Local Aspects of the Global Ranking of Web Pages

Fabien Mathieu, Laurent Viennot

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Local Aspects of the Global Ranking of Web Pages

Fabien Mathieu and Laurent Viennot

No 5192

Mai 2004

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--5192--FR+ENG

Local Aspects of the Global Ranking of Web Pages

Fabien Mathieu and Laurent Viennot

Thème 1 — Réseaux et systèmes
Projet Gyroweb

Rapport de recherche n 5192 — Mai 2004 — 14 pages

Abstract: Started in 1998, the search engine *Google* sorts pages using several parameters. *PageRank* is one of those. Precisely, *PageRank* is a distribution of probability on the web pages that depends on the web graph. Our purpose is to show that the PageRank can split into two terms, an internal and an external PageRank. These two PageRanks allow a better comprehension of the PageRank signification inside and outside a site. A first application is a local algorithm to estimate the PageRank of a given site pages. We will also show quantitative results on the possibilities for a site to boost its own PageRank.

Key-words: Web, PageRank decomposition, blocks, flow

(Résumé : tsvp)

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : 01 39 63 55 11 - International : +33 1 39 63 55 11
Télécopie : (33) 01 39 63 53 30 - International : +33 1 39 63 53 30

Estimation locale de l'importance globale des pages d'un site web

Résumé : Depuis sa création en 1998, le moteur de recherche *Google* trie les pages web à l'aide de plusieurs paramètres. *PageRank* est l'un des plus connus. Le *PageRank* est une distribution de probabilité sur les pages web qui dépend uniquement du graphe du web. Nous nous proposons de montrer que le *PageRank* peut se décomposer en deux termes qui sont les PageRank internes et externes. Cette décomposition permet d'avoir une meilleure compréhension de la signification du PageRank à l'intérieur et à l'extérieur d'un site. Une première application est un algorithme local pour estimer le PageRank des pages d'un site donné. Nous montrons également quelques résultats quantitatifs sur les possibilités qu'a un site de modifier son propre PageRank.

Mots-clé : Web, décomposition du PageRank, blocs, flot

1 Introduction

PageRank [14] was a major algorithmic breakthrough for evaluating the importance of web pages achieved by exploiting the topology of the web induced by hyperlinks. Numerous work has then been devoted to better understand the relation between this web graph structure and the quality of web pages. Some authors have proposed alternative methods for ranking pages [10, 17] based on similar matrix computations. Other results propose different computation of an approximation of the PageRank either to obtain a faster algorithm [8] or an incremental algorithm [1].

This paper tries to model how the PageRank decomposes with regards to the site partition of the web. A site can be seen as the collection of pages on a given web server or more generally as a set of pages tightly related. As noted by [12, 11, 15], a block structure of the web adjacency matrix can be observed from an url-induced ordering of the pages, showing how an intrinsic site partition could be defined. This paper assumes that a site partition is given.

Even if one can naturally state that the web graph structure is tightly related to the site partition (most of the links are local), the web graph has mainly been studied disregarding this property. This is the case for the PageRank computation. In [8], a site partition is exploited to efficiently compute an approximated PageRank. On the other hand, this paper makes an exact decomposition of the PageRank computation, showing how the PageRank can be split into an internal PageRank (related to internal links of a site) and an external PageRank (related to inter-site links). In [13, 3], the sum of the PageRanks of the pages of a site is decomposed according to internal, incoming links, outgoing links and sinks. The authors give basic hints on how the link structure of site can alter its PageRanks. A stability property of the overall PageRanks when a site changes its internal link structure is also shown. Our model of decomposed PageRank allows to push forward their analysis to better understand how a site can alter its own PageRanks.

Another contribution of our site decomposition model of PageRank is a framework for evaluating locally the global PageRank. This could be useful for a local search engine to rank the pages of a site according to a global importance knowing only locally the web structure.

The paper is organized as follows. Section 2 introduces more formally the PageRank. Section 3 introduces our model for decomposing the PageRank according to a site partition of the web. Section 4 shows how to locally estimate the global PageRank of the pages of a site. Finally, Section 5 analyzes how a site administrator can alter the PageRank of its pages by modifying the links inside the site.

2 PageRank Definition

Let $G = (V,E)$ be an oriented aperiodic strongly connected graph, without self-loop, and $\mathcal{S} = (S_1, \ldots, S_k)$ a partition of G, with $k > 1$. G is supposed to be a web graph, and $\mathcal{S}$ a site partition of G (elements of $\mathcal{S}$ are sites).

If $d^+(v)$ is the out degree of $v \in V$, we can define the following stochastic matrix $A : V \times V \to \mathbb{R}^+$

$$A = (a_{i,j})_{i,j \in V}, \text{ with } a_{(i,j)} = \begin{cases} \frac{1}{d^+(i)} \text{ if } i \text{ links to } j \\ 0 \text{ else} \end{cases}$$

According to Markov processes theory[16], there is a unique probability P on V such that:

$$\forall v \in V,\ P(v) = \sum_{w \to v} \frac{P(w)}{d^+(w)} \tag{1}$$

The matrix version of this is:

$$P = A^t P, \tag{2}$$

where A^t is the transposed matrix of A.

The distribution probability P defines the PageRank of the graph G. This concept of PageRank was introduced by [14] in 1998 and used by the search engine *Google*[6].

Remark The web graph is far from being strongly connected[5]. Nevertheless there is techniques to override this by either altering the web graph or the calculation of P:

- *Page et al.* [14] suggests to compensate the flow leak in A by normalizing P at each iteration.
- *Haveliwala et al.* [7] turns A explicitly into a stochastic matrix by removing recursively pages without link.
- The dumping factor, introduced by [4], is used by Google on a graph where leaves are non-recursively removed and reinjected after P converged. The principle of the dumping factor is to replace A by $d.A + \frac{1-d}{|V|}\mathbf{1}.\mathbf{1}^t$, where $\mathbf{1}$ is a vector filled with ones and d the dumping factor. The new matrix represents a weighted strongly connected graph (Normalization is always needed, but the process converges faster).
- Finally, *Abiteboul et al.* [1] adds a virtual dumping page that links to and is linked to every other page.

Convention In the rest of this article (except 3.3, 5.2 and 5.3), we will suppose that A is aperiodic, strongly connected and without self-loop. The PageRank will be unambiguously the probability vector P solution of $P = A^t P$.

3 Internal PageRank, external PageRank

3.1 Notations

For $v \in V$, we call $S(v)$ the element of $\mathcal{S}$ such as $v \in S(v)$. We also define $\delta_{\mathcal{S}}$: $V \times V \rightarrow \{0,1\}$ as follows:

$$\delta_{\mathcal{S}}(v,w) = \begin{cases} 1 \text{ if S(v)=S(w)} \\ 0 \text{ else} \end{cases}$$

Let $A_{\mathcal{S}}$ be the matrix of the projection of A on the elements of $\mathcal{S}$: $A_{\mathcal{S}} = (a_{v,w}\delta_{\mathcal{S}}(v,w))_{v,w \in V}$.

We also need to define the internal degree d_i^+ (resp. the external degree d_e^+) of a vertex v as its out degree in the graph induced by $S(v)$ (resp. $\{v\} \cup (V \setminus S(v))$).

Lastly we can define the notions of internal and external PageRank, deduced from the PageRank P seen on formula (2):

– The incoming internal PageRank P_{ii} (resp. incoming external PageRank P_{ie}) of $v \in V$ is the probability to come in v from a page of $S(v)$ (resp. $V \setminus S(v)$), that is:

$$P_{ii} = A_{\mathcal{S}}^t P \tag{3}$$
$$P_{ie} = (A - A_{\mathcal{S}})^t P = P - P_{ii} \tag{4}$$

– The outgoing internal PageRank P_{oi} (resp. outgoing external PageRank P_{oe}) is the term $P(v)\frac{d_i^+(v)}{d^+(v)}$ (resp. $P(v)\frac{d_e^+(v)}{d^+(v)}$).

3.2 Conservation laws

Using the definitions, we have the following equation:

$$P = P_{ie} + P_{ii} = P_{oe} + P_{oi} \tag{5}$$

We can now give the internal and external conservation laws. For each $S \in \mathcal{S}$, we see that

$$\sum_{v \in S} P(v) = \sum_{v \in S} \sum_{w \rightarrow v} \frac{P(w)}{d^+(w)} = \sum_{(w,v) \in E \cap (V \times S)} \frac{P(w)}{d^+(w)} \tag{6}$$
$$= \sum_{(w,v) \in E \cap S^2} \frac{P(w)}{d^+(w)} + \sum_{(v,w) \in E \cap (S \times V \setminus S)} \frac{P(w)}{d^+(w)} \tag{7}$$
$$= \sum_{w \in S} P_{oi}(w) + \sum_{v \in S} P_{ie}(v) \tag{8}$$

We can deduce from (5) and (8) the external conservation law:

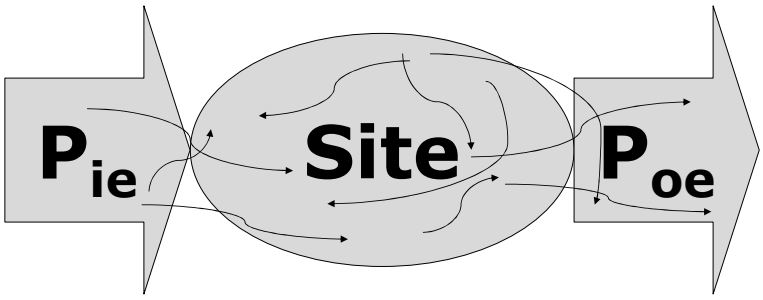

FIG. 1 – *External PageRank conservation:* $\sum_{v\in S} P_{ie}(v) = \sum_{v\in S} P_{oe}(v)$

$$\sum_{v\in S} P_{ie}(v) = \sum_{v\in S} P_{oe}(v) \tag{9}$$

and the internal conservation law.

$$\sum_{v\in S} P_{ii}(v) = \sum_{v\in S} P_{oi}(v). \tag{10}$$

The relation (9) shows that a site gives as much PageRank (outgoing external) that he receives (incoming external). If PageRank is a random surfer flow, there is a conservation of the external PageRank flow on the graph $G/\mathcal{S}$ (see figure 1). That remark will lead us to an intra-site and an inter-sites calculation of PageRank.

Remark If we formalize carefully the PageRank as a flow, we have another proof of (9): the PageRank is actually a stationary flow, so the flow on every subset S is stationary, therefore we have (9).

3.3 Dumping factor and flow

In this section, we keep supposing that G is leafless, but it is not necessarily strongly connected nor aperiodic.

As said in 2, using a dumping factor consists in replacing A by $d.A + \frac{1-d}{|V|}\mathbf{1}\mathbf{1}^t$. We have then a superposition of classic transitions ($d.A$) and dumping transitions ($\frac{1-d}{|V|}\mathbf{1}\mathbf{1}^t$). Dumping transitions are supposed to model the action of moving anywhere in the web without following any static link (use of *Bookmarks* or search engines, keyboard input,...).

Instead of spitting the dumping flow into an external one and an internal one, we find more interesting to introduce the notions of induced PageRank P_{ind} and dissipated PageRank P_{dis}.

We have now six different PageRanks corresponding to three types of flow as shown in figure 2 ($\times$ is the element by element product):

flow	incoming	outgoing
internal	$P_{ii} = dA_{\mathcal{S}}^t P$	$P_{oi} = d(A_{\mathcal{S}}\mathbf{1}).\times P$
external	$P_{ie} = d(A - A_{\mathcal{S}})^t P$	$P_{oe} = d((A - A_{\mathcal{S}})\mathbf{1}).\times P$
dumping	$P_{ind} = \frac{1-d}{\lvert V\rvert}\mathbf{1}$	$P_{dis} = (1-d)P$

FIG. 2 – *The different flows of PageRank in the dumping factor case*

We will assume that the whole dumping flow is external. Of course there are internal dumping transitions, but we choose to consider them external. Thus the internal flow conservation law does not change, but we have a new external flow conservation law:

$$\sum_{v\in S}(P_{ie}(v) + P_{ind}(v)) = \sum_{v\in S}(P_{oe}(v) + P_{dis}(v))\,,$$

that we will note

$$P_{ie}(S) + P_{ind}(S) = P_{oe}(S) + P_{dis}(S) \tag{11}$$

3.3.1 PageRank stability

The equation (11) shows the stability of the classic flow at the site level. From $P_{ind}(S) = (1-d)\frac{|S|}{|V|}$ and $P_{dis}(S) = (1-d)P(S)$, we can tell that for a site whose PageRank $P(S)$ is above (resp. below) the average (which is $\frac{|S|}{|V|}$ for a site of size $|S|$), the outgoing external PageRank $P_{oe}(S)$ is inferior (resp. superior) to the incoming external PageRank $P_{ie}(S)$. In other words, a *rich* site (in term of PageRank) will be greedy and will give less than he receives (dumping excluded), and vice versa. The dumping factor causes a retro-action that limit the phenomena of over-amplification that we will see in 5.2.

4 Local computation of the global Ranking

4.1 Relation between external PageRank and PageRank

From (3) and (4) we can write $A_{\mathcal{S}}^t.P = P - P_{ie}$, and then $P_{ie} = (Id - A_{\mathcal{S}}^t)P$, where Id is the identity matrix.

Lemma 1 *The matrix $(Id - A_{\mathcal{S}}^t)$ is inversible.*

Proof As G is strongly connected, there is links between sites. We have then $0 < A_{\mathcal{S}} < A$. $A_{\mathcal{S}}$ is strictly sub-stochatic, so its spectral radius is strictly inferior to 1. Therefore $(Id - A_{\mathcal{S}}^t)^{-1}$ exists.

■

Lemma 1 permits to express P as a function of P_{ie}:

$$P = (Id - A_{\mathcal{S}}^t)^{-1} P_{ie} \tag{12}$$

Knowing the incoming external PageRank P_{ie} of a site S, we can theoretically compute the PageRank of the pages of S with only the local graph G_S.

Remark $(Id - A_{\mathcal{S}}^t)^{-1} = \sum_{k=0}^{\infty} (A_{\mathcal{S}}^t)^k$ is a diagonal by blocks matrix, that can be interpreted as the transition matrix of all the internal paths.

4.2 External PageRank matrix

We want to translate the intuition of figure 1 in a conservation law with P_{ie} only. From (4) and (12), we have:

$$P_{ie} = (A - A_{\mathcal{S}})^t P = (A - A_{\mathcal{S}})^t (Id - A_{\mathcal{S}}^t)^{-1} P_{ie} \tag{13}$$

We have then the external PageRank transition matrix A_e:

$$A_e^t = (A - A_{\mathcal{S}})^t (Id - A_{\mathcal{S}}^t)^{-1}$$

Lemma 2 *The matrix A_e is stochastic.*

Proof We just have to show that the sum of each column of A_e^t is 1. First, we rewrite A_e^t:

$$\begin{aligned} A_e^t &= \sum_{k=0}^{\infty} \left(A^t (A_{\mathcal{S}}^t)^k - (A_{\mathcal{S}}^t)^{k+1} \right) \\ &= A^t + \sum_{k=1}^{\infty} \left(A^t (A_{\mathcal{S}}^t)^k - (A_{\mathcal{S}}^t)^k \right) \\ &= A^t + A^t M - M, \text{ with } M = \sum_{k=1}^{\infty} (A_{\mathcal{S}}^t)^k \end{aligned}$$

Then we consider the sum s_w of a column w in $A^t M$.

$$s_w = \sum_{u \in V} \sum_{v \in V} A_{u,v}^t M_{v,w} = \sum_{v \in V} \left(\sum_{u \in V} A_{u,v}^t \right) M_{v,w} = \sum_{v \in V} M_{v,w}$$

So the sum of each column of $A^t M - M$ is null; then A_e^t is stochastic as A^t.

■

4.3 Partially distributed PageRank algorithm

From (12) and (13) we can suggest a half-distributed algorithm for computing the PageRank:

- Each site S computes from its block of $A_{\mathcal{S}}$ a block of the matrix $(Id - A_{\mathcal{S}}^t)^{-1}$.
- The coefficients of A_e are centralized.
- The external PageRank P'_e associated with A_e is centrally computed using $A_e^t P'_e = P'_e$.
- Each site S gets its own PageRank thanks to the relation $P' = (Id - A_{\mathcal{S}}^t)^{-1} P'e$ applied to its block.

Lemma 3 *The vector P' we obtain is, once normalized, the PageRank P of G.*

Proof We have to show that P' is an eigenvector of A^t, and that its eigenvalue is 1:

$$\begin{aligned} A^t P' &= A^t (Id - A_{\mathcal{S}}^t)^{-1} P'_e \\ &= (A^t - A_{\mathcal{S}})(Id - A_{\mathcal{S}}^t)^{-1} P'_e + A_{\mathcal{S}}^t (Id - A_{\mathcal{S}}^t)^{-1} P'_e \\ &= A_e^t P'_e + \left((Id - A_{\mathcal{S}}^t)^{-1} - (Id - A_{\mathcal{S}}^t)(Id - A_{\mathcal{S}}^t)^{-1} \right) P'_e \\ &= P'_e + \left((Id - A_{\mathcal{S}}^t)^{-1} - Id \right) P'_e \\ &= P'_e + P' - P'_e = P' \end{aligned}$$

As the principal eigenvalue of A, that is 1, is unique, P and P' are homothetic, so $P = P'$ (after normalization).

■

4.4 Estimation of a site PageRank

A natural question is to ask if a site S can estimate the ranking of its pages only knowing local data. This can be very valuable for an internal search engine to be able to estimate the global ranking of its pages without crawling all the web or asking an external search engine. From (12), all we need is an estimation of the incoming external PageRank.

According to [14], PageRank models the statistic behaviour of surfers crawling the web. It seems then natural to estimate the PageRank of a page by the average hits it gets. More specifically, the incoming external PageRank should be proportional to the average hits from outside the site. So each site can get an estimation of the incoming external rank from analysing the logs files of its web server.

Abiteboul et al. [1] states that the incoming degree is a good estimation of the PageRank. Thus the number of external references for each page (obtained from the logs files) is another estimation of P_{ie}.

Both estimation methods of the incoming external PageRank will be furthered studied in future work.

Once P_{ie} known, you just have to compute $P = (Id - A_{\mathcal{S}}^t)^{-1} P_{ie}$. In fact, you do not have to calculate $(Id - A_{\mathcal{S}}^t)^{-1}$ explicitly. It is better to resolve $P = A_{\mathcal{S}}^t P + P_{ie}$ using iterative methods.

For example, choosing P_0 and iterating

$$P_{n+1} = A_{\mathcal{S}}^t P_n + P_{ie}$$

converges, because the spectral radius of $A_{\mathcal{S}}$ is strictly inferior to 1. Empirical results from [14] suggest a fast convergence of that sort of algorithm applied to web graphs.

Remark There are lot of methods to improve the convergence of that sort of iterative computation ([2], [9]). The purpose of this paper is not to optimize this part of the computation, so we will stay with the basic Jacobi method.

4.4.1 Interest of our method

Why are we not keeping the average hits per page as an estimation of the PageRank? We believe our method can give a better PageRank to pages newly created, that do not get a lot of hits yet but are well linked and will surely get known.

Another advantage is that the P_{ie} input can be very flexible. By example, assuming the incoming degree is often a good estimation of the PageRank, it can be set to the number of external references for each page (obtained from the logs files). The webmaster could also manually alter P_{ie} to promote some pages while keeping a minimum of ranking.

5 Locally altering the PageRank

Our decomposition of the PageRank explains some results about the ability that a site has to alter its own PageRank. A first approximation is to say that if a site can not alter the external PageRank, it is not the same for the internal PageRank.

5.1 Amplification factor

Let S be a site, $P(S) := \sum_{v \in S} P(v)$ and $P_{ie}(S) := \sum_{v \in S} P_{ie}(v)$. We can define the amplification factor of S by $\alpha(S) = \frac{P(S)}{P_{ie}(S)}$. This factor depends on both S and the direction of P_{ie}, but knowing S we can estimate $\alpha(S)$.

Lemma 4 *The amplification factor can be estimated by:*

$$\frac{1}{1-\omega} \leq \alpha(S) \leq \frac{1}{1-\Omega} \tag{14}$$

with $\omega = \min_{v \in S} \frac{d_i(v)}{d(v)}$ *and* $\Omega = \max_{v \in S} \frac{d_i(v)}{d(v)}$.

Proof For each basic vector e_v, $v \in S$, we have
$\|A_{\mathcal{S}}(e_v)\|_1 = \frac{d_i(v)}{d(v)}$, therefore $\omega \|(X)\|_1 \leq \|A_{\mathcal{S}} X\|_1 \leq \Omega \|X\|_1$ for each vector $X > 0$ defined in S.

The first inequality of (14) is obtained as follows:

$$\begin{aligned} P(S) &= \sum_{v \in S} P(v) = \left\| \sum_{k \in \mathbb{N}} (A_S^t)^k (P_{ie}) \right\|_1 \\ &\geq \sum_{k=0}^{\infty} \omega^k \|P_{ie}\|_1 = \frac{1}{1-\omega} P_{ie}(S) \end{aligned}$$

We get the second inequality the same way.

■

The consequences of this amplification system is that a site can arbitrarily increase its PageRank. In the limit case where there is no external link [1], we have a short-circuit phenomena. This fact is well-known: if there is some subsets strongly connected, those subsets will absorb all the PageRank (sink hole phenomena).

Fortunately, we will see how the dumping factor reduces this effect.

5.2 Dumping and amplification

The assumptions are those of 3.3. In particular, the transition matrix is $d.A + \frac{1-d}{|V|}\mathbf{1}\mathbf{1}^t$; previous results stay valid replacing A by dA and P_{ie} by the total incoming external PageRank $P_{ie} + P_{ind}$.

Lemma 5 *The estimation for the amplification factor* $\alpha'(S) = \frac{P(S)}{P_{ie}(S)+P_{ind}(S)}$ *is:*

$$\frac{1}{1-d\omega} \leq \alpha'(S) \leq \frac{1}{1-d\Omega}. \tag{15}$$

Proof It is the same that for (14); we can write:

$$\begin{aligned} P(S) &= \sum_{v \in S} P(v) = \left\| \sum_{k \in \mathbb{N}} (dA_S^t)^k (P_{ie} + \frac{1-d}{|V|}\mathbf{1}) \right\|_1 \\ &\leq \sum_{k=0}^{\infty} (d\Omega)^k (\|P_{ie}\|_1 + (1-d)\frac{\|\mathbf{1}\|_1}{|V|}) \\ &\leq \frac{1}{1-d\Omega}(P_{ie}(S) + (1-d)\frac{|S|}{|V|}) \end{aligned}$$

1. A real site does not have to respect the assumptions of this article. In particular, many commercial sites do not have any external link[5].

We get the second inequality the same way.

■

5.2.1 Numerical Value

It is not impossible for a real site to have $\omega = \Omega = 0$ (site without internal link) or $\omega = \Omega = 1$ (site without external link). So the amplification factor can vary between 1 and $\frac{1}{1-d}$. The empirical value of d being 0.85, we deduce that with a fixed incoming external PageRank, the PageRank of a site can fluctuate up to a factor $\frac{20}{3}$...

5.2.2 PageRank robustness

Bianchini et al. [3] states that the effect that a site can produce onto the web is bounded by the PageRank of this site. If we consider two instants t and $t+1$, they suggest that:

$$\sum_{v \in V} |P_t(v) - P_{t+1}(v)| \leq \frac{2d}{1-d} \sum_{s \in S} P_t(s)$$

Lemma 5 leads to this result: if the site S changes between t and $t+1$, the PageRank variation inside S is at most $\frac{d}{1-d}P(S)$, implying a variation up to another $\frac{d}{1-d}P(S)$ outside the site, since the total PageRank stays equal to 1.

5.3 Amplification of a given page

When a surfer uses *Google*, the results are not sites but pages. So what is important for a site is not to have a big average PageRank, but to be able to concentrate this PageRank on a few pages, or even on a single one. Then we can ask: let S be a site of $n+1$ pages and P_{ie} its incoming external PageRank. How can we maximize the PageRank of a given page $v_0 \in S$?

The answer is not difficult once we remark the optimal link structure is when v_0 links to all other pages of S and all other pages of S link to v_0 and only v_0 [2]. It is not hard then to have a limitation of $P(v_0)$:

$$P(v_0) \leq \frac{P_{ie}(S)}{1-d^2} + \frac{1+nd}{(1+d)\,|V|}, \tag{16}$$

with equality if, and only if $P_{ie}(S) = P_{ie}(v_0)$.

(16) shows some strategies to improve its PageRank [3]. For example:

– If v_0 links to all other pages without return [4], adding the links to v_0 can increase the PageRank of v_0 up to $\frac{1}{1-d^2} \simeq 3{,}6$.

2. PageRanks algorithms systematically remove self-loops, so a single page cannot amplify itself.

3. In fact, it seems that *Google* is rather aware of these strategies, so they do not work as well as they should in theory...

4. A typical situation with sites using *frames*.

– The optimal strategy ensures for v_0 a minimal PageRank at least equal to the average PageRank $\frac{1}{|V|}$ even if P_{ie} is null.
– If $1 \ll n \leq |V|$ (dynamically generated pages linking to v_0), the ratio $\frac{P(v_0)}{P_{average}}$ is about $\frac{d}{1+d}n$.

6 Conclusion

We have proposed a decomposition of the PageRank flow in accordance with the notion of site, showing how to use it for estimating locally the global PageRanks inside a site. However, this relies on estimating the incoming PageRank either with real user hits or external referer counts. Further experiments are needed for fully validating this approach. Another interesting research direction includes distributed computation of the PageRank: assuming that several sites collaborate, how to compute the PageRank induced by their union? Our model is certainly the first step for that. It can also be useful for evaluating approaches that alter the PageRank computation based on a site decomposition as proposed by [8] for speeding up the computation. Another related issue is the identification and the ranking of sites rather than pages.

At least, the flow decomposition has allowed to analyze some strategies that the webmasters could use if an unrefined version of PageRank was used by search engines. We have shown that the PageRank defined in [14] can be very versatile when subject to non-cooperative strategies. It also seems that P_{ie} can be more robust, assuming we are able to find a site partition $\mathcal{S}$ that reflects the reality.

Références

[1] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proceedings of the twelfth international conference on World Wide Web*, pages 280–290. ACM Press, 2003.

[2] G. Allaire and S. M. Kaber. *Algèbre linéaire numérique*. Ellipses, 1998.

[3] M. Bianchini, M. Gori, and F. Scarselli. Pagerank: A circuital analysis.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[5] A. B. et al. Graph structure in the web. In *Proc. 9th International World Wide Web Conference*, pages 309–320, 2000.

[6] Google. http://www.google.com/, 1998.

[7] T. Haveliwala. Efficient computation of PageRank. Technical report, Computer Science Department, Stanford University, 1999.

[8] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank, 2003.

[9] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[10] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, California, 25–27 Jan. 1998.

[11] F. Mathieu and L. Viennot. Structure intrinsèque du web.

[12] F. Mathieu and L. Viennot. Local structure in the web. In *12-th international conference on the World Wide Web*, 2003. poster.

[13] M. G. Monica Bianchini and F. Scarselli. Inside pagerank. In *ACM Transactions on Internet Technology*, 2003. To be published.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Computer Science Department, Stanford University, 1998.

[15] S. Raghavan and H. Garcia-Molina. Representing web graphs, 2003.

[16] L. Saloff-Coste. Lectures on finite Markov chains. In G. G. E. Giné and L. Saloff-Coste, editors, *Lecture Notes on Probability Theory and Statistics*, number 1665 in LNM, pages 301–413. Springer Verlag, 1996.

[17] P. P. Senellart and V. D. Blondel. Automatic discovery of similar words. In M. W. Berry, editor, *A Comprehensive Survey of Text Mining*. Springer-Verlag, 2003. To be published.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

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