

Predictive Ranking: A Novel Page Ranking Approach by Estimating the Web Structure

Haixuan Yang
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Shatin, N.T., Hong Kong
hxyang@cse.cuhk.edu.hk

Irwin King
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Shatin, N.T., Hong Kong
king@cse.cuhk.edu.hk

Michael R. Lyu
Dept. of Comp. Sci. and Eng.
The Chinese Univ. of HK
Shatin, N.T., Hong Kong
lyu@cse.cuhk.edu.hk

ABSTRACT

PageRank (PR) is one of the most popular ways to rank web pages. However, as the Web continues to grow in volume, it is becoming more and more difficult to crawl all the available pages. As a result, the page ranks computed by PR are only based on a subset of the whole Web. This produces inaccurate outcome because of the inherent incomplete information (dangling pages) that exist in the calculation. To overcome this incompleteness, we propose a new variant of the PageRank algorithm called, *Predictive Ranking* (PreR), in which different classes of dangling pages are analyzed individually so that the link structure can be predicted more accurately. We detail our proposed steps. Furthermore, experimental results show that this algorithm achieves encouraging results when compared with previous methods.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Search and Retrieval

General Terms: Algorithms, Experimentation, Theory

Keywords: PageRank, Link Analysis, Predictive Ranking

1. INTRODUCTION

The PageRank (PR) algorithm has proven to be very effective for ranking Web pages when given a well-defined and accurate Web structure. However, as the Web continues to grow, it becomes less possible to crawl all existing Web pages due to the large volume, bandwidth constraints, and dynamic nature. Consequently, the page ranks computed by PR are only based on a subset of the whole Web. This causes inaccurate results because of its incomplete information about the Web structure. Can we increase the accuracy of the PR? In this poster, we propose a solution to this problem by formulating a new model called, *Predictive Ranking* (PreR), in which the Web structure is estimated more accurately leading to a better and more accurate PR result.

In [3], PR gives the relative importance of a Web page based on the link structure of the Web. Formally presented in [1], the Web is modeled by a directed graph $G = (V, E)$, and the rank x_i for page $i \in V$ is defined recursively in terms of pages which point to it: $x_i = \sum_{(j,i) \in E} a_{ij}x_j$; in matrix terms, $x = Ax$.

There are three types of pages that a crawler could encounter while exploring the web structure. They are: Type

1—those that are found, but not visited or not visited successfully, Type 2—those that are visited but from which there is no outlink, and Type 3—those that are visited and from which there is at least one outlink. In [3], only pages of Type 3 are considered in the graph model while pages of Type 1 and Type 2 (dangling pages) are computed at the last iteration. In [2] and [1], all pages of Type 1 to 3 are included in the graph model, but pages of Type 1 and 2 are combined together. In our proposed model, we separate pages of Type 1 from those of Type 2, and analyze each statistically to obtain a more accurate estimated Web structure, based on which, PreR will produce better ranking results.

Including dangling pages in the overall ranking may have significant effect on the ranks of non-dangling pages [1]. Moreover, ranking on the dangling pages enables us to return to the users some useful sites, which contains the matching words in their names and are found (but not visited) by the crawler currently. In [1], dangling pages are handled by adding a virtual page so that the computation is efficient. Our model is different from the models in [2] and [1] in that some information in our model is obtained by prediction.

2. PREDICTIVE RANKING MODEL

In general, it is difficult to estimate link structure accurately; however, some elementary estimation is possible. We can estimate the in-degree of each page, and thus some information about the link structure can be inferred statistically. The dogma in PreR is this—the more we know about the structure of the Web, the more accurate we can infer about it. We formulate our PreR model in the following six steps:

Step 1. Partition all the pages V of the graph ($|V| = n$) into three subsets: S , D^1 , and D^2 , where S ($|S| = m$) denotes the subset of all pages of Type 3; D^1 ($|D^1| = m_1$) denotes the subset of all pages of Type 2; and D^2 ($|D^2| = n - m - m_1$) denotes the set of all pages of Type 1.

Step 2. Predict the in-degree $d^-(v_i)$ by the number of found links $fd^-(v_i)$ from visited pages to the page v_i . With the breadth-first crawling method, we assume that the number of found links $fd^-(v_i)$ from visited pages to the page v_i is proportional to the real number of links from all pages in V to the page v_i , and thus $d^-(v_i) \approx n/(m + m_1) \cdot fd^-(v_i)$. This assumption is meaningful since although the crawler crawls the Web from a given Web site to other sites in a definite way, its ability of finding new link to a given page v_i depends on the density of these links. The density of these links to the page v_i is equal to $d^-(v_i)/n$. Since the crawler has found $fd^-(v_i)$ such kind of links from m pages,

$fd^-(v_i)/(m+m_1)$ is an approximate estimate of the density. Following this, we obtain the above approximate equality.

Step 3. Estimate matrix A . All links ($fd^-(v_i)$) are from the pages in S , and the remaining links ($d^-(v_i) - fd^-(v_i)$) are from the pages in D^2 (it is impossible that some of these links are from the pages in D^1). Without any prior information about the distribution of these remaining links, we assume that they are distributed uniformly from the pages in D^2 to the page v_i , i.e., these remaining links are shared by all the pages in D^2 . So matrix A , modelling the users' behavior in following the actual links, is estimated to be

$$A = \begin{pmatrix} C & P & M \\ D & Q & N \end{pmatrix}. C \text{ and } D \text{ are used to model the known}$$

link structure from S to V in [3], where c_{ij} in C and d_{ij} in D are defined as $1/d_j$ if there is a link from j to i and 0 if not, d_j is the out-degree of page j ; M and N , modelling the link structure from D^2 to V , are defined as: $(M \ N)^T = \text{diag}\{l_1, \dots, l_n\}$, where $l_i = (d^-(v_i) - fd^-(v_i))/m_2\Sigma$. m_2 is used to share the remaining inlinks $d^-(v_i) - fd^-(v_i)$ uniformly by all pages in D^2 and $\Sigma = \sum_{i=1}^n d^-(v_i) - fd^-(v_i)$ is multiplied to the denominator to make the matrix to be stochastic. P and Q , modelling the link structure from D^1 to V , will be defined in Step 5.

Step 4. Model the users' teleportation. Assume that the users will jump to page v_i with a probability of f_i when they get bored in following the actual links. So the matrix modelling the teleportation is fe^T . We denote vector $(f_1 \ f_2 \ \dots \ f_n)^T$ by f . Previous suggestions include the choice of a uniform distribution among all pages, among a set of trusted "seed sites", uniformly among a set of all "top-level" pages of sites, or a personalized set of preferred pages.

Step 5. Set matrix $(P \ Q)^T = \text{diag}\{f_1, \dots, f_n\} \mathbf{1}_{n \times m_1}$. When the user encounters a page of Type 2, there is no outlink that the users can follow. In this case, we assume that the same kind of teleportation as in Step 4 will happen.

Step 6. Rank x_i should satisfy $x = [(1 - \alpha)fe^T + \alpha A]x$, where α is the probability of following an actual out-link from a page and $1 - \alpha$ is that of taking a "random jump".

This model is thus named as Predictive Ranking (PreR) algorithm. Different from [1, 2], $(M \ N)^T$ is decomposed from A in PreR and is constructed by predicted information. This gives rise to a more accurate estimate of the Web structure and subsequently a more accurate result.

3. EXPERIMENTS

For our experiments, we choose a relatively small subset of the Web, the network within the domain `cuhk.edu.hk` because we are able to obtain a relatively complete structure about the pages, and therefore the relatively accurate ranks can be calculated to make an easier comparison.

Because the importance of a Web page is an inherently subjective matter, it is difficult to measure whether a link analysis algorithm is better than another. However, we design a novel comparison by calculating the difference between the early results (less accurate) and the final results (relatively accurate). More specifically, we take snapshots of the 11 matrices during process of crawling the page, namely, A_1, \dots, A_{11} . The numbers of pages visited successfully and the total numbers of pages only found at time 11 are 502,610 and 607,170 respectively. At time 1 to 10, those numbers are: 7712, 18542; 78662, 120970; 109383, 157196; 160019, 234701; 252522, 355720; 301701, 404728; 373579, 476961; 411724, 515534; 444974, 549162; 471684, 576139.

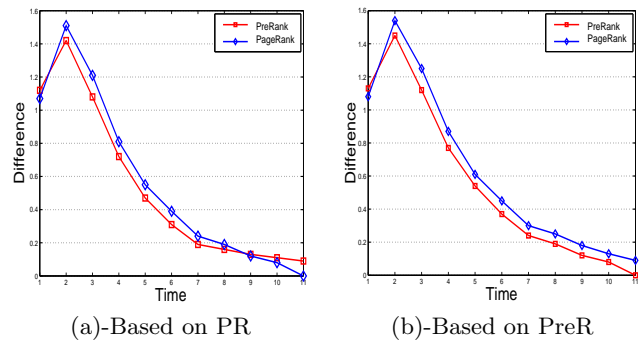


Figure 1: Comparison Results

By applying both the PreR algorithm and the modified PageRank Algorithm (PR) in [2] to these 11 data sets, we get rank results $PreR[t]$ and $PR[t]$ ($t = 1, 2, \dots, 11$). We set $\alpha = 0.85$ and set f to be uniform distribution in both algorithms.

The difference $D1[t]$ between $PreR[t]$ and $PR[11]$ is measured as $\|PreR[t] - Cut(t, PR[11])\|_1 / Sum[t]$, and the difference $D2[t]$ between $PR[t]$ and $PR[11]$ is calculated similarly. Where $cut(t, PR[11])$ means the vector cut from $PR[11]$ such that it has the same dimension as $PR[t]$, and $Sum[t]$ means the sum of values in vector $cut(t, PR[11])$. The results of $D1[t]$ and $D2[t]$ are shown in figure 1(a). In the figure, at time 1, $PR[1]$ is closer than $PreR[1]$ to $PR[11]$, this happens because at time 1, the data set is so small that the statistic estimation is not accurate sometimes. But as the time grows, from time 2 to time 8, $PreR[t]$ is closer than $PR[t]$ to $PR[11]$. As we expected, as time t is near to the end of 11, $PR[t]$ again is closer than $PreR[t]$ to $PR[11]$, this happens because we use the $PR[11]$ as comparison reference and so it is biased against $PreR[t]$. Even so, in 7 out of 11 (63%) cases, $PreR[t]$ is closer to final PR result $PR[11]$. If we use $PreR[11]$ as comparison reference, at all time t except 1, i.e., in 91% cases, $PreR[t]$ is closer to $PreR[11]$ than $PR[t]$. The results can be seen in Fig. 1(b).

4. CONCLUSION

The results of PreR is more accurate (closer to the final result) than those of PR. Even when we consider the results of PR as the reference (bias against our model), early results calculated by PreR are closer to the reference than the results calculated by PR.

5. ACKNOWLEDGMENTS

We thank Mr. Patrick Lau for his contributions to the experiments. This work is supported by grants from the Research Grants Councils of the HKSAR, China (CUHK4205/04E and CUHK4351/02E).

6. REFERENCES

- [1] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the Web frontier. In *Proc. of the 13th World Wide Web Conference*, pages 309–318, 2004.
- [2] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub. Exploiting the block structure of the Web for computing pagerank. Technical report, Stanford University, 2003.
- [3] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report, Stanford University, 1999.