PageRank
A Circuital Analysis

Monica Bianchini, Marco Gori, Franco Scarselli
Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Siena
Via Roma, 56 — 53100 Siena, ITALY
e–mail: {monica,marco,franco}@ing.unisi.it

Abstract

PageRank is a topological measure of the authority of Web pages which is adopted by Google search engine to sort all documents matching a given query. In this paper, we provide a circuital analysis which allows us to understand how the connectivity of the Web affects PageRank. Based on the given results, we formulate some rules which turn out to be very useful to construct pages with high PageRank. Finally, we prove that PageRank exhibits a nice robustness property, in the sense that communities with a small authority cannot change significantly the PageRank of other communities.

Keywords: PageRank, Google search engine, Web page scoring systems, Web visibility

1 Introduction

Google search engine combines the relevance of the document content with PageRank [1], which is a measure of the authority of a page simply based on the Web connectivity. The ideas behind PageRank can be traced back to the notion of citation in 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 by a forward link. Moreover, selective citations from q to p provide more contribution to the score of p than spread citations. In [1], PageRank is formally defined by

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

where $d \in (0,1)$ is a *dumping factor*, h_q is the number of hyperlinks outcoming from q, and pa$[p]$ is the set of pages linked to p. Eq. (1) is slightly different in [2], but the corresponding versions of PageRank are tightly related [3]. The PageRank algorithm evaluates eq. (1) using the Jacobi algorithm to solve linear systems [2, 3].

In this paper, we study how the connectivity of the Web affects the distribution of PageRank. Our goal is not to provide a direct method to acquire visibility on Google. In fact, Google's scoring system is a complex mechanism using a query relevance measure and many heuristics along with PageRank. On the other hand, our analysis of PageRank will allow us to disclose interesting properties of the method and to provide the basis for extensions and improvements.

2 Basic Results

In order to carry out our analysis, let us define the concept of **community** and the associated notion of **energy**. A community $\boldsymbol{G}_I = (I, W_I)$ is a subgraph of the Web which is defined by a set of pages I and by all the hyperlinks W_I between the pages in I. The energy E_I of $\boldsymbol{G}_I$ is the sum $E_I = \sum_{i \in I} x_i$ of the PageRanks of all its pages. A community is connected to the rest of the Web by a set of internal pages out(I) that point to other communities and by a set of external pages in(I) that point to $\boldsymbol{G}_I$. In our analysis, an important role is also played by the set sink(I), which collects the pages that do not contain hyperlinks (e.g. full text documents). Finally, a community isolated from the rest of the Web is an **island**. Due to space limitations, the proofs of the theoretical results shown in the following are collected in [3].

2.1 Communities and their interactions

Intuitively, a community can be any set of pages on a given topic, the set of pages of a domain or a Web site. The energy is a measure of the authority of the community. A detailed analysis points out that the energy is due to the interaction among communities established by the hyperlinks. The theory that describes the energy flow through the communities resembles theory of digital circuits. The following theorem shows that energy E_I depends on four components.

Theorem 2.1 *The energy E_I of a community $\boldsymbol{G}_I$ satisfies*

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{sink}. \tag{2}$$

Here, $|I|$ denotes the number of pages in $\boldsymbol{G}_I$ and represents the default energy assigned to the community. The component E_I^{in} is the energy that comes from the other communities pointing to $\boldsymbol{G}_I$. The presence of E_I^{in} in eq. (2) is coherent with the fact that communities with many references have a high authority. The term E_I^{out} is the energy that the community spread to the Web. In fact, the presence of hyperlinks from $\boldsymbol{G}_I$ to outside $\boldsymbol{G}_I$ leads to decrease its energy. Finally, also sinks yields a loss of energy, which can be calculated by E_I^{sink}. More precisely, it can be proved [3] that

$$\begin{aligned} E_I^{in} &= \frac{d}{1-d} \sum_{i \in \text{in}(I)} f_i x_i, & (3) \\ E_I^{out} &= \frac{d}{1-d} \sum_{i \in \text{out}(I)} (1-f_i) x_i, & (4) \\ E_I^{sink} &= \frac{d}{1-d} \sum_{i \in \text{sink}(I)} x_i, & (5) \end{aligned}$$

where f_i is the fraction of the hyperlinks of page i that point to pages in $\boldsymbol{G}_I$ w.r.t. the total number of hyperlinks outgoing from i.

Notice that 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^{out}, the external hyperlinks (those pointing to out of $\boldsymbol{G}_I$) should be in pages with a small PageRank and that have many internal hyperlinks. A similar reasoning works also for sinks. The above discussion provides the following guidelines for the design of Web communities with a high energy.

1. *The same content divided into many small pages yields a higher score than if it is concentrated into a single large page.*

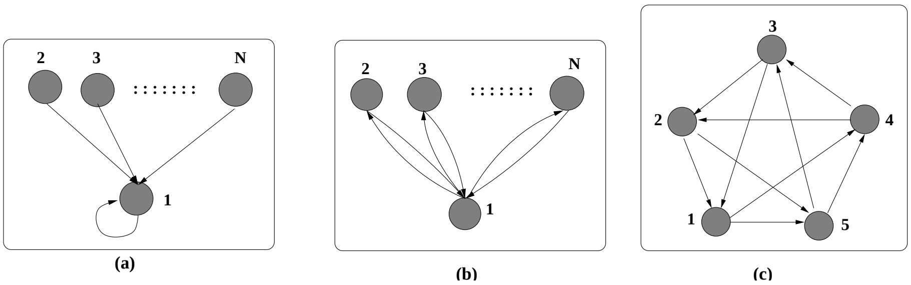

Figure 1: Two islands with maximum PageRank and a regular graph.

2. *Sinks should be avoided or carefully limited.*

3. *External hyperlinks must be limited.*

4. *External hyperlinks should belong to pages with many internal hyperlinks and/or with small PageRank.*

5. *Pages that point to sinks should have a small score and/or many internal hyperlinks.*

2.2 PageRank and degree of regularity

An interesting question related to the above discussion, but not answered by Theorem 2.1, is the following one: How can the energy be concentrated into a single page? Said another words, which is the island that is constituted of exactly N documents and contains the page with the largest PageRank? In fact, it is reasonable that a Web designer wants to concentrate PageRank into one or few important pages (e.g. homepages), instead of spreading the energy over all the documents. Formally, the answer to this question is different according whether self hyperlinks were admitted or not. According to [3], Fig. 2.2(a) and Fig. 2.2(b) show the two islands with the largest PageRanks. Moreover, let us consider a regular graph (of degree k), i.e. a graph where each node has exactly k incoming and k outgoing hyperlinks (see Fig. 2.2(c)). If a community is an island and its connectivity is represented by a regular graph, then it is easy to verify that all the PageRanks are 1. On the other hand, notice that the island of Fig. 2.2(a) gets the "maximal irregularity": the difference between the hyperlinks received by page 1 and the other pages is maximal. Thus, practically speaking, the irregularity of the graph can be a trick by which a Web designer boosts the PageRank of certain pages.

2.3 PageRank is robust

Even if an appropriate organization of a community can avoid energy loss, however, small communities with few references cannot have pages with large scores. Such a claim is straightforwardly confirmed by Theorem 2.1, which proves that

$$E_I \leq |I| + E_I^{in}. \tag{6}$$

Moreover, also the effect that a community can produce onto the rest of the Web is bounded. In fact, the following theorem describes what happens when a community alters its pages.

Theorem 2.2 *Let C be a set of pages whose content changes in time. Then,*

$$\sum_{p \in W} |x_p(t+1) - x_p(t)| \leq \frac{2d}{1-d} E_C(t), \tag{7}$$

where W denotes the whole Web.

Theorem 2.2 actually extends a similar result in [4]. It points out that the change in PageRank is proportional to the score of the modified pages. Thus, global PageRank can change linearly at most (by a factor $2d/(1-d)$) w.r.t. the energy of the community that contains the altered pages. Finally, we can conclude that PageRank is robust, since non–authoritative communities cannot influence significantly the global PageRank.

References

[1] S. Brin and L. Page, "The anatomy of a large–scale hypertextual Web search engine," in *Proceedings of WWW7*, 14–18 Apr. 1998.

[2] S. Brin, L. Page, R. Motwani, and T. Winograd, "The page rank citation ranking: Bringing order to the Web," Tech. Rep. 1999–66, Stanford Digital Libraries Working Paper, 1999. Available at `http://dbpubs.stanford.edu:8090/pub/1999-66`.

[3] M. Bianchini, M. Gori, and F. Scarselli, "Inside Google Web page scoring system," Tech. Rep. DII /2001, Dipartimento di Ingegneria dell'Informazione, Università di Siena, 2001.

[4] A. Y. Ng, A. X. Zheng, and M. I. Jordan, "Link analysis, eigenvectors and stability," in *Proceedings of SIGIR 2001*, ACM Press, 2001.