-1}U[I - cP_2]^{-1}\right\}. \tag{15}$$

From formula (15) one can see that the pages that are "far away" from the pages $i_1, i_2, \ldots, i_m$ suffer less from the inappropriate linking. Indeed, for $i, j = 1, \ldots, n_2$, denote by s_{ij} a minimal number of hyperlink transitions that is needed in order to reach j from i, and put $s_{ii} = 0$. For $i \neq j$, the entry (i, j) of the matrix P_2^k equals zero for all $k = 1, 2, \ldots, s_{ij} - 1$. Now, since

$$[I - cP_2]^{-1} = \sum_{k=0}^{\infty}(cP_2)^k, \tag{16}$$

we see that

$$U[I - cP_2]^{-1}e_j = c^{s_j}UP_2^{s_j}[I - cP_2]^{-1}e_j,$$

where $s_j = \min_{1\leq\nu\leq m}\{s_{i_\nu j}\}$. Thus, the proportion of the PageRank lost by a page j is reduced at least by factor c with every extra step needed to reach j from $i_1, i_2, \ldots, i_m$. Naturally, if there is no hyperlink path from any of these pages to j (it can happen, for

example, if j is a transient state with respect to the transition matrix P_2) then the rank of page j does not change. In this case, all entries in the jth column of $U[I - cP_2]^{-1}$ equal zero.

Let π_2^j be the jth coordinate of π_2, $j = 1, 2, \ldots, n_2$. We conclude that for any page j, $\hat{\pi}_2^j < \pi_2^j$ if there is a hyperlink path from any of the pages $i_1, i_2, \ldots, i_m$ to j, and $\hat{\pi}_2^j = \pi_2^j$, otherwise. In other words, inappropriate linking by pages $i_1, i_2, \ldots, i_m$ results in a reduced ranking for all pages in Community 2 who have a hyperlink path from $i_1, i_2, \ldots, i_m$. Moreover, the closest pages will be affected the most.

4.2 Special case: one page with inappropriate links

One can get much more insight, by studying the case when only one of the pages of Community 2 has inappropriate links. Suppose that some page $i \in \{1, \ldots, n_2\}$ initially had k_2 links to the pages in Community 2 and then added k_1 links to Community 1. Assume further that other pages of Community 2 link only to each other. Denote $k = k_1 + k_2$. In this case, matrix U equals to a row vector u whose non-zero entries equal $1/k_2 - 1/k = k_1/(kk_2)$. One can see that u is merely the ith row of P_2 multiplied by k_1/k. Furthermore, V^T is now equal to the column vector e_i. In this case, formula (15) becomes

$$\hat{\pi}_2 = \pi_2 \left(I - \frac{ce_i u[I - cP_2]^{-1}}{1 + cu[I - cP_2]^{-1}e_i} \right) = \pi_2 - \frac{c\pi_2^i u[I - cP_2]^{-1}}{1 + cu[I - cP_2]^{-1}e_i}, \tag{17}$$

where π_2^i is the ith coordinate of π_2.

From (17), one can see how much probability mass the Community 2 loses when one page gives inappropriate links. Multiplying both parts of (17) by the column vector $\underline{1}$ and taking into account that

$$[I - cP_2]^{-1}\underline{1} = \sum_{l=0}^{\infty} (cP_2)^l \underline{1} = [1/(1-c)]\underline{1},$$

we get

$$(\pi_2\underline{1} - \hat{\pi}_2\underline{1}) = \frac{c\pi_2 e_i u[I - cP_2]^{-1}\underline{1}}{1 + cu[I - cP_2]^{-1}e_i} = \frac{ck_1\pi_2^i}{(1-c)k\,(1 + cu[I - cP_2]^{-1}e_i)}. \tag{18}$$

The concept of communities was also used in [2], where the authors interpret the probability mass as an energy, and they express the lost energy in terms of the new PageRank $\hat{\pi}_2$. In this paper, we determine the lost probability mass in terms of the old ranking π_2. Using these results, one can anticipate the consequence of the inappropriate linking. Besides, these results can be used in the PageRank update.

An interesting special case of formula (17) is when the page i has links to all pages in Community 2 (including the page i itself). In this case, we have

$$u = \frac{k_1}{kk_2}\underline{1}^T,$$

and due to (6), formula (17) becomes

$$\hat{\pi}_2 = \pi_2 \left(I - \frac{ck_1 e_i \underline{1}^T [I - cP_2]^{-1}}{kk_2 + ck_1 \underline{1}^T [I - cP_2]^{-1} e_i} \right) = \pi_2 \left(1 - \frac{ck_1 \pi_2^i}{kk_2 + ck_1 \pi_2^i} \right). \tag{19}$$

We see that in this case, the ranking lost by page j in Community 2 is proportional to π_2^j and is an increasing concave function of π_2^i.

Formula (17) has a clear probabilistic interpretation. According to the transition matrix $\tilde{P} = cP + (1-c)(1/n)E$, at each step, the hyperlink transitions governed by matrix P occur with probability c and the random transitions governed by matrix $(1/n)E$ occur with probability $(1-c)$. Now consider a sample path of hyperlink transitions in Community 2 until the first random transition occurs. Denote by z_{ij} the expected number of visits to page j on such a path started from page i when transitions are defined by the matrix cP_2. Note that z_{ij} is equal to the entry (i,j) of the matrix $[I - cP_2]^{-1}$, that is,

$$z_{ij} = e_i^T [I - cP_2]^{-1} e_j, \quad i,j = 1, \ldots, n_2.$$

Further, if the path starts in i then $cu[I - cP_2]^{-1} e_i$ can be interpreted as an expected number of visits to i that are lost as a consequence the inappropriate links. Hence, we can interpret formula (17) as follows:

$$\pi_2^j - \hat{\pi}_2^j = \frac{\pi_2^i \mathbb{E}(\#\text{lost visits from } i \text{ to } j)}{1 + \mathbb{E}(\#\text{lost visits from } i \text{ to } i)}, \tag{20}$$

for $j = 1, \ldots, n_2$. In the following theorem we determine the amount of ranking lost by the page i.

Theorem 5 *If some page $i \in \{1, 2, \ldots, n_2\}$ of Community 2 has k_1 links to Community 1 and k_2 links to Community 2, and all other pages of Community 2 link only to each other, then*

$$\pi_2^i - \hat{\pi}_2^i = \frac{\pi_2^i z_{ii} \{k_1 z_{ii} + k_2 - k\}}{k_1 z_{ii}^2 + k_2}. \tag{21}$$

Proof. We only need to determine the expected number of lost visits from i to i. Obviously,

$$\begin{aligned} \mathbb{E}(\#\text{lost visits from } i \text{ to } i) = \mathbb{E}(\#\text{visits from } i \text{ to } i \text{ with } cP_2) \\ - \mathbb{E}(\#\text{visits from } i \text{ to } i \text{ with } c\hat{P}_2). \end{aligned} \tag{22}$$

Now, let q_{ii} be a probability to make a random transition on the way from i to i with cP_2. Then, because of the Markov property,

$$[\#\text{visits from } i \text{ to } i \text{ with } cP_2] + 1$$

has a geometric distribution with parameter q_{ii}. For the same reason,

$$[\#\text{visits from } i \text{ to } i \text{ with } c\hat{P}_2] + 1$$

has a geometric distribution with some other parameter $\hat{q}_{ii}$. Since z_{ii} is the average number of visits to i starting from i with the transition matrix cP_2, we have

$$q_{ii} = z_{ii}^{-1}. \tag{23}$$

On the other hand,

$$1 - q_{ii} = c\mathbb{P}(\text{reaching } i \text{ from } i \text{ with } cP_2 | \text{ the first transition follows } cP_2). \tag{24}$$

Denoting the last conditional probability by γ_{ii}, we get

$$\gamma_{ii} = (1 - z_{ii}^{-1})/c.$$

Furthermore, in a similar fashion,

$$\begin{aligned} 1 - \hat{q}_{ii} &= c\left(1 - \frac{k_1}{k}\right)\mathbb{P}(\text{reaching } i \text{ from } i \text{ with } c\hat{P}_2 | \text{the 1st trans. with } c\hat{P}_2) \\ &= c\left(1 - \frac{k_1}{k}\right)\gamma_{ii} = \frac{k_2}{k}(1 - z_{ii}^{-1}), \end{aligned} \tag{25}$$

where the second equality follows since $\hat{P}_2$ differs from P_2 only by its ith row. Now, substituting (23) and (25) in (22), we get

$$cu[I - cP_2]^{-1}e_i = \mathbb{E}(\#\text{lost visits from } i \text{ to } i) = \frac{1}{q_{ii}} - \frac{1}{\hat{q}_{ii}} = z_{ii}\left(1 - \frac{k}{k_1 z_{ii} + k_2}\right). \tag{26}$$

Substitution of (26) into (17) or (20) gives the desired result. □

Using the probabilistic approach, we could also tackle the problem of one page with inappropriate links in a different manner. Assume again that some page $i \in \{1, 2, \ldots, n_2\}$ which had k_2 links to Community 2 added k_1 links to Community 1. For page j, denote by μ_j and $\hat{\mu}_j$ the mean first passage time from j to j in the old and in the new situation, respectively. Then we have

$$\frac{\pi_2^j}{\hat{\pi}_2^j} = \frac{\hat{\mu}_2^j}{\mu_j} = 1 + \frac{\hat{\mu}_j - \mu_j}{\mu_j} = 1 + (\hat{\mu}_j - \mu_j)\pi_2^j,$$

and thus

$$\hat{\pi}_2^j = \frac{\pi_2^j}{1 + (\hat{\mu}_j - \mu_j)\pi_2^j}. \tag{27}$$

Note that $\hat{\mu}_j$ differs from μ_j only in contribution of the paths that are passing through i. For such paths, in the new situation, there is a probability ck_1/k to make a transition to Community 1 instead of Community 2. After such a transition, the only possibility to return back to j is to make a random transition to the Community 2. At each step, a probability of such random transition is $(1 - c)n_2/n$, and the transition to each of the pages

in Community 2 is equally likely. Hence, the average number of transitions needed to return to the state j from Community 1 equals

$$\frac{n}{(1-c)n_2} + \frac{1}{n_2}\sum_{l\neq j}\hat{m}_{lj},$$

where $\hat{m}_{lj}$ is the mean first passage time from l to j in the new situation. Let $\hat{m}_{\cdot j} = (1/n_2)\sum_{l\neq j}\hat{m}_{lj}$ be the average time needed to reach the page j starting from the uniform distribution over the Community 2. Now it follows from (27) that

$$\pi_2^j - \hat{\pi}_2^j = \left[\frac{\hat{\beta}_{ji}\left\{1+\frac{n}{(1-c)n_2}+(\hat{m}_{\cdot j}-m_{ij})\right\}\pi_2^j}{1+\hat{\beta}_{ji}\left\{1+\frac{n}{(1-c)n_2}+(\hat{m}_{\cdot j}-m_{ij})\right\}\pi_2^j}\right]\pi_2^j,$$

where m_{ij} is the mean first passage time from i to j in the old situation, and $\hat{\beta}_{ji}$ is a probability that the path from j to j goes via i and makes a transition by using the inappropriate link. The term

$$\hat{\beta}_{ji}\left(1+\frac{n}{(1-c)n_2}+(\hat{m}_{\cdot j}-m_{ij})\right)\pi_2^j \tag{28}$$

determines the proportion of rank lost by the page j. This term characterizes the change in the distance from page i to page j. Naturally, this term is larger for pages j that are 'close' to i. If the rank of i is high, then, in particular, it means that i is 'close' to other pages, and thus (28) should increase with π_2^i. For the same reason, the term (28) increases with π_2^j. However, the closeness to i plays more important role in (28) than the popularity of j. For instance, when i refers to everybody in Community 2 then the change in the distance (28) is also the same for everybody and therefore $\pi_2^j - \hat{\pi}_2^j$ is simply proportional to π_2^j as we saw in (19).

Finally, consider what happens if Community 2 is small and highly inter-connected whereas Community 1 is large (that can happen, for example, when a newly constructed Web site gives links to the outside world). Then,

$$1+\frac{n}{(1-c)n_2} >> (\hat{m}_{\cdot j}-m_{ij}),$$

and we obtain the following estimate

$$\pi_2^j - \hat{\pi}_2^j \approx \frac{\hat{\beta}_{ji}\left\{1+\frac{n}{(1-c)n_2}\right\}(\pi_2^j)^2}{1+\hat{\beta}_{ji}\left\{1+\frac{n}{(1-c)n_2}\right\}\pi_2^j} < \frac{c^{s_{ji}+1}(k_1/k)\alpha_{ji}\{(1-c)n_2+n\}(\pi_2^j)^2}{(1-c)n_2+c^{s_{ji}+1}(k_1/k)\alpha_{ji}\{(1-c)n_2+n\}\pi_2^j},$$

where s_{ji} is the length of the shortest path from j to i, and α_{ji} is the probability of the most likely path from j to i when the path is induced by the transition matrix P_2.

References

[1] Avrachenkov, K.E., and Litvak, N., PageRank as a Stationary Distribution of a Singularly Perturbed Markov Chain. In preparation.

[2] Bianchini, M., Gori, M., Scarselli, F., (2002) PageRank: A circuital analysis. In the Proceedings of the 11-th WWW Conference.

[3] Bianchini, M., Gori, M., Scarselli, F., (2002) Inside PageRank. ACM Trans. Internet Technology, In press.

[4] Brin, S., Page, L., Motwami, R. and Winograd, T. (1998) The PageRank citation ranking: bringing order to the web. Stanford University Technical Report.

[5] Fagin, R., Karlin, A.R., Kleinberg, J., Raghavan, P., Rajagopalan, Rubinfeld, R., Sudan, M., and Tomkins, A., (2000), Random walks with "back buttons", In the Proceedings of 32nd ACM Symposium on Theory of Computing.

[6] Golub, G. H. and Van Loan, C. F. (1996) Matrix Computations, 3rd ed. Baltimore, MD: Johns Hopkins.

[7] Haveliwala, T.H., (2002) Topic-Sensitive PageRank, In Proceedings of the Eleventh International World Wide Web Conference.

[8] Kamvar, S.D., Haveliwala, T.H., Manning, C.D. and Golub, G.H. (2003) Exploiting the Block Structure of the Web for Computing PageRank. Stanford University Technical Report.

[9] Kemeny, J.G. and Snell, J.L., Finite Markov Chains, The University Series in Undergraduate Mathematics, Van Nostrand, Princeton, NJ, 1960,

[10] Kumar, V., Grama, A., Gupta, A., and Karypis, G., (1994), Introduction to Parallel Computing: Design and Analysis of Algorithms, The Benjamin/Cummings Publishing Company, Inc.

[11] Langville, A.N., and Meyer, C.D. (2003) Deeper Inside PageRank. Preprint, North Carolina State University.

[12] Moler, C.D., and Moler, K.A., (2003) Numerical Computing with MATLAB, SIAM.

[13] http://www.searchenginewatch.com/

[14] Sedgewick, R., (1988), Algorithms, 2-nd Ed., Addison-Wesley Publishing Company.

[15] Schweitzer,P.J., (1968) Perturbation theory and finite Markov chains. Journal of Applied Probability, 5(3):401–404.

Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399