VNU Journal of Science, Mathematics - Physics 23 (2007) 35-46

Applying probabilistic model for ranking Webs in multi-context

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

[1]*Department of Mathematics, Hue University of Sciences, Vietnam*
77 Nguyen Hue, Hue city
[2]*Department of Computer Science, ELTE University, Hungary*

Received 15 May 2007

Abstract. The PageRank algorithm, used in the Google search engine, greatly improves the results of Web search by applying probabilistic model on the link structure of Webs to evaluate the "importance" of Webs. In PageRank probabilistic model, the links and webs are uniform, so the rank score of webs are quite independent from their content. In practice, the researchers often hope that the web results can be ranked by their proposed topics. Moreover, when computer's techniques solve given problems ineffectively, it's necessary to do better research in theoretical problems. From this judgement, in this paper, we introduce and describe the MPageRank based on a new probabilistic model supporting multi-context for ranking Webs. A Web now has different ranking scores, which depends on the given multi topics. The basic idea in establishing the new MPageRank model is that partition our Web graph into smaller-size sub Web graph. 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. Similar to the PageRank, the multi ranking scores in the MPageRank are pre-computed and reflect the hyperlink of Web environment.

1. Introduction

Nowadays the World Wide Web has became very large and heterogeneous, with an extraordinary grow rate. It creates many new challenges for information retrieval. One of the interesting problems is that evaluating the importance of a Web. The 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 used in the Google search engine is the most famous and effective one in practice. The underlying idea of PageRank is that using the stationary distribution of a *random surfer* on the Web graph in order to assign relating ranks to the pages. The link structure of the Web graph is an abundant source of information about the authority of the Webs. It encodes a considerable

* Corresponding author. Tel: 84-054-822407.
E-mail: hieukien@hotmail.com

amount of latent human judgment, and we claim that this type of judgment is necessary to formulate a notion of authority. In the probabilistic model of PageRank algorithm, the *random surfer* surfs indefinitely from page to page, following all *outlinks* with equal probability and the score of a page is the probability that the random surfer would visit that page. PageRank scores act as overall authority values of pages which are independent of any topic.

In practice, a user himself often has a proposed topic when he retrieves information in the internet. In fact, at first, the surfer seems to visit from the pages, which their 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. Moreover, the most difficult problem in PageRank is the rapid development of environment World Wide Web. When computer's techniques solve problems inffectively; obviously, theoretical problems should be studied more thoroughly. One of studying theoretical problems is the research of the *topological structure of Web graph* and the *partition Web graph*.

From the above observations, we introduce and describe the MPageRank algorithm. We assume 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 weighed and this weight is determined by the given popular topic. The probabilistic model in the MPageRank doesn't behavior uniform for all outlinks and nodes, it is improved by supporting the weight of web nodes. The rank scores of a Web are multi-values. The user can choose his proposed topic from the collection of given topics, and the chosen rank score is suitable for this topic. Certainly, the probabilistic model in MPageRank not only enables the user to choose his prefer topic but also models surf-Web process more precisely than the PageRank's. However, the main aim in building new MPageRank model is that weighting the Web graph; so thank to this, we study more effectively about the theory of partition Web graph. As we know, if our Web graph is partition into subgraphs which don't connect together, the calculation in algorithms will be reduced remarkably. From the definiton of the set (or node) ϵ-weak in Section 3.2, which evaluates the influence rate of one page to other pages, and several results in the Section 3.3 about approximating the rank score of original Web graph through partition Web graph, we can make the MPageRank algorithm to be cheaper.

The two best-know algorithms which improved Web search results by using the information hyperlink structure are HITS [1] and PageRank [2]. Given a query, HITS invokes a traditional search engine to obtain a set of pages relevant to it, expands this set with its inlinks and outlinks, and then attempts to find two types of pages, *hubs* and *authorities*. Because this computation is carried out at query time, it is not feasible for today's search engines, which need to handle billions of queries per day. In contrast, PageRank computes a single measure of quality for a page at crawl time so it is feasible for today's search engines as Yahoo!, Google, etc. But PageRank has the restriction that its score of a page ignores topic corresponding to the query and computation is too complex.

More recently, there are many approachs for surmount the probability score of page ignores topic corresponding to the query. M. Richardson and P. Domingos [3] 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. T. Haveliwala [4] 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, S. Kamvar,

T. Haveliwala et al. [5, 6] 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 random? There are interesting questions, they help us to realize Web environment from other way. 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 [7]. Based on model random graphs, R. Albert and A. Barabási [8] discovered the small-world property and the clustering coefficient of World Wide Web. Specially, they discovered that the degree distribution of the web pages follows a power law over several orders of magnitude. D. Callaway et al.[9] 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 [10-13] investigate the property of partition Web graph; most results have theoretical character. J. Kleinberg [10] introduced the notion (ϵ, k)-*detection set* play a role as the evidence for existence of sets which don't 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 ϵ fraction of the Web graph, that are disconnected from one another. J. Fakcharoenphol [11] showed that the (ϵ, k)-detection set for node failures can be found with probability at least $1-\delta$ by randomly chossing a subset of nodes of size $O(\frac{1}{\epsilon}k \log k \log \frac{k}{\epsilon} + \frac{1}{\epsilon}\log \frac{1}{\delta})$. F. Chung [12, 13] 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 F. Chung research.

The remainder of the paper is organized as follows: Section 2 is the preliminary. The result of the paper is all in Section 3. In this section, we introduce the MPageRank, present the set of Web pages having weak inffuence on other Webs. Then we give the result approximate to the rank score of the original Web graph from the rank score of the new Web graph after destroys all of *weak-pages*. Finally, section 4 will be the conclusion.

2. Preliminary

In this section, we give an outline of the probabilistic model of PageRank (2.1), the PageRank computation (2.2) and discuss the relationship between the content of a page and a given popular topic to supplement to PageRank algorithm (2.3).

2.1. Probabilistic Model of PageRank

PageRank is the algorithm that evaluates the authority of web pages based on the link structure. Link structure can be modelled by a 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 u to v $(u, v \in V)$. The *rank score vector* $r : V \to [0, 1]$ denotes the rank

score of pages, $r(u)$ is the score of page u. PageRank builds the rank score vector based on two following assumptions:

- The web pages, which are linked by many others pages, have a high score. In literature, we evaluate the authority of a page 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 then 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.

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. 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 to some other pages. Formally, time by time the surfer does two following actions:

(1) Generally, with probability $1-p$, the surfer surfs following all outlinks with equal probability.
(2) When the surfer feels bored, with the probability p, it jumps to all nodes in Web graph with an equal probability.

p is called *jump probability* ($0 < p < 1$), in practice we choose $p = 0.1$.

Hence, we can give the following intuitive description of PageRank: a page has a high rank if the sum of the ranks of its inlinks is high.

2.2. Rank score vector in PageRank

Let $N = |V|$ be the number of nodes in Web graph. Let u be a web page, F_u be the set of pages u points to, B_u be the set of pages that point to u and $O_u = |F_u|$ be the number of links from 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.

From the probabilistic model in MPageRank algorithm, the probability of event that the surfer is on page u at step i is given by the formula:

$$r_u^i = \frac{p}{N} + (1-p)\sum_{v \in B_u} \frac{r_v^{i-1}}{O_v}$$

Let $R = p\left[\frac{1}{N}\right]_{N\times N} + (1-p)M$, with $M_{uv} = \begin{cases} \frac{1}{O_u} & \text{if } (u,v) \in E \\ 0 & \text{otherwise} \end{cases}$

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

$$r^i = R^T r^{i-1}$$

The above formula shows that $(r^i)_N$ is a Markov chain with the state space V, corresponding the transition probability matrix R. It is well-know, see e.g. [14, Chap XV], that a Markov chain has uniquely a stationary probability distribution if, and only if, it is irreducible and aperiodic. Based on this knowledge, we have an important result:

Proposition 1. *The Markov chain $(r^i)_N$ exists uniquely the stationary probability distribution, be denoted r.*

[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, $(B_u \cup \{s\})$

Proof. Thus, our Web graph G has probability move from node u to node v: $R_{uv} > 0$ so $(r^i)_N$ is *an irreducible chain*. Moreover, each node $u \in V$, since $p_{vu} = R_{vu} \geqslant p$ so u has a period $t = 1$. Therefore node u is aperiodic for $u \in V$, so the state space V has only one positive recurrence class (it means that this is *an aperiodic chain*). In fact, the Markov chain $(r^i)_N$ exists uniquely the stationary probability distribution, r.

This stationary distribution r, itself is a rank score vector in PageRank. Rank score vector in PageRank 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 probability matrix R correspond with *eigenvalue 1*.

2.3. *Supplement to PageRank algorithm*

Generally, while user retrieves information in internet, he would like to find information related to the determined topic. Hence, he has a tendency to retrieve web pages which have content related to this topic. For example, when a user find information about the Manchester United football team, certainly he prefers to find some web pages having content related to sport topic.

From the above observation, we propose the third assumption that supplements the two assumption of PageRank:

- With a given topic, a page having its content related to this topic will have a high score.

However, how to evalute the relating rate of a Web page with a given topic based on its content? This is a big and complex problem which attract 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 & Sons has published the book [15] and has one chapter to present this problem. The techniques are 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 is that: "Let a topic T, we can have an *evaluation function* $f_T : V \longrightarrow [0, 100]$ to evaluate how relationship between a page and this topic is." After constructing the evaluation function f_T for the topic T, where $f_T(u)$ evaluates how the page u related to the topic T, we introduce a new probabilistic model for ranking Webs, MPageRank, improvement of PageRank model based on the evaluation about Web page importance related to the given topic. Moreover, from the weighed Web graph technique, we present some new theoretical results to understand more clearly the partition property of Web graph.

3. The MPageRank

There are three problems we discuss in this section. The first, we will describe probabilistic model in MPageRank algorithm. Next, in theory, we will evaluate and propose quantitatives to partition

the set of Web pages in Web graph. The end, we will present basic results to suggest the direction of the cheap algorithm, MPageRank.

3.1. Probabilistic Model of MPageRank

Based on above discussion, we construct the MPageRank algorithm according to a new probabilistic model. To begin constructing the MPageRank, we choose k popular topics $T_1, T_2, \ldots, T_k$; (e.g. with $k = 5$, we can choose a collection of popular topics such as: Politics, Economics, Culture, Society, Others). For each topic T_i, we consider and give an evaluation function f_i to evaluate the relationship between the content of pages and this topic.

We build the MPageRank algorithm satisfies three following assumptions:

- The web pages, which are linked by many others pages, have a high score.
- If a high score page links to some pages then its destination has high score too.
- With a given topic, a page having its content related to this topic will have a high score.

We choose the rank score vector r_M as the the standing probability distribution of a random surfer on the Web graph. However, difference of PageRank, in MPageRank the surfer doesn't surf following all outlinks and choose all the pages when he feels boring with equal probability. It depends on the topic which the user choose. For each topic T_i, the surfer surfs following outlink $(u, v) \in E$ and jumps to page v when he feels bored with probability:

$$p_{uv} = \frac{f_i(v)}{\sum_{j \in F_u} f_i(j)} \quad ; \quad p_v = \frac{f_i(v)}{\sum_{j \in V} f_i(j)}$$

Formally, time by time this surfer does two following actions:

(1) Generally, with probability $1 - p$, the surfer stayed at page u 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, it jumps to all pages in Web graph, where page v is probability p_v.

Like to the calculation in PageRank, we calculate rank score function r_M in MPageRank as following:

The probability of event that the surfer is on page u at step i is given by the formula:

$$r_M^i(u) = pp_u + (1 - p) \sum_{v \in B_u} p_{vu} r_M^{i-1}(v)$$

Let $R_M = pR^1 + (1 - p)R^2$, where R^1, R^2 are a $N \times N$ matrix 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_M 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:

$$r_M^i = R_M^T r_M^{i-1}$$

Certainly, $(r_M^i)_N$ is a Markov chain with the state space V. Similar to PageRank, we have another result:

Proposition 2. *The Markov chain $(r_M^i)_N$ exists uniquely the stationary probability distribution, be denoted r_M.*

Proof. If the Markov chain $(r^i_M)_N$ has only one irreducible closed subset S, and if S is aperiodic, then the chain must have a unique the stationary probability distribution. So we simply must show that the Markov chain $(r^i_M)_N$ has a single irreducible closed subset S, and that this subset is aperiodic.

Let the set U be the states with nonzero components in $v = (p_u)_{N\times 1}$. Let S consist of the set of all states reachable from U along nonzero transition in the chain. S trivially forms a closed subset. Further, since every state 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 state in S has a self-transition, then the subset S is aperiodic. Let u be any state in U. By construction, there exists a self-transition from u to itself. Therefore S must be aperiodic, so the Markov chain $(r^i_M)_N$ exists uniquely the stationary probability distribution, r_M.

The stationary distribution r_M is the rank score vector in MPageRank and it is given by formula:

$$r_M = R^T_M r_M \tag{2}$$

R^T_M is the stochastic matrix so rank score vector r_M is equivalent to *primary eigenvector* of the transition matrix R_M correspond with *eigenvalue* 1.

The naive algorithm computing accurately multi-rank scores for all Webs is presented from equation (2). If our Web graph is connective 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 ($N \approx 6.10^9$). As we know, if our Web graph has an order N; however it partition into m subgraphs which has the corresponding order N_i, $(i = \overline{1,m})$ and don't connect to each other, so the complexity in computation of algorithm is $O(M^2)$, where $M = \max_{i=\overline{1,m}} N_i$. From this observation, we would like to submit a cheaper algorithm which approximates the rank score vector in MPageRank. Our basic idea in forming the cheap MPageRank algorithm is that rejects most of Web pages which influence weakly on MPageRank score of other pages. And Web graph can be partitioned by shrinking to a graph created from the remain of Web pages. The influence of one page on other pages according to topic depends on two factors: *the hyperlink structure* (specify in PageRank score) and *the content evaluation function* related to the topic. A central problem of forming the cheap MPageRank algorithm is answering a question ***"How the rank score of pages change when we rejects some special pages and their conjugate edges?"***. We will give the answer of this question in two subsection follows:

3.2. Classification of the Web pages

Definition 1. *Let a structure Web graph, a page is called* ***the strong structure*** *if its PageRank score taken in this Web graph is high, and a page is called* ***the weak structure*** *if its PageRank score is low.*

Let a given topic, a page is called ***related*** *if its evaluation function value is high, and a page is called* ***unrelated*** *if its evaluation function value is low.*

Defenition 2. *Let a set of Web pages having structure Web graph and a given topic.* ***The weakest authority*** *set is the set containing all of pages which are weak structure and unrelated.*

We classify the set V, the set all of web page in Web graph, according to two subsets. W is a set which contains all of pages in the weakest authority set, and S contains all that remains of page[2]. Certainly, if we define topic's score of a set is the sum of all topic's score of pages in it then the topic's score of W is too lower than the topic's score of S.

[2] $S = V \backslash W$

Let a Web graph $G = (V, E)$ and the given topic T. We have a transition matrix R_M and evaluation function f_T for all of pages in Web graph. From MPageRank algorithm we have rank score vector r_M. Let a subset U of V, we write $r_M(U) = \sum_{u \in U} r_M(u)$ and $f_T(U) = \sum_{u \in U} f_T(u)$, so we have some basic notions as follows:

Defenition 3. *A node u is called ϵ-weak if $r_M(u) \leqslant \epsilon$.*

A subset U of V is called ϵ-weak if $r_M(U) \leqslant \epsilon$.

Defenition 4. *A subset U is called* weak *if the transition probability from $V\backslash U$ to U is smaller than the transition probability from $V\backslash U$ to $V\backslash U$ and the transition probability from U to $V\backslash U$ is smaller than the transition probability from $V\backslash U$ to $V\backslash U$.*

It is easy to recognize the subset W is a *weak* set. Let $\epsilon = \frac{f_T(W)}{f_T(S)}$ (ϵ is too tiny), we have a result.

Theorem 1. *W is an ϵ-weak set.*

Proof. We can see the detail of solution to Theorem 1 in [16]. The set W is a weak set so the transition probability from S to W is smaller than the transition probability from S to S, and the transition probability from W to S is smaller than the transition probability from S to S. It is the main reason for doing $\frac{r_M(W)}{r_M(S)} \leqslant \frac{f_T(W)}{f_T(S)} = \epsilon$, so $r_M(W) \leqslant \frac{\epsilon}{\epsilon+1} \leqslant \epsilon$.

We see that the rank score of pages in set W is really tiny and doesn't have influence on rank score of other pages. Therefore, rank score vector in MPageRank is decided by pages in set S. Indeed, with a weak page $u \in W$, if we reject page u and its conjugate edges, we will have an interesting question that how the rank score of other pages will change? With the same question when we reject a set of really weak pages $U \subset W$. That is what we will answer in the next section.

3.3. Main results

Let a given popular topic T, we have a weight directed graph $G = (V, E)$ with a transition probability matrix in MPageRank algorithm is R_M. For $u \in V(G)$ is a weak vertex, get $G' = G\backslash u$ is a graph (V', E') where $V' = V\backslash\{u\}$ and $E' = \{v_1v_2 \mid v_1, v_2 \in V' \ , \ v_1v_2 \in E\}$. Let R'_M is a transition probability matrix corresponding to a random surfer in the new Web graphs G'. The new random surfer will have a stationary distribution, denote by r'_M. We have an interesting judgement that the random surfer on the graph G' with MPageRank transition probability matrix R'_M is equivalent to another random surfer on the graph G with MPageRank transition probability matrix R^*_M when the evaluation function value $f_T(u) = 0$. Let r^*_M is a stationary distribution of random surfer on the graph G corresponding the transition probability matrix R^*_M, and called r^*_M is *an expand MPageRank rank score vector* of Web graph G'; $\Delta R_M = R^*_M - R_M$, $\Delta r_M = r^*_M - r_M$.

As the question submited above, we would like to know how the rank score vector, $\Delta r_M = r^*_M - r_M$, will change when rejecting page u and its conjugate edges. Let G is a Web graph and a random surfer in MPageRank algorithm surf on its. We have a transition probability matrix R_M. If R_M has a stantionary distribution r_M, then let a matrix

$$\mathcal{L} = \mathbf{I} - \frac{D^{1/2}R_MD^{-1/2} + D^{-1/2}R_M^TD^{1/2}}{2}$$

where D is a diagonal matrix with entries $D(v, v) = r_M(v)$. $\mathcal{L}$ is called *an expand Laplacian matrix* of a directed Web graph G. Clearly, the expand Laplacian is real symmetric, 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 $\lambda = min_{i\neq 0}|\lambda_i|$ is *an expand algebraic connectivity* of Web graph G, so we have an important result[3]

Proposition 2. *For any tiny real number $\epsilon > 0$, and a weak page u, $r_M(u) \leqslant \epsilon$. If r^*_M is an expand rank score vector of Web graph when we reject page u and its conjugate edges, then*

$$\|\Delta r_M\|^2 = \|r^*_M - r_M\|^2 \leqslant \frac{2r_M(u)}{\lambda} \leqslant \frac{2\epsilon}{\lambda}.$$

Proof. To prove Theorem 2, we consider the Lemma:

Lemma 1. *We have*

$$|[\Delta R^T_M.r_M](i)| \leqslant r_M(u),\ \forall i \in V\backslash\{u\}.$$

Proof. Let $B^1_u = \{v \in B_u \mid F_v \neq \{u\}\}$, $B^2_u = B_u\backslash B^1_u = \{v \in B_u \mid F_v = \{u\}\}$, we have

- If $i \neq u$ and $i \notin F_u$

$$\begin{aligned}[\Delta R^T_M.r_M](i) &= \sum_{j\in B^1_u} \Delta R^{ji}_M.r_M(j) + \sum_{j\in B^2_u} \Delta R^{ji}_M.r_M(j) + \Delta R^{ui}_M.r_M(u)\\ &= \sum_{j\in B^1_u\cap B_i} \frac{f_T(i)}{f_T(F_j)-f_T(u)}\frac{f_T(u)r_M(j)}{f_T(F_j)} + \sum_{j\in B^2_u} \frac{f_T(j)}{f_T(V)-f_T(u)}\frac{f_T(u)r_M(j)}{f_T(F_j)}\end{aligned}$$

because when $j \in B^2_u$ so $F_j = \{u\} \Rightarrow f_T(u) = f_T(F_j)$. Clearly, $\frac{f_T(i)}{f_T(F_j)-f_T(u)} \leqslant 1$ and $\frac{f_T(j)}{f_T(V)-f_T(u)} \leqslant 1$, we have

$$\begin{aligned}|[\Delta R^T_M.r_M](i)| &\leqslant \frac{1}{1-p}\Big[(1-p)\sum_{j\in B_u} \frac{f_T(u)r_M(j)}{f_T(F_j)} + p\frac{f_T(u)}{f_T(V)}\Big] - \frac{p}{1-p}\frac{f_T(u)}{f_T(V)}\\ &\leqslant \frac{1}{1-p}r_M(u) - \frac{p}{1-p}\frac{f_T(u)}{f_T(V)}.\end{aligned}$$

From Theorem 1, if page u is weak, we have

$$r_M(u) \leqslant \frac{f_T(u)}{f_T(V)} \Rightarrow \frac{1}{1-p}r_M(u) - \frac{p}{1-p}\frac{f_T(u)}{f_T(V)} \leqslant r_M(u).$$

- If $i \neq u$ and $i \in F_u$

$$\begin{aligned}|[\Delta R^T_M.r_M](i)| &= \Big|\sum_{j\in B^1_u} \Delta R^{ji}_M.r_M(j) + \sum_{j\in B^2_u} \Delta R^{ji}_M.r_M(j) + \Delta R^{ui}_M.r_M(u) - \frac{f_T(i)}{f_T(F_u)}r_M(u)\Big|\\ &\leqslant \Big|\Big[\frac{1}{1-p}r_M(u) - \frac{p}{1-p}\frac{f_T(u)}{f_T(V)}\Big] - \frac{f_T(i)}{f_T(F_u)}r_M(u)\Big|\\ &\leqslant max\Big\{\frac{1}{1-p}r_M(u) - \frac{p}{1-p}\frac{f_T(u)}{f_T(V)}, \frac{f_T(i)}{f_T(F_u)}r_M(u)\Big\}\\ &\leqslant r_M(u).\end{aligned}$$

Lemma is proven.

[3] We can see carefully these conceptions in [16].

Now, we prove Theorem 2. We have

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

From Lemma 1 and $\sum_i r_M(i) = \sum_i r_M^*(i) = 1$, we have

$$|r_M^T \Delta R_M \Delta r_M| \leqslant 2r_M(u).$$

To prove

$$\|\Delta r_M\|^2 \leqslant \frac{2r_M(u)}{\lambda}$$

we consider the second Lemma

Lemma 2. [16] *For a stochastic matrix R with order n; d is a vector with same order n and satisfied $\sum d_i^2 = 1$. Let a diagonal matrix D, where $D_{ii} = d_i > 0$. So we have*

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

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

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

So if $\Delta' r_M$ is $(N-1)$-vector which produced from vector Δr_M by rejecting page u, then $\sum_i \Delta' r_M(i) = 0$ (vector $\Delta' r_M$ orthogonal with $e = (1, \ldots, 1)^T$).

Therefore we have

$$\begin{aligned} |\Delta r_M^T[\mathbf{I}_N - R_M^*]\Delta r_M| &= |\Delta' r_M^T[\mathbf{I}_N - R_M']\Delta' r_M| \geqslant \lambda \|\Delta' r_M\|^2 \\ \Rightarrow \quad \lambda \|\Delta' r_M\|^2 &= \lambda \|\Delta r_M\|^2 \leqslant 2r_M(u) \\ \Rightarrow \quad \|\Delta r_M\|^2 &\leqslant \frac{2r_M(u)}{\lambda} \leqslant \frac{2\epsilon}{\lambda}. \end{aligned}$$

The Theorem is proven.

As we know, the value λ is called an algebraic connectivity of Web graph G according to the transition probability matrix R_M. In the paper [16], we have a result to bound the value λ as follow:

Let a weight directed graph G which $f_T(v)$ is a weight value for each node v. The transition probability matrix R_M of random surfer in MPageRank surfed on graph G is defined as follows:

For a real number $p \in [0,1]$, $\forall i, j \in V(G)$ then

$$R_M(i,j) = \begin{cases} (1-p)\dfrac{f_T(j)}{\sum\limits_{k \in F_i} f_T(k)} + p\dfrac{f_T(j)}{\sum\limits_{k \in V(G)} f_T(k)} & \text{if } O_i > 0 \\ \dfrac{f_T(j)}{\sum\limits_{k \in V(G)} f_T(k)} & \text{if } O_i = 0 \end{cases}$$

p is a jump probability[4] .

Proposition 3. [16]. *If λ is an expand algebraic connectivity of G, then we have*

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

As a directed consequence of Theorem 2 and Proposition 3, we have two important results.

Corollary 1. *For a tiny real number $\epsilon > 0$, and a weak page u, $r_M(u) \leqslant \epsilon$. If r_M^* is an expand rank score vector of Web graph when we reject page u and its conjugate edges, then*

$$\|\Delta r_M\|^2 \leqslant \frac{16 r_M(u)}{p^2} \leqslant \frac{16\epsilon}{p^2}.$$

Corollary 2. *For a tiny real number $\epsilon > 0$, and a set of weak pages $W \subseteq V(G)$, $r_M(W) \leqslant \epsilon$. If r_M^* is an expand rank score vector of Web graph when we reject all of pages in W and their conjugate edges, then*

$$\|\Delta r_M\|^2 \leqslant \frac{16 r_M(W)}{p^2} \leqslant \frac{16\epsilon}{p^2}.$$

4. Conclusion

To highlight the consideration to user's purpose, we introduced and described MPageRank algorithm according to improved probabilistic model which allowed ranking Webs depending on the given topic. Different to PageRank just conforms only two assumptions, the model probability in MPageRank conforms three assumptions. In MPageRank model, we supplemented more assumption that is:

- Considering with a given topic, page having its content related to this topic will has a high score.

We believe that our model will model more exactly upon real surf-Web. Therefore in theory, our rank score of pages will satisfy more sufficient for the users.

Similar to the computation in PageRank, MPageRank model is preformed based on knowledge of Markov chain. Our transition matrix is irreducible and aperiodic so rank score function in MPageRank exists and itself is a primitive eigenvector of this transition matrix with eigenvalue 1. From the ideas that partition Web graph to many subgraphs to make the algorithm to be more simple, this paper introduces the way to approximate rank score vector when we reject some weakly influenced pages and their conjugate edges.

Of course, this paper doesn't give the way to known where the page, called *the bridge* of Web graph, which when we reject it and its conjugate edges, the Web graph will be disconnected, and

[4] we can see the definition of O_i in page 4 of this paper.

what an given popular topic making our Web graph having many bridges. It is difficult and important problems. This is our future works!

References

[1] J. Kleinberg, Authoritative Sources in a Hyperlinked Enviroment, *Journal of ACM*, 46 (1999) 604.
[2] L. Page, S. Brin, R. Motwani, T. Windograd, The PageRank Citation Ranking: Bring Order to the Web', *Technical report*, Stanford Digital Library Technologies Project 1999-0120, 1998.
[3] M. Richardson, P. Domingos, The intelligent surfer: Probabilistic combination of link and content information in PageRank, *In Proceedings of Advances in Neural Information Processing Systems 14*, Cambridge, Massachusetts, Dec. 2002.
[4] T. Haveliwala, Topic-Sensitive PageRank, *In Proceedings of the Eleventh International World Wide Web Conference*, Honolulu, Hawaii, May 2002.
[5] S. Kamvar, T. Haveliwala, C. Manning, G. Golub, Extrapolation methods for accelerating PageRank computations, *In Proceedings of the Twelfth International World Wide Web Conference*, 2003.
[6] S. Kamvar, T. Haveliwala, C. Manning, G. Golub, *Exploiting the Block Structure of the Web for Computing PageRank*, Stanford University Technical Report, 2003.
[7] B. Bollobás. *Random Graphs*, CAMBRIDGE University Press, 2001.
[8] R. Albert, A. Barabási, Statistical mechanics of complex networks, *Reviews of Modern Physics*, Vol 74, January 2002.
[9] D. Callaway, J. Hopcroft, J. Kleinberg, M. Newman, S. Stragatz, Are randomly grown graphs really random?, *Phys. Rev. E* 64 (2001) 041902.
[10] J. Kleinberg, Detecting a Network Failure *Proc. 41st Annual IEEE Symposium on Foundations of Computer Science*, 2002.
[11] J. Fakcharoenphol, *An Improved VC-Dimension Bound for Finding Network Failures*, Master's thesis, U.C. Berkeley, 2001.
[12] F. Chung, Laplacians and the Cheeger inequality for directed graphs, *Annals of Combinatorics*, 2002.
[13] F. Chung, *Spectral Graph Theory*, American Mathematical Society, No.92 in the Regional Conference Series in Mathematics, Providence, RI, 1997.
[14] William Feller. *An Introduction to Probability Theory and Its Applications*. Vol. 1, 3rd ed. John Wiley & Sons, Inc. New York, 1968.
[15] P. Baldi, P. Frasconi, P. Smyth, *Modeling the Internet and the Web*, John Wiley & Sons, Inc. New York, 2003.
[16] Le Trung Kien, *The probabilitic models for ranking Webs*, Graduate's thesis, Hue University of Sciences, May 2005.