*Vietnam Journal of Mathematics* **37**:4(2009) 475-489

Vietnam Journal
of
MATHEMATICS


# MPageRank: The Stability of Web Graph

**Le Trung Kien**[1], **Le Trung Hieu**[2], **Tran Loc Hung**[1], **and Le Anh Vu**[3]

[1] *Department of Mathematics, College of Sciences, Hue University, Vietnam*

[2] *Mathematics & Mechanics Faculty, Saint-Petersburg State University, Russia*

[3] *Department of Computer Science, ELTE University, Hungary*

Received November 04, 2008

**Abstract.** Problems which have a huge database always issue challenges to scientists. Evaluating importance of Webs is an interesting example. This problem should be very difficult about not only computation, but also storage since the Web environment contains around 10 billions Web pages. Basing on the "random surfer" idea of PageRank algorithm, MPageRank greatly improves results of Web search by applying a probabilistic model on the link structure of Webs to evaluate "authority" of Webs. Unlike PageRank, in MPageRank, a Web now has different ranking scores which depend on the given multi topics. By assigning a value characterizing a relationship between content of pages and a popular topic, we would like to introduce some new notions such as *the influence of page* and *the stability of rank score vector* to evaluate the stability of Web environment. However, the main idea of establishing the MPageRank model is to partition our Web graph into smaller-size Web subgraphs. As a consequence of evaluation and rejection about pages influence weakly to other pages, the rank score of pages of the original Web graph can be approximated from the rank score of pages in the new partition Web graph.

2000 Mathematics Subject Classification: 68W40, 05C75.

*Key words:* MPageRank algorithm, web graph, ranking web, probabilistic model.

## 1. Introduction

With an extraordinary grow rate of quantity and a heterogeneity of quality, World Wide Web creates a lot of new challenges for information retrieval. One

of the interesting challenges is evaluating the importance of a Web. Search engines have to choose from a huge number of the Web pages, which contain the information specified by the user, the "most important" ones, and bring them to the user.

The PageRank algorithm [13] used in the Google search engine is the most famous and effective one in practice. The link structure of the Web graph is an abundant source of information about the authority of Webs. It encodes a considerable amount of latent human judgment, and we claim that this type of judgment is necessary to formulate a notion of authority. The underlying idea of PageRank is to use the stationary distribution of a *random surfer* on the Web graph in order to assign relating ranks to the pages. PageRank scores act as overall authority values of pages which are independent on any topic. However, in practice, a user himself often has a proposed topic when he retrieves information from the internet. In fact, at first, the surfer seems to visit from the pages of which content are related to his proposed topic and while surfing from page to page following outlinks, he always give priority to surf these pages. This property is not considered in PageRank because its random surfer surfed indefinitely from page to page following all outlinks with equal probability.

From the above observations, we introduce and describe about the MPageRank algorithm. Assuming that we can find a finite collection of the most popular topics (music, sport, news, health, etc). For each topic, we can evaluate the correlation between Webs and the topic by scanning their text. Each node of the Web graph now is weighted and this weight is determined by the given popular topic. The probabilistic model in the MPageRank does not behave uniform for all outlinks and nodes, it is improved by supporting the weight of web nodes. The user can choose his proposed topic from the collection of given topics, and the chosen rank score is suitable to this topic. Certainly, the probabilistic model in MPageRank not only enables the user to choose his preferable topic but also models surf-Web process more precisely than the PageRank's. However, the main aim at building new MPageRank model is to weight the Web graph. So thanks to this, we study more the theory of partition Web graph effectively. As we know, if our Web graph is partitioned into subgraphs which do not connect together, the calculation in algorithms will be reduced remarkably. Deepening this opinion, in Sec. 3, we bring out the definition of the *influence* of Web page, which evaluates the influence rate of one page to other pages. Some results such as Theorem 3.3 and Corollary 3.6 in this section combining with the notion of *stability of rank score vector* and Proposition 4.2 in Sec. 4 have proved the effectiveness of MPageRank algorithm.

Recently, there have been many approaches surmounting the probability score of page ignoring topic corresponding to the query. Richardson and Domingos [14] have proposed the other probabilistic model, an intelligent random surfer, which approached for rank score function by generating a PageRank vector for each possible query term. Haveliwala [8] has approached by using categories "topic-sensitive" in Open Directory to bias importance scores, where the vectors and weights were selected according to the text query without the user's choice.

To speed up the computation of PageRank, Kamvar, Haveliwala, et al. [9, 10] have used successive intermediate iterates to extrapolate successively better estimates of the true local PageRank scores for each *host* which are computed independently using the link structure of that host. Then these local rank scores are weighted by the "importance" of the corresponding host, and the standard PageRank algorithm is then run using as its starting vector the weighted concatenation of the local rank score. This idea originated from exploiting a nested block structure of the Web graph.

What is the model Web graph? How does it grow randomly? These interesting questions help us realize Web environment from other ways. The complex network systems have been modeled as *random graphs*, it is increasingly recognized that the topology and evolution of real networks are governed by robust organizing principles. The basic knowledge of *random graphs* can find in [3]. Basing on model random graphs, Albert and Barabási [1] have discovered the small-world property and the clustering coefficient of World Wide Web. Especially, they have discovered that the degree distribution of the web pages follows a power law over several orders of magnitude. Callaway et al.[4] have introduced and analyzed a simple model of a growing network, *randomly grown graphs* that many of its properties are exactly solvable, yet it shows a number of non-trivial behaviors. The model demonstrates that even in the absence of preferential attachment, the fact that a Web environment is grown, rather than created as a complete entity, leaves an easily identifiable signature in the environment topology.

There have been many papers [5, 6, 7, 12] investigating the property of partition Web graph; most results have theoretical character. Kleinberg [12] has introduced the notion $(\epsilon, k)$-*detection set* play a role as the evidence for existence of sets which does not have as most $k$ elements (nodes or edges) and have the property: if an adversary destroys this set, after which two subsets of the nodes, each at least an $\epsilon$ fraction of the Web graph, that are disconnected from one another. Fakcharoenphol [7] has showed that the $(\epsilon, k)$-detection set for node failures can be found with probability at least $1-\delta$ by randomly chossing a subset of nodes of size $O(\frac{1}{\epsilon}k \log k \log \frac{k}{\epsilon} + \frac{1}{\epsilon} \log \frac{1}{\delta})$. Chung [5, 6] has studied partition property of a graph based on applications of eigenvalues and eigenvectors of graphs in combinatorial optimization. Basically, our new theoretical results in this paper originate from the direction of Chung's research.

The remainder of the paper is organized as follows: Sec. 2 is the MPageRank introducing the probabilistic model of MPageRank. In Sec. 3, we bring out the notion *influence of Web page* and some related results evaluating the influence of one page to the order of authority of others. The concept *stability of rank score vector* is introduced in Sec. 4 having the deeper appreciation of the influence notion. Finally, Sec. 5 will be a summary about the MPageRank algorithm and a conclusion.

## 2. The MPageRank

### *2.1. Probabilistic Model of MPageRank*

MPageRank is the algorithm that evaluates the authority of Web pages based on the link structure depending on the query's topic. Given a topic $T$, link structure can be modelled by a weighted directed graph, *Web graph*. Formally, we denote the Web graph as $G = (V, E)$, where the nodes, $V$, corresponding to the pages, and a directed edge $(u, v) \in E$ indicates the presence of a link from page $u$ to page $v$ $(u, v \in V)$. We call $N = |V|$ the number of nodes in Web graph, and for all page $u \in V$, $F_u$ be the set of pages $u$ points to, $B_u$ be the set of pages that point to $u$. For pages which have no outlinks we add a link to all pages in the graph[1]. In this way, rank which is lost due to pages with no outlinks is redistributed uniformly to all pages. The *rank score vector* $r : V \to [0, 1]$ denotes the rank score of pages, $r(u)$ is the score of page $u$. MPageRank builds the rank score vector based on three following assumptions:

- The web pages, which are linked by many others pages, have a high score. In literature, we evaluate the authority of pages from "the crowd". A web page is considered "high quality" if the crowd accepts to it.
- If a high score page links to some pages, its destination have a high score too. For example, a page just has only one link from Yahoo!, but it may be ranked higher than many pages with more links from obscure places.
- A page having its content related to the topic will have a high score. Clearly, Hue university's homepage should be ranked higher than Yahoo! with the purpose query "Hue university".

In the third assumption, how to evaluate the relating rate of a Web page with a given topic based on its content? This is a big and complex problem which attracts the attention of scientists in two recent decades. As we know, this problem is known with the name *Text Analysis*, which contains some techniques for analyzing the textual content of individual Web pages. Recently, the publisher John Wiley & Sons has published the book [2] which has one chapter to present this problem. The techniques presented in this book have been developed within the fields of *information retrieval* and *machine learning* and include indexing, scoring, and categorization of textual documents. Concretely, the main problem to evaluate the relating rate of Web's content with a given topic is that whether we can classify Web pages or not based on their content. Clearly, this technique is related to information retrieval technique, that consists of assigning a document of Web to one or more predefined categories.

In this paper, we have no intention of researching on the above problem thoroughly; however, in order to create theoretical base for results in the next section of the paper, we accept a judgement that: "Given a topic $T$, we can have an *evaluation function* $f_T$ to evaluate how relationship between a page and this topic is". After constructing the evaluation function $f_T$ for the topic $T$, the node $u$ of Web graph is weighted by the value of function $f_T(u)$ of the corresponding of page $u$. The Web graph will be a weighted directed graph now.

[1] For each page $s$ with no outlinks, we set $F_s = V$ be all N nodes, and for all other nodes augment $B_u$ with s, $(B_u \cup \{s\})$.

We choose the rank score vector as a standing probability distribution of a random walk on the Web graph. Intuitively, this can be thought as a result of the behavior model of a "random surfer". The "random surfer" simply keeps clicking on successive links at random, and gives priority to pages relevant to the given topic. However, if a real Web surfer ever gets into a small loop of web pages, it is unlikely that the surfer will be in the loop forever. Instead, the surfer will jump into some other pages. Formally, if the surfer is at Web $u$ at present, he will do two following actions in the next step:

1) Generally, with probability $1 - p$, the surfer surfs following all outlinks, where surfs to page $v$ ($v \in B_u$) with probability $p_{uv}$.

2) When the surfer feels bored, with probability $p$, he jumps to all pages in Web graph, where page $v$ is probability $p_v$.

$p$ is called *jump probability* ($0 < p < 1$). In practice, we choose $p = 0.1$. Through the above surfing formula, considering the relation to the topic of page, $p_{uv}$ is recognized as an importance of page $v$ among the pages which $u$ links to, and $p_v$ is considered as an importance page $v$ within the whole World Wide Web. Of course, $p_{uv}$ and $p_v$ are identified as:

$$p_{uv} = \frac{f_T(v)}{\sum\limits_{j \in F_u} f_T(j)} = \frac{f_T(v)}{f_T(F_u)}; \qquad p_v = \frac{f_T(v)}{\sum\limits_{j \in V} f_T(j)} = \frac{f_T(v)}{f_T(V)},$$

where we denote $f_T(A) = \sum\limits_{j \in A} f_T(j)$, for any set $A$.

*2.2. Rank Score Vector in MPageRank*

If let $r_i$ be the probabilistic distribution of random surfer visiting Web pages in Web graph at step $i$, basing on above surfing model, the probability of event that the surfer is on page $v$ at step $i$ is given by the formula

$$r_i(v) = pp_v + (1 - p) \sum_{u \in B_v} p_{uv} r_{i-1}(u).$$

Let $R = pR_1 + (1-p)R_2$, where $R_1, R_2$ are $N \times N$ matrices with $R_1(uv) = p_v$ and

$$R_2(uv) = \begin{cases} p_{uv} & \text{if } (u, v) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Matrix $R$ is the transition probability matrix of surfer when he surfs on the Web graph in probabilistic model of MPageRank. Rank score vector in MPageRank at step $i$ is given by the formula

$$r_i = R^T r_{i-1}.$$

Certainly, $(r_i)_{\mathbb{N}}$ is a Markov chain with the state space $V$. We have a result

**Proposition 2.1.** *Markov chain $(r_i)_{\mathbb{N}}$ has a unique stationary probability distribution, denoted by $r$.*

*Proof.* If Markov chain $(r_i)_{\mathbb{N}}$ has only one irreducible closed subset $S$, and if $S$ is aperiodic, the chain must have a unique stationary probability distribution. So we simply have to show that the Markov chain $(r_i)_{\mathbb{N}}$ has a single irreducible closed subset $S$, and that this subset is aperiodic.

Let the set $U$ be the pages whose evaluation function $f_T$ is nonzero. Concretely, $U = \{v \in V \mid f_T(v) \neq 0\}$. Let $S$ consist of the set of all pages reachable from $U$ along nonzero transition in the chain. $S$ trivially forms a closed subset. Further, since every page has a transition to $U$, no subset of $S$ can be closed. Therefore, $S$ forms an irreducible closed subset. Moreover, every closed subset must contain $U$, and every closed subset containing $U$ must contain $S$. So $S$ must be the unique irreducible closed subset of the chain.

On the other hand, all members in an irreducible closed subset have the same period, so if at least one page in $S$ has a self-transition, then the subset $S$ is aperiodic. Let $u$ be any page in $U$. By construction, there exists a self-transition from $u$ to itself. Therefore $S$ must be aperiodic, so the Markov chain $(r_i)_{\mathbb{N}}$ has a unique stationary probability distribution $r$. ■

This stationary distribution $r$, itself is the rank score vector in MPageRank. Rank score vector in MPageRank is given by formula

$$r = R^T r. \tag{1}$$

$R^T$ is the stochastic matrix so rank score vector $r$ is equivalent to *primary eigenvector* of the transition matrix $R$ corresponding with *eigenvalue* 1. Thus, in the MPageRank model, the authority of page $u$, $r(u)$, is recognized as the probability that the random surfer will surf page $u$ when he surfs Web during his searching topic $T$.

## 3. The Influence of Web Page

Evidently, in the problem of searching Web pages related a query's topic, *the order of authority* of pages is the most thing which this problem needs to be solved. From this point of view, *an influence* of Web page on World Wide Web is understood as the change of order of authority among Web pages when we change some properties of that page. In this paper, we consider that the order of authority of Web pages is typical to the rank score vector MPageRank $r$, and the property of any Web page which can be changed is the rejectable itself and its conjugate edges in Web graph. Within this idea, we have some results evaluating *the influence of Web page* in two next sections. From this interesting results, we can see that World Wide Web environment is really stable. In other words, if we reject the weak page (the page has small MPageRank value) and its conjugate edges, the order of authority of other pages will be an inappreciable difference.

### *3.1. Related Notions*

Let be given a popular topic $T$, we have a weighted directed graph $G = (V, E)$ with a transition probability matrix $R$ in MPageRank algorithm. For any $t \in V$, set $G' = G \backslash t$, a graph $(V', E')$ where $V' = V \backslash \{t\}$ and $E' = \{uv \mid u, v \in V',\ uv \in E\}$. Let $R^{'}$ be a transition probability matrix corresponding to a random surfer in the new Web graph $G'$. The new random surfer will have a stationary distribution, denoted by $r'$. We have an interesting judgement that the random surfer on the graph $G'$ with MPageRank transition probability matrix $R^{'}$ is equivalent to another random surfer on the graph $G$ with MPageRank transition probability matrix $R^*$ when the evaluation function value $f_T(t) = 0$. Let $r^*$ be a stationary distribution of random surfer on the graph $G$ corresponding to the transition probability matrix $R^*$, we called $r^*$ *an expand MPageRank rank score vector* of Web graph $G^{'}$. From this notion, a definition of *the influence* of $t$ is given as follows.

**Definition 3.1.** The influence of Web page $t$ is the difference between the expand MPageRank rank score vector $r^*$ and the MPageRank rank score vector $r$,

$$\mathcal{I}(t) = \frac{1}{N}\|r^* - r\| = \frac{\|\Delta r\|}{N}.$$

In order to identify $\|\Delta r\|$, we consider the differential matrix $\Delta R = R^* - R$. We easily see that

$$R(uv) = \begin{cases} p\frac{f_T(v)}{f_T(V)} & \text{if } uv \notin E, \\ p\frac{f_T(v)}{f_T(V)} + (1-p)\frac{f_T(v)}{f_T(F_u)} & \text{if } uv \in E \end{cases}$$

and

$$R^*(uv) = \begin{cases} 0 & \text{if } v = t, \\ p\frac{f_T(v)}{f_T(V)-f_T(t)} & \text{if } v \neq t\ ,\ uv \notin E, \\ p\frac{f_T(v)}{f_T(V)-f_T(t)} + (1-p)\frac{f_T(v)}{f_T(F_u)} & \text{if } v \neq t\ ,\ uv \in E\ ,\ t \notin F_u, \\ p\frac{f_T(v)}{f_T(V)-f_T(t)} + (1-p)\frac{f_T(v)}{f_T(F_u)-f_T(t)} & \text{if } v \neq t\ ,\ uv \in E\ ,\ t \in F_u \end{cases}$$

so, we have

$$\Delta R(uv) = \begin{cases} -p\frac{f_T(v)}{f_T(V)} & \text{if } v = t\ ,\ u \notin B_t, \\ -p\frac{f_T(v)}{f_T(V)} - (1-p)\frac{f_T(v)}{f_T(F_u)} & \text{if } v = t\ ,\ u \in B_t, \\ p\frac{f_T(t)}{f_T(V)}\frac{f_T(v)}{f_T(V)-f_T(t)} & \text{if } v \neq t\ ,\ u \notin B_t \cap B_v, \\ p\frac{f_T(t)}{f_T(V)}\frac{f_T(v)}{f_T(V)-f_T(t)} + (1-p)\frac{f_T(t)}{f_T(F_u)}\frac{f_T(v)}{f_T(F_u)-f_T(t)} & \text{if } v \neq t\ ,\ u \in B_t \cap B_v. \end{cases}$$

Let $G$ be a Web graph and a MPageRank random surfer surfs on it. We have a transition probability matrix $R$. If $R$ has a stationary distribution $r$, then set a matrix

$$\mathcal{L} = \mathbf{I} - \frac{D^{1/2}RD^{-1/2} + D^{-1/2}R^T D^{1/2}}{2},$$

where $D$ is a diagonal matrix with entries $D(v,v) = r(v)$. $\mathcal{L}$ is called *a Laplacian matrix* of a directed Web graph $G$. Clearly, the Laplacian is a real symmetric matrix, so its has $N = |V(G)|$ real eigenvalues $\lambda_0 \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_{N-1}$ (repeated according to their multiplicities). We define $\lambda = \min_{i\neq 0} |\lambda_i|$ to be *an algebraic connectivity* of Web graph $G$, so we have two important results in the next section.

*3.2. The Evaluating Results*

**Theorem 3.2.** *We have*

$$\left|[\Delta R^T.\ r](v)\right| \leqslant r(t),\ \forall v \in V.$$

*Proof.* We consider two cases

- If $v = t$

$$\begin{aligned} \Delta R(ut) = -R(ut) \quad &\Rightarrow \quad \left[\Delta R^T.r\right](t) = -\sum_{u\in V} R(ut)r(u) = -r(t) \\ &\Rightarrow \quad \left|\left[\Delta R^T.r\right](t)\right| = r(t). \end{aligned}$$

- If $v \neq t$

$$\begin{aligned} &\left[\Delta R^T.r\right](v) = \sum_{u\in V} \Delta R(uv)r(u) \\ &= \sum_{u\in B_t\cap B_v} \left[p\frac{f_T(v)}{f_T(V)-f_T(t)}\frac{f_T(t)}{f_T(V)} + (1-p)\frac{f_T(v)}{f_T(F_u)-f_T(t)}\frac{f_T(t)}{f_T(F_u)}\right] r(u) \\ &\quad + \sum_{u\notin B_t\cap B_v} p\frac{f_T(v)}{f_T(V)-f_T(t)}\frac{f_T(t)}{f_T(V)}r(u). \end{aligned}$$

We easily see that $\frac{f_T(v)}{f_T(V)-f_T(t)} \leqslant 1$ and $\frac{f_T(v)}{f_T(F_u)-f_T(t)} \leqslant 1$, so we have

$$\begin{aligned} &\left[\Delta R^T.r\right](v) \leqslant \sum_{u\in B_t\cap B_v} \left[p\frac{f_T(t)}{f_T(V)} + (1-p)\frac{f_T(t)}{f_T(F_u)}\right] r(u) + \sum_{u\notin B_t\cap B_v} p\frac{f_T(t)}{f_T(V)}r(u) \\ &\leqslant \sum_{u\in B_t} \left[p\frac{f_T(t)}{f_T(V)} + (1-p)\frac{f_T(t)}{f_T(F_u)}\right] r(u) + \sum_{u\notin B_t} p\frac{f_T(t)}{f_T(V)}r(u) \\ &= \sum_{u\in V} R(ut)r(u) = r(t). \end{aligned}$$

Thus, $\left|\left[\Delta R^T.r\right](v)\right| \leqslant r(t)$, for all $v \in V$. ∎

Theorem 3.2 shows that there is a little effect of the *differential matrix* $\Delta R$ on the rank score vector $r$, depending on the value MPageRank $r(t)$ of page $t$. Applying Theorem 3.2, we can infer a result reflected directly the value of influence of $t$.

**Theorem 3.3.** *For any tiny real number $\epsilon > 0$, and a weak page $t$ with $r(t) \leqslant \epsilon$, the influence of page $t$, $\mathcal{I}(t)$, is bounded as follows*

$$\mathcal{I}(t) \leqslant \frac{1}{N}\left(\frac{2r(t)}{\lambda}\right)^{1/2} \leqslant \frac{1}{N}\left(\frac{2\epsilon}{\lambda}\right)^{1/2}.$$

*Proof.* We have

$$\begin{aligned}
& r^* = R^{*T}.r^* \\
\Rightarrow \quad & r^* = R^T.r + R^T.\Delta r + \Delta R^T.r + \Delta R^T.\Delta r \\
\Rightarrow \quad & [\mathbf{I}_N - R^T - \Delta R^T]\Delta r = \Delta R^T.r \\
\Rightarrow \quad & \Delta r^T[\mathbf{I}_N - R^*] = r^T.\Delta R \\
\Rightarrow \quad & \Delta r^T[\mathbf{I}_N - R^*]\Delta r = r^T.\Delta R.\Delta r.
\end{aligned}$$

From Theorem 3.2 and $\sum_u r(u) = \sum_u r^*(u) = 1$, we have $\left|r^T.\Delta R.\Delta r\right| \leqslant 2r(t)$. To prove

$$\|\Delta r\|^2 \leqslant \frac{2r(t)}{\lambda},$$

we consider the lemma:

**Lemma 3.4.** [11] *Let be given a stochastic matrix $R$ with order $n$; a vector $d$ with the same order $n$ and satisfying $\sum d_i^2 = 1$, a diagonal matrix $D$, where $D_{ii} = d_i > 0$. Then we have*

$$\begin{aligned}
\min_{\substack{xe=0 \\ \|x\|=1}} \left\{\left|x^T(\mathbf{I}_n - R)x\right|\right\} &= \min_{\substack{xd=0 \\ \|x\|=1}} \left\{\left|x^T(\mathbf{I}_n - DRD^{-1})x\right|\right\} \\
&= \min_{\substack{xd=0 \\ \|x\|=1}} \left\{x^T(\mathbf{I}_n - \frac{DRD^{-1} + (DRD^{-1})^T}{2})x\right\}.
\end{aligned}$$

Lemma 3.4 is correctly proven based on the basic knownledge of eigenvector. From Lemma 3.4, for a case with $d = r^{\frac{1}{2}}$ $(d(v) = r^{\frac{1}{2}}(v))$, we have

$$\begin{aligned}
\min_{xe=0, x\neq 0}\left\{\frac{\left|x^T(\mathbf{I}_{N-1} - R')x\right|}{\|x\|^2}\right\} &= \min_{xd=0, x\neq 0}\left\{\frac{\left|x^T(\mathbf{I}_{N-1} - D^{\frac{1}{2}}R'D^{-\frac{1}{2}})x\right|}{\|x\|^2}\right\} \\
&= \min_{xd=0, x\neq 0}\left\{\frac{x^T\mathcal{L}x}{\|x\|^2}\right\} = \lambda.
\end{aligned}$$

So if $\Delta^{'}r$ is $(N-1)$-vector which produced from vector $\Delta r$ by rejecting page $t$, then $\sum_u \Delta^{'}r(u) = 0$ (vector $\Delta^{'}r$ is orthogonal to $e = (1, \ldots, 1)^T$).

Therefore we have

$$\begin{aligned} & \left|\Delta r^T[\mathbf{I}_N - R^*]\Delta r\right| = \left|\Delta^{'}r^T[\mathbf{I}_{N-1} - R^{'}]\Delta^{'}r\right| \geqslant \lambda\|\Delta^{'}r\|^2 \\ \Rightarrow \quad & \lambda\|\Delta r\|^2 = \lambda\|\Delta^{'}r\|^2 \leqslant 2r(t) \\ \Rightarrow \quad & \|\Delta r\|^2 \leqslant \frac{2r(t)}{\lambda} \leqslant \frac{2\epsilon}{\lambda}. \end{aligned}$$

Theorem 3.3 is proven. ■

As we know, the value $\lambda$ is called an algebraic connectivity of Web graph $G$ according to the transition probability matrix $R$. In the paper [11], we have a result to bound the value $\lambda$ as follows.

**Proposition 3.5.** [11] *If $\lambda$ is an expand algebraic connectivity of $G$, then we have*

$$\lambda \geqslant \frac{p^2}{8}.$$

As a direct consequence of Theorem 3.3 and Proposition 3.5, we have an important result:

**Corollary 3.6.** *For a tiny real number $\epsilon > 0$, and a page $t$, $r(t) \leqslant \epsilon$, the influence of page $t$, $\mathcal{I}(t)$, is bounded as follows*

$$\mathcal{I}(t) \leqslant \frac{1}{N}\left(\frac{16r(t)}{p^2}\right)^{1/2} \leqslant \frac{4\epsilon^{\frac{1}{2}}}{pN}.$$

## 4. The Stability of Rank Score Vector

Concept of *stability*, opposite of *chaos*, is identified as differential meanings due to purposes. However, on the whole, the nature of this concept of stability or chaos comes from evaluating the influence of separate individual on a system. When one or a lot of properties of the individual may change, does a disordered situation of the system happen?

A natural question whether the World Wide Web is stable or chaotic is posed immediately. With an extraordinary grown rate of new Webs and their heterogeneous contents, beside the unpredictable decrease of bad Webs, we can point out that World Wide Web is too chaotic. However, intuitively, basing on Theorem 3.3 and Proposition 3.3, World Wide Web is predicted to be quite stable. In order to make this clear, we keep on developing this problem in this section.

A real vector $x = (x_1, x_2, \ldots, x_n)$ is called an *ordered rank score vector*, if it satisfies:

$$\begin{cases} 0 \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n \leqslant 1, \\ x_1 + x_2 + \cdots + x_n = 1. \end{cases}$$

Having an ordered rank score vector $x = (x_1)_n$, we get a new concept:

**Definition 4.1.** The real number

$$\mu(x) = \frac{1}{n-1}\Big[(x_2 - x_1) + (x_3 - x_2) + \cdots + (x_n - x_{n-1})\Big] = \frac{1}{n-1}(x_n - x_1)$$

is called the stability of ordered rank score vector $x$.

Clearly, $\mu(x)$ is a mean value of difference among the elements value in $x$. If $\mu(x)$ is large, the stability of element's order of rank score vector $x$ will be high. Vice versa, if $\mu(x)$ is small, the difference between two of any elements of $x$ is low, the tiny change of element's value will disorder the order of them. It means the order of rank score vector $x$ will be changed.

We have a question that what is an interval of value in which the stability of ordered of rank score vector $\mu(x)$ will be? How's level of interval which we claim this ordered rank score vector is stable? And how's level of interval which we claim chaotic? For example, when $n = 2$, an ordered rank score vector $x = (x_1, x_2)$ satisfies $0 \leqslant x_1 \leqslant x_2 \leqslant 1$ and $x_1 + x_2 = 1$. So, we easily see that $\mu(x) = x_2 - x_1$ is in $[0, 1]$ and its expectation $\mathbb{E}(\mu(x)) = \frac{1}{2}$. In this case, we can judge on one's own intiative that: If $\mu(x) > \frac{3}{4}$, the ordered rank score vector $x$ is stable, and if $\mu(x) < \frac{1}{4}$, we claim that $x$ is chaotic. To solve the general problem $(n > 2)$, we use some techniques as follows:

Let a sequence of variables having the uniform distribution $X_1, X_2, \ldots, X_n$ satisfy:

$$\begin{cases} 0 \leqslant X_1 \leqslant X_2 \leqslant \cdots \leqslant X_n \leqslant 1, \\ X_1 + X_2 + \cdots + X_n = 1. \end{cases}$$

If we think $X_i$ as the variable taking the $x_i$'s value in the ordered rank score vector $x$, so

$$\mathbb{E}(Y_n) = \frac{1}{n-1}\mathbb{E}(X_n - X_1) = \frac{1}{n-1}\left[1 - \mathbb{E}(2X_1 + X_2 + \cdots + X_{n-1})\right] \qquad (2)$$

will be exactly $\mathbb{E}(\mu)$. To compute $\mathbb{E}(2X_1 + X_2 + \cdots + X_{n-1})$, we suppose

$$X_1 \sim \mathrm{U}\left[0, \frac{1}{n}\right]; \quad X_2 \sim \mathrm{U}\left[X_1, \frac{1 - X_1}{n-1}\right]; \quad X_3 \sim \mathrm{U}\left[X_2, \frac{1 - X_1 - X_2}{n-2}\right] \quad \cdots$$

$$X_{n-1} \sim \mathrm{U}\big[X_{n-2}, \frac{1 - X_1 - \cdots - X_{n-2}}{2}\big]; \quad X_n = 1 - X_1 - \cdots - X_{n-1}.$$

From the formula of conditional expectation, we have

$$\mathbb{E}X_{n-1} = \frac{1}{2}\left(\frac{1 - X_1 - \cdots - X_{n-2}}{2} + X_{n-2}\right)$$

$$= \frac{1}{2}\left(\frac{1 - X_1 - \cdots - X_{n-3}}{2} + \frac{1}{2}X_{n-2}\right)$$

$$\cdots$$

$$\mathbb{E}X_k = \frac{1}{2}\left(\frac{1 - X_1 - \cdots - X_{k-1}}{n-k+1} + X_{k-1}\right)$$

$$= \frac{1}{2}\left(\frac{1 - X_1 - \cdots - X_{k-2}}{n-k+1} + \frac{n-k}{n-k+1}X_{k-1}\right)$$

$$\cdots$$

$$\mathbb{E}X_2 = \frac{1}{2}\left(\frac{1 - X_1}{n-1} + X_1\right) = \frac{1}{2}\left(\frac{1}{n-1} + \frac{n-2}{n-1}X_1\right)$$

$$\mathbb{E}X_1 = \frac{1}{2n}.$$

Step by step replacing from the formula $\mathbb{E}X_{n-1}$ to the formula $\mathbb{E}X_1$ in the formula (2), basing on the inductive method we have the result

$$\mathbb{E}(Y_n) = \frac{a_{n-1}+1}{2(n-1)n} \qquad \forall n \geqslant 4, \tag{3}$$

where the sequence $\{a_n\}_{\mathbb{N}}$ is born by inductive formula:

$$a_1 = 1;\ a_2 = \frac{5}{4};\ a_3 = 1 + \frac{1}{2}\left(\frac{2}{3}a_2 - \frac{1}{2}a_1\right);\ a_4 = 1 + \frac{1}{2}\left(\frac{3}{4}a_3 - \frac{2}{3}a_2 - \frac{1}{2}a_1\right);\ \ldots$$

$$a_n = 1 + \frac{1}{2}\left(\frac{n-1}{n}a_{n-1} - \frac{1}{n-1}a_{n-2} - \frac{1}{n-2}a_{n-3} - \cdots - \frac{1}{3}a_2 - \frac{1}{2}a_1\right),\ \forall n \geqslant 5. \tag{4}$$

Formula (4) can be reduced to a simple inductive formula by using the differential method:

$$a_1 = 1;\ a_2 = \frac{5}{4}; a_3 = \frac{7}{6}; a_n = \left[\frac{n-1}{2n} + 1\right]a_{n-1} - \frac{1}{2}a_{n-2},\ \forall n \geqslant 4. \tag{5}$$

From formula (5), we easily recognize that when $n \geqslant 4$, the sequence $\{a_n\}_{\mathbb{N}}$ is decreasing and has the upper bound 1 and lower bound 0. Thus, we have an important in this section result.

**Proposition 4.2.** *Given any ordered rank score vector $x$ with length $n$ ($n \geqslant 5$), the expectation of its stability will be bounded as follows*

$$\frac{1}{2(n-1)n} \leqslant \mathbb{E}[\mu(x)] \leqslant \frac{1}{(n-1)n}.$$

The proof of Proposition 4.2 can be clearly seen through $\mathbb{E}[\mu(x)] = \mathbb{E}(Y_n) = \frac{a_{n-1}+1}{2(n-1)n}$ and $1 \geqslant a_n \geqslant 0$ for all $n \geqslant 4$.

## 5. MPageRank Algorithm and Conclusion

The naive algorithm computing accurately rank scores for all Webs is presented from equation (1). If our Web graph is connected so the complexity of the naive algorithm is $O(N^2)$, where $N$ is the number of pages in Web graph. In practice, this complexity is extremely high ($N \approx 10^{10}$). As we know, if our Web graph has an order $N$; however it is partitioned into $m$ subgraphs which have the corresponding order $N_i$, $(i = 1, \ldots, m)$ and do not connect to each other, so the complexity in computation of algorithm is $O(M^2)$, where $M = \max_{i=1,\ldots,m} N_i$. In addition, the storing to compute the rank score vector in formula (1) within the matrix having the size $N \approx 10^{10}$ requires a very advanced technology. It is so more difficult than computing the rank score vector within the matrix having the size $M \approx 10^8$.

From this observation and some results in the above section such as Corollary 3.6, Proposition 4.2, we would like to submit a cheap algorithm which approximates the rank score vector in MPageRank as follows.

*MPageRank Algorithm.*

Step 1: Computing a rank score vector MPageRank, $r_{T_0}$, with the topic $T_0$ being *general*. In other words, this MPageRank score vector is exactly the PageRank score vector in [13]. Let

$$W_0 = \left\{ t \in V \;\middle|\; r_{T_0}(t) < N^{-\frac{3}{2}} \right\}.$$

Step 2: Choose $k$ popular topics $T_1, T_2, \ldots, T_k$. For example, with $k = 5$, we can choose a collection of popular topics such as: *Politics, economics, culture, society, others*.

Step 3: For each topic $T_i$, finding an evaluation function $f_{T_i}$ to evaluate the relationship between the content of pages and this topic. Giving the set of pages which is unrelated to the topic $f_{T_i}$:

$$W_i = \left\{ t \in V \;\middle|\; \frac{f_{T_i}(t)}{f_{T_i}(V)} < N^{-\frac{3}{2}} \right\}.$$

Step 4: Constructing an expand rank score vector $r'_{T_i}$ on the Web graph $G_i = G \backslash (W_0 \cap W_i)$. Note that, we have the judgment from the practical experience, without proof:

$$r_{T_0}(t) < N^{-\frac{3}{2}} \quad \text{and} \quad \frac{f_{T_i}(t)}{f_{T_i}(V)} < N^{-\frac{3}{2}} \qquad \Rightarrow \qquad r_{T_i}(t) < N^{-3}.$$

It means that

$$\mathcal{I}(t) \leqslant \frac{1}{N} \left( \frac{16 r_{T_i}(t)}{p^2} \right)^{1/2} \leqslant \frac{4}{p N^{\frac{5}{2}}} \approx \frac{\mathbb{E}[\mu(r_{T_i})]}{N^{\frac{1}{2}}}.$$

Thus, for any $t \in W_0 \cap W_i$, the infuence $\mathcal{I}(t)$ is too tiny to change the order of rank score vector $r_{T_i}$. So we believe that the order of $r'_{T_i}$ is the same as the order of $r_{T_i}$.

Step 5: For each query $\mathcal{Q}$ from the users, evaluating the relationship between the query's topic and the collection of popular topics $\{T_1, T_2, \ldots, T_k\}$, we have the vector $(q_1, q_2, \ldots, q_k)$, $(q_1 + q_2 + \cdots + q_k = 1)$.

Step 6: A rank score vector $r_{\mathcal{Q}}$ to evaluate the authority of pages corresponding with query $\mathcal{Q}$ is constructed as follows

$$r_{\mathcal{Q}} = \sum_{i=1}^{k} q_i r'_{T_i}.$$

Step 7: Choosing some pages which have the most authority, and bring them to the users.

Although not being applied into reality, from its obvious theoretic results, the MPageRank algorithm is really effective in evaluating not only the authority, but also the computation and storage of pages. The results of this paper obviously are proved within simple theory. We hope that MPageRank algorithm in the near future is researched more effectively so that getting an exact point of view!

## References


1. R. Albert and A. Barabási, Statistical mechanics of complex networks, *Physica* **74** (2002), 47–97.
2. P. Baldi, P. Frasconi, and P. Smyth, *Modeling the Internet and the Web*, John Wiley & Sons, Inc. New York, 2003.
3. B. Bollobás, *Random Graphs*, Cambridge University Press, 2001.
4. D. Callaway, J. Hopcroft, J. Kleinberg, M. Newman, and S. Stragatz, Are randomly grown graphs really random?, cond-mat/0104546 v2, 14 Jun 2001.
5. F. Chung, Laplacians and the Cheeger inequality for directed graphs, *Annals of Combinatorics*, 2002, to appear.
6. F. Chung, *Spectral Graph Theory*, Amer. Math. Soc. No. 92 in the Regional Conference Series in Mathematics, Providence, RI, 1997.
7. J. Fakcharoenphol, *An Improved VC-Dimension Bound for Finding Network Failures*, Master's thesis, U.C, Berkeley, 2001.
8. T. Haveliwala, Topic-sensitive PageRank, in: *Proc. of the Eleventh Intern. World Wide Web Conference*, Honolulu, Hawaii, May 2002.
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. S. Kamvar, T. Haveliwala, C. Manning, and G. Golub, *Exploiting the Block Structure of the Web for Computing PageRank*, Stanford University Technical Report, 2003.
11. L. T. Kien, The probabilitic models for ranking Webs, Dep. Math. Hue Univ. Sci., May 2005, 91–110.
12. J. Kleinberg, *Detecting a Network Failure,* Proc. 41st Annual IEEE Symposium on Foundations of Computer Science, 2002.

13. L. Page, S. Brin, R. Motwani, and T. Windograd, *The PageRank Citation Ranking: Bring Order to the Web'*, Technical report, Stanford Digital Library Technologies Project, 1998.
14. M. Richardson and P. Domingos, The intelligent surfer: Probabilistic combination of link and content information in PageRank, in: *Proc. Advances in Neural Information Processing Systems 14*, Cambridge, Massachusetts, December 2002.