



The Effect of New Links on Google PageRank

Nelly Litvak, Konstantin Avrachenkov

► To cite this version:

Nelly Litvak, Konstantin Avrachenkov. The Effect of New Links on Google PageRank. [Research Report] RR-5256, INRIA. 2004, pp.12. inria-00070742

HAL Id: inria-00070742

<https://inria.hal.science/inria-00070742>

Submitted on 19 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Effect of New Links on Google PageRank

Konstantin Avrachenkov — Nelly Litvak

N° 5256

July 2004

Thème COM



*rapport
de recherche*

The Effect of New Links on Google PageRank

Konstantin Avrachenkov^{*}, Nelly Litvak[†]

Thème COM — Systèmes communicants
Projet MAESTRO

Rapport de recherche n° 5256 — July 2004 — 12 pages

Abstract: PageRank is one of the principle criteria according to which Google ranks Web pages. PageRank can be interpreted as a frequency of visiting a Web page by a random surfer and thus it reflects the popularity of a Web page. We study the effect of newly created links on Google PageRank. We discuss to what extent a page can control its PageRank. Using the asymptotic analysis we provide simple conditions that show if new links bring benefits to a Web page and its neighbors in terms of PageRank or they do not. Furthermore, we show that there exists an optimal linking strategy. We conclude that a Web page benefits from links inside its Web community and on the other hand irrelevant links penalize the Web pages and their Web communities.

Key-words: Google, PageRank, Rank One Update, Optimal Linking Strategy, Markov chains

The authors gratefully acknowledge the financial support of the Netherlands Organization for Scientific Research (NWO) under the grant VGP 61-520 and the French Organization EGIDE under Van Gogh grant no.05433UD.

^{*} 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

L'Effet de Nouveaux Liens sur le PageRank de Google

Résumé : PageRank est l'un des principaux critères selon lesquels Google ordonne les pages Web. PageRank peut être interprété comme la fréquence des visites d'une page Web par un utilisateur aléatoire, et reflète donc la popularité d'une page Web. Nous étudions l'effet des nouveaux liens sur le PageRank de Google. Nous discutons jusqu'à quel point une page peut modifier son PageRank. En utilisant l'analyse asymptotique, nous fournissons les conditions simples qui montrent si les nouveaux liens apportent des avantages à une page Web et aux pages voisines. Puis, 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, Rank One Update, Stratégie optimale d'établissement de liens, Chaînes de Markov

1 Introduction

Surfers on the Internet frequently use search engines to find pages satisfying their query. However, there are typically hundreds or thousands of relevant pages available on the Web. Thus, 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 [3] 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, the probability is spread among all pages of the Web, namely, $p_{ij} = 1/n$. 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

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

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

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

where $\underline{1}$ is a column vector of ones. The row 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 the surfer is at page i .

In order to keep up with constant modifications of the Web structure, Google updates its PageRank at least once per month. According to publicly available information Google still uses simple power iterations to compute the PageRank. Several proposals [1, 5, 6, 7, 10, 11, 12, 13] (see also an extensive survey paper by Langville and Meyer [9]) have recently been put forward to accelerate the PageRank computation. This research direction can be regarded as a system point of view. On contrary, here we take a user point of view and try to answer the following questions: When do new links benefit the Web page from which they emanate? When do the pages from the same Web community benefit from the link creation? Is there optimal linking strategy?

The paper is organized as follows: In Section 2, we study a question to what extent a page can control its PageRank. Then in the ensuing Section 3 we quantify the preliminary analysis of Section 2 with the help of Sherman-Morrison-Woodbury updating formula and derive the expression of new PageRank after the addition of new links. In Section 4 using asymptotic analysis we obtain natural conditions that show if a newly created link is beneficial or not. In Section 5 we demonstrate that there exists an optimal linking strategy. Finally, we provide conclusions in Section 6.

2 To what extend a page can control its PageRank?

The PageRank defined in (2) clearly depends on both incoming and outgoing links of a page. Thus, the easiest way for a page to change its ranking is to modify the outgoing links. Below we shall show that the PageRank can be written as a product of three terms where only one term depends on outgoing links. It will allow us to estimate to what extend a page can control its PageRank. To this end, we first recall the following useful expression for the PageRank [2, 9, 13]:

$$\pi = \frac{1-c}{n} \mathbf{1}^T [I - cP]^{-1}. \quad (3)$$

Let z_{ij} denote an (i, j) element of the matrix $Z = [I - cP]^{-1}$. Namely, we have

$$z_{ij} = e_i^T [I - cP]^{-1} e_j, \quad i, j = 1, \dots, n, \quad (4)$$

where e_k is the k -th column of the identity matrix I . Now, define a discrete-time absorbing Markov chain $\{X_t, t = 0, 1, \dots\}$ with the state space $\{0, 1, \dots, n\}$, where transitions between the states $1, \dots, n$ are conducted by the matrix cP , and the state 0 is absorbing. Let N_i be the number of visits to state $j = 1, \dots, n$ before absorption. Formally,

$$N_j = \sum_{t=0}^{\infty} \mathbf{1}_{\{X_t=j\}}, \quad j = 1, \dots, n,$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. It is easy to see that the value z_{ij} is the conditional expectation of N_j given that the initial state is i . Let q_{ij} be the probability to reach the state j before absorption if the initial state is i . Using the strong Markov property, we can now establish the following decomposition result.

Proposition 1 *The PageRank of page $i = 1, \dots, n$ is given by*

$$\pi_i = \frac{1-c}{n} z_{ii} \left(1 + \sum_{\substack{j=1 \\ j \neq i}}^n q_{ji} \right), \quad i = 1, \dots, n. \quad (5)$$

Proof: It follows from (3) that

$$\pi_i = \frac{1-c}{n} \mathbf{1}^T [I - cP]^{-1} e_i = \frac{1-c}{n} \sum_{j=1}^n z_{ji}. \quad (6)$$

Further, we have $\mathbb{P}(N_i = 0 | X_0 = j) = 1 - q_{ji}$, and using the strong Markov property, we can write for any $k \geq 1$

$$\begin{aligned} \mathbb{P}(N_i = k | X_0 = j) &= \mathbb{P}(N_i = k, N_i \geq 1 | X_0 = j) \\ &= \mathbb{P}(N_i \geq 1 | X_0 = j) \mathbb{P}(N_i = k | X_0 = j, N_i \geq 1) = q_{ji} \mathbb{P}(N_i = k | X_0 = i). \end{aligned}$$

Consequently, for any $i, j = 1, \dots, n; i \neq j$, we have

$$z_{ji} = \mathbb{E}(N_i | X_0 = j) = q_{ji} \mathbb{E}(N_i | X_0 = i) = q_{ji} z_{ii}.$$

Substituting the last equation in (6) we immediately obtain (5). \square

The decomposition formula (5) represents the PageRank of page i as a product of three multipliers where only the term z_{ii} depends on the outgoing links of page i . Hence, by changing the outgoing links, a page can control its PageRank up to a multiple factor $z_{ii} = 1/(1 - q_{ii}) \in [1, (1 - c^2)^{-1}]$, where $q_{ii} \in [0, c^2]$ is a probability to return back to i starting from i . Note that the upper bound $(1 - c^2)^{-1}$ (approximately 3.6 for $c = .85$) is hard or rather not possible to achieve because in this case a page i points to pages that are linking only to i . Such a policy makes i and its neighbors isolated from the rest of the Web.

If page i does not have outgoing links, then z_{ii} is very close to the lower bound 1, since there is almost no chance to return from i back to i before absorption. Bianchini et al. [2] used a circuit analysis to study in detail the influence of pages without outgoing links (leaves, dangling nodes) on the PageRank of a Web community. The authors of [2] came to the same conclusion that dangling causes a considerable loss in ranking. The formula (5) helps to quantify this loss.

We note that even a threefold increase of the PageRank might not be considered as a significant improvement, since Google measures the PageRank on a logarithmic scale [15]. In the ensuing sections we show how a Web page should use its scarce resources to increase its PageRank.

3 Rank One update of Google PageRank

Let us consider a Web page with k_0 old hyperlinks and k_1 newly created hyperlinks. Without loss of generality, we assume that the page with new links has index 1 and the pages towards which new links are pointed have indices from 2 to $k_1 + 1$. Then, the addition of new links can be regarded as one rank update of the hyperlink matrix

$$\tilde{P} = P + e_1 v^T, \tag{7}$$

with

$$v^T = \frac{k_1}{k} \left(\frac{1}{k_1} \sum_{i=2}^{k_1+1} e_i^T - p_1^T \right),$$

and where $k = k_0 + k_1$, p_1^T is the first row of matrix P . In [10] the authors use updating formulae to accelerate the PageRank computation. By restricting ourselves to the case of rank one update, we are able to perform a more comprehensive analysis. The next theorem provides updating formulae for the PageRank elements.

Theorem 1 *Let k_1 new links emanating from page 1 be added. Then, the elements of new PageRank vector are given by the following updating formulae*

$$\tilde{\pi}_1 = \frac{\pi_1}{1 - \frac{k_1}{k} \left(1 + \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{i1} - z_{11}\right)} \quad (8)$$

$$\tilde{\pi}_j = \pi_j + \pi_1 \frac{\frac{k_1}{k} \left(\frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{ij} - z_{1j}\right)}{1 - \frac{k_1}{k} \left(1 + \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{i1} - z_{11}\right)}, \quad j = 2, \dots, n. \quad (9)$$

Proof: Applying the Sherman-Morrison-Woodbury updating formula [4] to $[I - c\tilde{P}]^{-1}$, we can write

$$[I - c\tilde{P}]^{-1} = [I - cP]^{-1} + c \frac{[I - cP]^{-1} e_1 v^T [I - cP]^{-1}}{1 - cv^T [I - cP]^{-1} e_1}.$$

Then, premultiplying the above equation by $\frac{1-c}{n} \mathbf{1}^T$ and using (3), we get

$$\tilde{\pi} = \pi + \pi_1 \frac{cv^T [I - cP]^{-1}}{1 - cv^T [I - cP]^{-1} e_1},$$

and consequently,

$$\tilde{\pi}_1 = \pi_1 \frac{1}{1 - cv^T [I - cP]^{-1} e_1}, \quad (10)$$

$$\tilde{\pi}_j = \pi_j + \pi_1 \frac{cv^T [I - cP]^{-1} e_j}{1 - cv^T [I - cP]^{-1} e_1}, \quad j = 2, \dots, n. \quad (11)$$

We evaluate next $cv^T [I - cP]^{-1} e_j$ for $j = 1, \dots, n$,

$$cv^T [I - cP]^{-1} e_j = cv^T Z e_j = c \frac{k_1}{k} \left(\frac{1}{k_1} \sum_{i=2}^{k_1+1} e_i^T - p_1^T \right) Z e_j = \frac{k_1}{k} \left(\frac{c}{k_1} \sum_{i=2}^{k_1+1} e_i^T Z e_j - cp_1^T Z e_j \right)$$

Since $cPZ = Z - I$, we have $cp_1^T Z = z_1^T - e_1^T$, where z_1^T is the 1-st row of matrix Z , and hence $cp_1^T Z e_j = z_{1j} - e_1^T e_j$. Thus, we get

$$cv^T [I - cP]^{-1} e_j = \frac{k_1}{k} \left(\frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{ij} - (z_{1j} - e_1^T e_j) \right).$$

Substituting the above expression for $cv^T [I - cP]^{-1} e_j, j = 1, \dots, n$, into (10) and (11), we obtain (8) and (9). \square

The results in Theorem 1 are in line with formula (5). If page 1 updates its outgoing links then in the decomposition formula for π_1 only the second multiplier will be affected.

In the new situation, the probability \tilde{q}_{11} to return to page 1 starting from this page, is given by

$$\tilde{q}_{11} = \frac{k - k_1}{k} q_{11} + \frac{c}{k} \sum_{i=2}^{k_1+1} q_{i1}.$$

Substituting this expression in

$$\tilde{\pi}_1 = \frac{\tilde{z}_{11}}{z_{11}} \pi_1 = \frac{1 - q_{11}}{1 - \tilde{q}_{11}} \pi_1,$$

we get the updating formula (8). Note that according to (8) the ranking of page 1 increases when

$$1 + \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{i1} - z_{11} > 0, \quad (12)$$

which is equivalent to

$$\frac{1}{k_1} \sum_{i=2}^{k_1+1} q_{i1} > q_{11}.$$

Hence, the page 1 increases its ranking when it refers to pages that are characterized by a high value of q_{i1} . These must be the pages that refer to page 1 or at least belong to the same Web community. Here by a Web community we mean a set of Web pages that a surfer can reach from one to another in a relatively small number of steps.

Let us now consider formula (9). First, we see that the difference between the old and the new ranking of page j is proportional to $\tilde{\pi}_1$. Naturally, hyperlink references from pages with high ranking have a greater impact on other pages. Furthermore, the PageRank of page j increases if

$$\frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{ij} > z_{1j}, \quad (13)$$

Indeed, if this inequality holds then the link update by page 1 leads to higher probabilities of reaching page j from any page that has a path to j via page 1. Thus, in formula (5), the third multiplier will increase and the second multiplier will either increase or remain the same depending on whether there is a hyperlink path from j back to j passing through page 1. Note that if several links are added by page 1 then it might happen that some pages receive new links and loose in ranking at the same time. Such situation occurs when a new incoming link is created simultaneously with several other links, which point to “irrelevant” pages. The asymptotic analysis in the next section allows us to further clarify this issue.

4 Asymptotic analysis

Let us apply the asymptotic analysis to formulae (8) and (9) when c is sufficiently close to one and the Markov chain induced by the hyperlink matrix P is irreducible. The asymptotic

approach allows us to derive simple natural conditions that show if a Web page with newly created links and its neighbors benefit from these new links.

Theorem 2 *Let c be sufficiently close to one and let page 1 has k_1 new links to pages $\{2, \dots, k_1+1\}$. Assume that the Markov chain induced by the hyperlink matrix P is irreducible. Then, we have the following conditions:*

1. *if $m_{11} > 1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} m_{i1}$, the creation of new links $\{1 \rightarrow 2, \dots, 1 \rightarrow k_1+1\}$ increases the PageRank of page 1;*
2. *if $m_{1j} > 1 + \frac{1}{k_1} \sum_{i=2, i \neq j}^{k_1+1} m_{ij}$, the creation of new links $\{1 \rightarrow 2, \dots, 1 \rightarrow k_1+1\}$ increases the PageRank of page j from the set $\{2, \dots, k_1+1\}$;*
3. *if $m_{1l} > 1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} m_{il}$, the creation of new links $\{1 \rightarrow 2, \dots, 1 \rightarrow k_1+1\}$ increases the PageRank of page l , for $l > k_1+1$;*

where m_{ij} is the mean first passage time from page i to page j if $c = 1$.

Proof: First, we make a change of variable $c = 1/(1 + \rho)$, with $\rho > 0$ in the formula for $Z(c)$ as follows:

$$Z(c) = [I - cP]^{-1} = [I - \frac{1}{1+\rho}P]^{-1} = (1+\rho)[\rho I - (P - I)]^{-1}.$$

Then, we use the resolvent Laurent series expansion for the Markov chain generator (see e.g., [14, Theorem 8.2.3])

$$Z(c) = [I - cP]^{-1} = (1+\rho)[\rho I - (P - I)]^{-1} = (1+\rho)\left[\frac{1}{\rho}\Pi + H + \sum_{k=1}^{\infty} \rho^k (-1)^k H^{k+1}\right],$$

where $\Pi = \underline{1}\pi$ is the ergodic projection and H is the deviation matrix of the Markov chain induced by P . Since we consider a finite state Markov chain, the above series has a non-zero radius of convergence.

Let us now prove the first statement of the theorem. It is enough to show that Condition 1 guarantees that

$$1 + \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{i1}(c) - z_{11}(c) > 0$$

for all c sufficiently close to one, or equivalently, for all ρ sufficiently close to zero.

$$\begin{aligned} 1 + \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{i1}(c) - z_{11}(c) &= 1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} \left(\frac{1}{\rho} \pi_1 + h_{i1} + o(\rho) \right) - (1+\rho) \left(\frac{1}{\rho} \pi_1 + h_{11} + o(\rho) \right) \\ &= 1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} h_{i1} - h_{11} - \pi_1 + o(\rho) \end{aligned}$$

Since $m_{ij} = (h_{jj} - h_{ij})/\pi_j$ for $i \neq j$ [8], we have

$$1 + \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{i1}(c) - z_{11}(c) = 1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} (h_{11} - m_{i1}\pi_1) - h_{11} - \pi_1 + o(\rho) = 1 - \pi_1 \left(1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} m_{i1} \right).$$

Then, the condition

$$1 - \pi_1 \left(1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} m_{i1} \right) > 0$$

is equivalent to

$$m_{11} > 1 + \frac{1}{k_1} \sum_{i=2}^{k_1+1} m_{i1},$$

since $m_{11} = 1/\pi_1$.

Next we prove the second statement of the theorem. To prove the second statement, it is enough to show that Condition 2 implies that for a given $j \in \{2, \dots, k_1 + 1\}$

$$\frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{ij}(c) - z_{1j}(c) > 0$$

for all c sufficiently close to one, or equivalently, for all ρ sufficiently close to zero.

$$\begin{aligned} \frac{c}{k_1} \sum_{i=2}^{k_1+1} z_{ij}(c) - z_{1j}(c) &= \frac{1}{k_1} \sum_{i=2}^{k_1+1} \left(\frac{1}{\rho} \pi_j + h_{ij} + o(\rho) \right) - (1 + \rho) \left(\frac{1}{\rho} \pi_j + h_{1j} + o(\rho) \right) \\ &= \frac{1}{k_1} \sum_{i=2}^{k_1+1} h_{ij} - h_{1j} - \pi_j + o(\rho) = \frac{1}{k_1} \sum_{i=2, i \neq j}^{k_1+1} (h_{ij} - h_{1j} - \pi_j) + \frac{1}{k_1} (h_{jj} - h_{1j} - \pi_j) + o(\rho) \\ &= \frac{1}{k_1} \sum_{i=2, i \neq j}^{k_1+1} (h_{jj} - m_{ij}\pi_j - h_{jj} + m_{1j}\pi_j - \pi_j) + \frac{1}{k_1} (m_{1j}\pi_j - \pi_j) + o(\rho) \\ &= \frac{\pi_j}{k_1} \left(\sum_{i=2, i \neq j}^{k_1+1} (m_{1j} - m_{ij} - 1) + (m_{1j} - 1) \right) + o(\rho) \end{aligned}$$

The latter expression is positive for all sufficiently small ρ , if

$$\sum_{i=2, i \neq j}^{k_1+1} (m_{1j} - m_{ij} - 1) + (m_{1j} - 1) > 0,$$

which is equivalent to Condition 2. The proof of Condition 3 is very similar to the proofs of Conditions 1 and 2.

□

All conditions of Theorem 2 have very clear probabilistic interpretations. These interpretations become even more transparent in the case when only one new link is added. For instance, in the case of a single new link Condition 1 takes the form $m_{11} > m_{21} + 1$. The latter means that the average path from page 2 to page 1 should be shorter at least by one hop than the average return path from page 1 to itself. Condition 2 becomes $m_{12} > 1$, which is always true. This is not surprising as we know that an addition of a single new link pointing to a Web page always increases its PageRank (see e.g., [2]). However, as mentioned at the end of the previous section, we would like to note that if a page creates several new links, the pages towards which these new links point might not always benefit from the addition of new links. To clarify this idea, assume that the new links are been created one by one. When some page gives a link to our page, we will always benefit from it. However, if after pointing to us, this page also links to many other pages, which, on average, have very long paths back to our page, then it might eventually damage our ranking. To summarize, incoming links are especially valuable if they are received from popular pages that are dedicated to our Web community.

There is a striking similarity between Conditions 1–3 from Theorem 2 and conditions (12) and (13) derived in the previous section. However, in the asymptotic case, the conditions of improving the PageRank are expressed in terms of mean first passage times rather than mean number of visits before absorption.

5 Optimal linking strategy

Let us show that there exists in fact an optimal linking strategy. Consider a page $i = 1, \dots, n$ and assume that i has links to the pages i_1, \dots, i_k where $i_l \neq i$ for all $l = 1, \dots, k$. Then for the Google matrix \tilde{P} , we have

$$\mu_{ii} = 1 + \frac{c}{k} \sum_{l=1}^k \mu_{i_l i} + \frac{1}{n} (1 - c) \sum_{\substack{j=1 \\ j \neq i}}^n \mu_{ji}, \quad (14)$$

where μ_{ij} is the mean first passage time from page i to page j and c is the Google constant. The objective now is to choose k and i_1, \dots, i_k such that μ_{ii} becomes as small as possible and consequently $\pi_i = 1/\mu_{ii}$ becomes as large as possible. From (14) one can see that μ_{ii} 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 (14). As was also noted in Section 2, 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. Thus, we have the following statement

Theorem 3 *The optimal linking strategy for a Web page is to have only one outgoing link pointing to a Web page with a shortest mean first passage time back to the original page.*

Of course, in reality, it is not practical to link only to one page. However, the message from the above theorem is that one has to link to pages that are relevant and belong to the same Web community. 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 [16]. Indeed, it follows from our arguments that such policy will certainly increase the PageRank of all pages in a group.

6 Conclusions

A PageRank of a Web page is defined by the incoming links up to a multiple factor, which can be controlled by adding or removing outgoing links. We studied in detail the case when one Web page adds several new links and derived exact updating formulae for the PageRank. Using asymptotic analysis, we obtained simple natural conditions that show if a Web page with newly created links and its neighbors benefit from these new links. We conclude that having no outgoing link is a bad policy and that the best policy is to link to pages from the same Web community. Surprisingly, a new incoming link might be not a good news if a page that points to us gives many other irrelevant links at the same time.

References

- [1] A. ARASU, J. NOVAK, A. TOMKINS, AND J. TOMLIN (2002) PageRank Computation and the Structure of the Web: Experiments and Algorithms. In Proceedings of the 11-th International World Wide Web Conference.
- [2] BIANCHINI, M., GORI, M., SCARSELLI, F., (2002) Inside PageRank. ACM Trans. Internet Technology, In press. A shorter version has also appeared in the Proceedings of the 11-th International World Wide Web Conference.
- [3] BRIN, S., PAGE, L., MOTWAMI, R. AND WINOGRAD, T. (1998) The PageRank citation ranking: bringing order to the web. Stanford University Technical Report.
- [4] GOLUB, G. H. AND VAN LOAN, C. F. (1996) Matrix Computations, 3rd ed. Baltimore, MD: Johns Hopkins.
- [5] IPSEN, I. C. F., AND KIRKLAND, S. (2004) Convergence Analysis of an improved PageRank algorithm. North Carolina State University Technical Report CRSC-TR04-02.

- [6] 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.
- [7] KAMVAR, S.D., HAVELIWALA, T.H., MANNING, C.D. AND GOLUB, G.H. (2003) Extrapolation Methods for Accelerating PageRank Computations. In Proceedings of the 12-th International World Wide Web Conference.
- [8] KEMENY, J.G. AND SNELL, J.L., Finite Markov Chains, The University Series in Undergraduate Mathematics, Van Nostrand, Princeton, NJ, 1960,
- [9] LANGVILLE, A.N., AND MEYER, C.D. (2003) Deeper Inside PageRank. Preprint, North Carolina State University. Accepted to the Internet Mathematics Journal.
- [10] LANGVILLE, A.N., AND MEYER, C.D. (2004) Updating PageRank with Iterative Aggregation. In Proceedings of the 13-th World Wide Web Conference.
- [11] LANGVILLE, A.N. AND MEYER, C.D. (2004) A Reordering for the PageRank problem. NCSU CRSC Technical Report CRSC-TR04-16.
- [12] LEE, C.P.-C., GOLUB, G.H., AND ZENIOS, S.A. (2004) A Fast Two-Stage Algorithm for Computing PageRank. Stanford University Technical Report.
- [13] MOLER, C.D., AND MOLER, K.A., (2003) Numerical Computing with MATLAB, SIAM.
- [14] PUTERMAN, M.L., (1994) Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley.
- [15] <http://toolbar.google.com/>
- [16] <http://www.searchenginewatch.com/>



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