PageRank and Web Communities

M. Bianchini, M. Gori, and F. Scarselli
Dipartimento di Ingegneria dell’Informazione
Università degli Studi di Siena
Via Roma, 56 — 53100 Siena (ITALY)
Email:{monica,marco,franco}@dii.unisi.it

Abstract

The definition of the ordering of the Web pages, returned on a given query, is a crucial topic, which gives rise to the notion of Web visibility. A fundamental contribution towards the conception of appropriate ordering criteria has been given by means of the introduction of PageRank, which takes into account only the hyperlinked structure of the Web, regardless of the content of the pages. In this paper, we introduce a circuit analysis which allows us to understand the distribution of PageRank, and show some basic results for understanding the way it migrates amongst communities. In particular, we highlight some topological properties which suggest methods for the promotion of Web communities. These results confirm the importance and the effectiveness of PageRank for discovering relevant information but, at the same time, point out its vulnerability to spamming.

1 Introduction

Scoring algorithms, derived from Information Retrieval (IR), employ similarity measures that only take into account the page content and neglect the graphical structure of the Web. These approaches are prone to be cheated, so that pages can be highly ranked if they contain irrelevant but popular words, appropriately located in the page (e.g. in the title).

PageRank [6] is used by Google together with a number of different factors, including standard IR measures, proximity, and anchor text (text of links pointing to Web pages), in order to find the most relevant answers to a given query. The way these factors are computed with PageRank is not of public domain, but the ordering returned on the queries is affected by both the page content and the hyperlinks.

In spite of its relevance, the theoretical properties of PageRank are only partially understood. In order to explain the computational properties of the algorithm, most of the authors [6, 14, 18] cite the general theory of Markov chains [12, 16]. However, such a theory can be applied to PageRank only under the assumption that the Web does not contain dangling pages 1. A related popular ranking algorithm, proposed in [11] and called HITS, computes two values for each page: the degree of authority and the degree of hubness. Whereas the authority is a measure of the importance of the page, the hubness is a measure of the usefulness of the page to act as a starting point for a surfer which wants to find important documents. PageRank and HITS belong to a large class of ranking algorithms, where the scores can be computed as a fixed point of a linear equation [9]. A completely different approach, rooted to statistics, was proposed in [7, 8]. Such approach is based on Bibliometrics, a sequence of techniques that seeks to quantify the process of written communication. However, starting from PageRank and HITS, some extensions have been proposed by hybrid solutions [1, 3, 9, 10, 15, 18].

In this paper, we look inside PageRank to disclose its fundamental properties concerning the score distribution in Web communities. We introduce the notion of energy, which simply represents the sum of the score of all the pages of a given community, and propose a general circuit analysis which allows us to understand the distribution of PageRank. In addition, the derived energy balance equations make it possible to understand the way different Web communities interact each other and to disclose some secrets for promotion of Web pages. In particular, it is pointed out that the energy of a given target community can be driven by a “promoting community” so as to grow linearly with the number of its pages. This property holds regardless of the structure of the promoting community, which makes it very hard its detection. The results of this pa-

1Dangling pages are those with no hyperlinks.
per suggest the importance of combining topological-based methods like PageRank with intelligent models for checking the quality of the pages. Moreover, they motivates recent efforts towards the conception of focused page ranking schemes which propagate the score of topic selected pages [9, 10, 15, 17]. In particular, the limitations pointed out in this paper concerning the linear growth of PageRank can be overcome by any related algorithm which propagates only the score of pages meeting some quality requirements. This places the problem of page scoring naturally in the field of artificial intelligence.

The paper is organized as follows. In the next section, PageRank is briefly reviewed. In Section 3, the interaction among communities is discussed, while in Section 4, we analyze community promotion. Finally, some conclusions are drawn in Section 5.

2 PageRank

The basic idea of PageRank is that of introducing a notion of page authority which is independent of the page content. Such an authority measure only emerges from the topological structure of the Web. In PageRank, the authority resembles the notion of citation in the scientific literature. In particular, the authority of a page $p$ depends on the number of incoming hyperlinks (number of citations) and on the authority of the page $q$ which cites $p$ with a forward link. Moreover, selective citations from $q$ to $p$ are assumed to provide more contribution to the score of $p$ than uniform citations.

Hence, PageRank $x_p$ of $p$ is computed by taking into account the set of pages $pa[p]$ pointing to $p$. According to [5]:

$$x_p = d \sum_{q \in pa[p]} \frac{x_q}{h_q} + (1 - d). \tag{1}$$

Here $d \in (0, 1)$ is a DAMPING FACTOR and $h_q$ is the out-degree of $q$, that is the number of hyperlinks outgoing from $q$. When stacking all the $x_p$ into a vector $x$, we get

$$x = dWx + (1 - d)\mathbf{1}_N, \tag{2}$$

where $\mathbf{1}_N = [1,\ldots,1]^T$ and $W = \{w_{i,j}\}$ — the transition matrix — is such that $w_{i,j} = 1/h_i$ if there is a hyperlink from $j$ to $i$ and $w_{i,j} = 0$, otherwise.\(^2\)

3 Communities and energy balance

In the following, the Web is represented by a directed graph $G_W = (P, H)$, where each page $p \in P$ is a node and a hyperlink between two pages $h \in H$ is an edge between the corresponding nodes. The set of pages pointed by $p$ is denoted by $cp[p]$, whereas the pages that do not contain hyperlinks are called dangling pages. Moreover, $|\cdot|$ denotes the cardinality operator on sets and the modulus operator on reals, and $||V||_1 = \sum |v_i|$ is the 1-norm of the array $V = [v_1,\ldots,v_n]^T$.

Given a set of pages $I \subset P$, a community $G_I$ is the subgraph of the Web, $G_I = (I, H_I)$, that contains all the hyperlinks between pages in $I$. If the pages belonging to $I$ are disconnected from the rest of the Web, then we say that $G_I$ is an island. The energy $E_I$ of a community $G_I$ is given by $E_I = \sum_{p \in I} x_p$, where $x_p$ is defined by eq. (1). Finally, $out(I) \subset I$ denotes the pages of the community that point to other pages not in $I$, and $in(I)$ represents the pages not belonging to $I$ and pointing to pages in $I$.

Energy balance

The remainder of this section is devoted to a simple analysis of the mechanisms that are involved in the interaction among communities. Let $dp(I)$ be the set of the dangling pages in $G_I$. The following theorem describes how the energy of each community depends on its cardinality, on the dangling pages, and on the energy that the community receives from the Web and spreads onto it.

**Theorem 3.1** Given a community $G_I$, let $f_p$ be the fraction of the hyperlinks of page $p$ that point to pages in $G_I$ with respect to the total number of hyperlinks outgoing from $p$. Let $E^{in}_I$, $E^{out}_I$, and $E^{dp}_I$ be defined by

$$E^{in}_I = \frac{d}{1 - d} \sum_{i \in in(I)} f_i x_i, \tag{3}$$

$$E^{out}_I = \frac{d}{1 - d} \sum_{i \in out(I)} (1 - f_i) x_i, \tag{4}$$

$$E^{dp}_I = \frac{d}{1 - d} \sum_{i \in dp(I)} x_i, \tag{5}$$

where $E^{in}_I$ is the energy injected into the community from the outside Web, $E^{out}_I$ is the energy spread by the community onto the rest of the Web, and $E^{dp}_I$ is the energy lost in the dangling pages. Then, the energy $E_I$ satisfies

$$E_I = |I| - E^{dp}_I + E^{in}_I - E^{out}_I. \tag{6}$$
Proof: Let $o(I) = I \setminus (\text{out}(I) \cup \text{dp}(I))$. Without loss of generality, let the components of $x$ be ordered so that $x = [x_{\text{out}(I)}, x_{\text{dp}(I)}, x_{o(I)}, x_{\text{in}(I)}]^T$. First, we consider the following system

$$Y = dQY + U,$$  \hspace{1cm} (7)

with

$$Q = \begin{pmatrix}
W_{I,\text{out}(I)} & 0 & W_{I,o(I)} & 0 \\
F & 0 & 0 & 1 \\
0 & \mathbb{I}_{|\text{dp}(I)|} & 0 & 0
\end{pmatrix},$$

$$F = [1 - f_1, \ldots, 1 - f_{|\text{out}(I)|}],$$

$$U = [U'_{I}, 0, (1 - d)(N - E_{I}^{\text{in}} + E_{I}^{\text{out}})]^T,$$

$$U_{I} = dW_{I,\text{in}(I)}x_{\text{in}(I)} + (1 - d)\mathbb{I}_{|I|}.$$  \hspace{1cm} (8)

The proof consists of three phases. In the first step, we describe the solution $Y$ of eq. (7). Then, we prove that $Q$ is a stochastic matrix. Finally, we collect both results to prove the energy balance equation.  

**Description of $Y$**

If $x$ is a solution of (2), then

$$Y = [x_{I}, E_{I}^{\text{out}} + E_{I}^{\text{dp}} + (N - E_{I}^{\text{in}} + E_{I}^{\text{out}})]$$

is a solution of (7). In fact, $E_{I}^{\text{out}} = \frac{d}{1 - d}F_{x_{\text{out}(I)}}$ and $E_{I}^{\text{dp}} = \frac{d}{1 - d}\mathbb{I}_{|\text{dp}(I)|}x_{\text{dp}(I)}$, by definition, so that

$$Y_{I} = dQ_{I,I}Y_{I} + U_{I} = x_{I},$$

$$y_{I+1} = dF_{x_{\text{out}(I)}} + \frac{d^2}{1 - d}F_{x_{\text{out}(I)}} = E_{I}^{\text{out}},$$

$$y_{I+2} = E_{I}^{\text{dp}} + (N - E_{I}^{\text{in}} + E_{I}^{\text{out}})$$

follows (see [2], for more details).

**$Q$ is a stochastic matrix**

Let $Q_{p}$ and $W_{p}$ be the $p$-th columns of $Q$ and $W$, respectively. The equality $\|Q_{p}\|_{1} = 1$ must be proved for $p \in \text{dp}(I)$, $p \in o(I)$, and $p \in \text{out}(I)$. If $p \in \text{dp}(I)$, then $\|Q_{p}\|_{1} = \|0, \ldots, 0, 1\|_{1} = 1$ by definition of $Q_{p}$. Assume $p \in o(I)$. Then, $\|W_{p}\|_{1} = 1$ holds since $p$ is not a dangling page, $W_{p} = [0, \ldots, 0, Q_{p}, 0, \ldots, 0]^T$ holds by definition, and $\|Q_{p}\|_{1} = 1$ is an immediate consequence. On the other hand, let $p \in \text{out}(I)$. Then

$$\|Q_{p}\|_{1} = \sum_{q \in \text{ch}(p) \cap I} \frac{1}{h_{p}} + 1 - f_{p} = 1.$$

**Energy balance equation**

Since $Q$ is a stochastic matrix, the solution of (7) fulfills $\|Y\|_{1} = d\|Q\|_{1}, \|Y\|_{1} + \|U\|_{1}$, and, as a consequence,

$$\|Y\|_{1} = \|U\|_{1}/(1 - d).$$  \hspace{1cm} (9)

Combining (9) with the definition of $U$,

$$\|Y\|_{1} = E_{I}^{\text{out}} + |I| + (N - E_{I}^{\text{in}} + E_{I}^{\text{out}}).$$

Finally, by the definition of $Y$,

$$\|Y\|_{1} = \|x_{I}\|_{1} + E_{I}^{\text{out}} + E_{I}^{\text{dp}} + (N - E_{I}^{\text{in}} + E_{I}^{\text{out}}),$$

so that, matching eqs. (10) and (11),

$$\|x_{I}\|_{1} = |I| + E_{I}^{\text{in}} - E_{I}^{\text{out}} - E_{I}^{\text{dp}},$$

which proves the theorem. \hfill $\blacksquare$

Theorem 3.1 suggests that communities with many pages and many references have high energy. Moreover, it also shows that dangling pages give rise to a loss of energy. The energy lost is small provided that the dangling pages have a small score (see the definition of $E_{I}^{\text{dp}}$). Theorem 3.1 finally asserts that hyperlinks which point outside the community originate a loss of energy, which is high when the hyperlinks belong to pages with high PageRank. The energy loss depends also on the fraction of all the links which point outside the community. Hence, this energy is small whenever the pages pointing outside have many hyperlinks to pages of the community (see the definition of $E_{I}^{\text{out}}$).

**Example 3.1** Let us consider the simple community of two pages represented in Fig. 1(a). It has no dangling page, no incoming hyperlink, and no outgoing hyperlink. Hence, the related energies are zero, i.e, $E_{I}^{\text{out}} = 0$, $E_{I}^{\text{in}} = 0$, $E_{I}^{\text{dp}} = 0$. From eq. (6), the energy of the community equals the number of pages, that is $E_{I} = |I|$. In fact, it can easily be verified that the PageRanks are $x_{A} = x_{B} = 1$ and $E_{I} = 2$. In Fig. 1(b), a hyperlink was removed, thus transforming page $B$ into a dangling page which, according to eq. (5), causes a loss of energy. We have $x_{A} = 1 - d$.  

![Figure 1. Some examples of communities.](image-url)
and \( x_B = 1 - d^2 \). Hence, it follows that \( E_{I_x}^{dp} = d + d^2 \) and, consequently,
\[
E_{I_x} = 2 - d - d^2,
\]
which expresses clearly the loss of energy. Moreover, if we extend the community with a hyperlink pointing to an external page, the energy becomes even smaller due to the loss \( E_{I_x}^{out} \). In fact, in Fig. 1(c), we have \( x_A = 1 - d \) and \( x_B = 1 - d + d(1 - d)/2, f_A = 1/2 \). Hence, \( E_{I_x}^{dp} = d + d^2/2, E_{I_x}^{out} = d/2 \) and, as a consequence,
\[
E_{I_x} = 2 - 3d/2 - d^2/2.
\]
On the contrary, if we extend Fig. 1(b) with a page that points to the community, as in Fig. 1(d), the energy will grow because of the presence of the term \( E_{I_x}^{in} \). In such a case, \( x_A = 1 - d + d(1 - d), x_B = 1 - d + d[(1 - d) + d(1 - d)], \) and \( x_C = 1 - d \). Then \( E_{I_x}^{in} = d, E_{I_x}^{dp} = d + d^2 + d^3 \) and
\[
E_{I_x} = 2 - d^2 - d^3.
\]
Finally, we can easily check that
\[
E_{I_x} > E_{I_x} > E_{I_x} > E_{I_x}.
\]
Eq. (4) shows that the loss of energy due to each page \( i \in \text{out}(I) \) depends also on \( f_i \). In particular, it turns out that in order to minimize \( E_{I_x}^{out} \), the hyperlinks pointing outside \( G_I \) should be in pages with a small PageRank and with many internal hyperlinks.

**Example 3.2** Let us consider the community of Fig. 1(e). It is the same as the one in Fig. 1(c), but there are two links from page \( A \) to page \( B \). The new hyperlink modifies the factor \( f_1 \), which becomes 2/3. Consequently, \( E_{I_x}^{out} \) is smaller than \( E_{I_x}^{out} \). In particular we find \( x_A = 1 - d \) and \( E_{I_x}^{out} = d/3 \).

**Entrapping energy**

The above result establishes no explicit relationship between the incoming energy and the energy entrapped in \( G_I \). The next theorem states that the energy \( E_I \) is directly proportional to the input energy \( E_{I_x}^{in} \). In particular, the proportionality factor depends upon the length of the shortest internal path \( L \) from \( p \in \text{in}(I) \) to the pages belonging to \( \text{dp}(I) \cup \text{out}(I) \).

**Theorem 3.2** Let \( \Theta_I \subseteq I \) s.t. there exists a path from \( \text{in}(I) \) to each \( q \in \Theta_I \) which is internal to \( G_I \) and has length exactly \( L \). Let \( L = \min \{l|\Theta_I \cap (\text{dp}(I) \cup \text{out}(I)) \neq \emptyset \} \). Then,
\[
E_I \geq (1 - d^{L-1}) E_{I_x}^{in}.
\]
Figure 3. Promotion of a Web community.

(b) Dangling pages give rise to a loss of energy in the community they belong to (see Fig. 2(b)). The lost energy is small provided that the dangling pages have a small score. This can be deduced from eqs. (5) and (6).

(c) Hyperlinks which point outside the community originate a loss of energy, which is high when the hyperlinks belong to pages with high PageRank. Moreover, the lost energy is small whenever the hyperlinks pointing outside are in pages with many internal hyperlinks. This can be obtained from eq. (4). For instance, in Fig. 2(c), the hyperlink departing from node “A” is preferable to the hyperlink departing from node “B”.

Detection of artificial communities

In the following, we will prove that an analogous increase in the energy of a target community can be obtained by any artificial promoting community, regardless of its connectivity pattern, with the only constraint that all its nodes have outlinks to the target community (see Fig. 3). Such a result proves that artificial promoting communities cannot be distinguished from other communities. Thus, no algorithm can be designed which guarantees the detection of all the spamming communities.

**Theorem 4.1** Let us consider two communities $C$ and $D$ such that every node of $C$ is connected to at least one of $D$, and let $F$ be defined as $F = \max_{p \in C} h_p$. Then,

$$E_D \geq d \frac{1-d}{F} |C|$$  \hspace{1cm} (12)

**Proof:** Let $L$ be defined as in Theorem 3.2. Since all the paths from the nodes in $\mathrm{in}(L)$ to a page in $I$ contains at least an arc, then $L \geq 1$, and it follows that

$$E_D \geq (1-d) E_D^{\text{in}}.$$  \hspace{1cm} (13)

Moreover, using $x_L \geq 1 - d$ and by straightforward algebra,

$$E_D^{\text{in}} = \sum_{p \in \mathrm{in}(L)} \frac{d}{1-d} f_p x_p \geq \sum_{p \in \mathrm{in}(L)} d f_p \geq d \sum_{p \in C} \frac{1}{F} = \frac{d}{F} |C|.$$  \hspace{1cm} (14)

The theorem follows putting together eqs. (13) and (14). ■

A bound on the effect of Web evolution

The above results analyze the energy migration under the assumption that the Web remains static. On the other hand, even a simple change to the Web connectivity, like the introduction or the removal of a hyperlink, may cause a modification on the whole PageRank. However, Web changes actually cause only a redistribution of the energy associated with the altered pages, such that the distance between the old and the new rank is bounded accordingly.

**Theorem 4.2** Suppose that $C$ is a set of pages where we change the outlinks and denote by $\tilde{x}$ the PageRank after the changes were carried out. Then,

$$||\tilde{x} - x||_1 \leq \frac{d}{1-d} \sum_{p \in C} \delta_p x_p \leq \frac{2d}{1-d} E_c,$$  \hspace{1cm} (15)

where $\delta_p \leq 2, \forall p$. More precisely,

$$\delta_p = \left\{ \begin{array}{ll} \left| \frac{1}{h_p} - \frac{1}{\bar{h}_p} \right| u_p + \frac{n_p}{h_p} + \frac{r_p}{\bar{h}_p}, & \text{if } h_p \neq 0, \bar{h}_p \neq 0, \\ 1, & \text{if } h_p = 0 \text{ or } \bar{h}_p = 0, \end{array} \right.$$ \hspace{1cm}

$h_p$ and $\bar{h}_p$ being the number of hyperlinks in $p$ before and after the change, respectively, $n_p$ the number of new hyperlinks, $r_p$ the number of removed hyperlinks, and $u_p$ the number of unchanged hyperlinks.

**Proof:** Let $\tilde{W}$ be the transition matrix of the changed Web and let us define $D = W - W$. Then, by eq. (2),

$$||\tilde{x} - x||_1 = ||d\tilde{W}(\tilde{x} - x) + dD||_1,$$

$$\leq xd||\tilde{W}||_1 ||\tilde{x} - x||_1 + d||Dx||_1,$$

holds. Thus, it follows that

$$||\tilde{x} - x||_1 \leq \frac{d}{1-d} ||Dx||_1 = \frac{d}{1-d} \sum_{p=1}^N D_p x_p,$$  \hspace{1cm} (16)
where $D_p$ is the $p$-th column of $D$. By definition of $D$ and straightforward algebra, we get $||D_p||_1 = \delta_p$, which, along with eq. (16), implies eq. (15).

Finally, assuming $r_p \geq n_p$ and $h_p \neq 0, \tilde{h}_p \neq 0$, we get $\delta_p = 2r_p/h_p \leq 2$. A similar analysis can be carried out if $r_p < n_p$.

Theorem 4.2 extends a similar result given in [13], where the authors prove that $||x - x||_1 \leq \frac{2}{1-d} E$. In fact, Theorem 4.2 gives a tighter bound — it replaces the factor $2/(1-d)$ with $2d/(1-d)$ — and yields a more detailed analysis — it introduces the $\delta_i$s.

Theorem 4.2 highlights a nice property of PageRank, namely that, regardless the way they change, non-authoritative communities cannot affect significantly the global PageRank.

The bound of Theorem 4.2 gives also indications on the frequency of PageRank updating. Assuming that every day the Web dynamics alters $\alpha$ pages and that the affected pages are random, then $\frac{2d}{1-d}\alpha t$ is an approximate measure of $||x_0 - x_t||_1$, the difference between the PageRanks computed at days 0 and $t$. Thus, for example, if we want that the error remains bounded by $\varepsilon$, then PageRank must be computed approximately every $\frac{1-d}{2d\alpha}\varepsilon$ days.

5 Conclusions

In this paper, we have analyzed the distribution of PageRank throughout the Web with special emphasis on its migration amongst communities. In particular, we derive a general energy balance equation which makes it possible to understand how the score migrates through different Web communities. Based on the proposed circuit equation, some simple rules to improve Web visibility are discussed, which only take into account the topological structure of the community to be promoted.

Acknowledgments

We thank Michelangelo Diligenti, Marco Mazzini, Edmondo Trentin (University of Siena), Paolo Frasconi (University of Florence), Ah Chung Tsoi (University of Wollongong), Prabhakar Raghavan (Verity Inc.), and Klaus Meer (University of Odense) for fruitful discussions which contributed to shed lights on some topics covered in this paper.


References


