DRANK+: A DIRECTORY BASED PAGERANK PREDICTION METHOD FOR FAST PAGERANK CONVERGENCE

Hung-Yu Kao, Chia-Sheng Liu, Yu-Chuan Tsai
Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC

Chia-Chun Shih, Tse-Ming Tsai
Innovative Digitech-Enabled Applications & Services Institute (IDEAS), Institute for Information Industry, Taiwan

Keywords: Search engine, Link Analysis, PageRank, Web Graph, Hierarchical Structure, Page Quality.

Abstract: In recent years, most part of search engines use link analysis algorithms to measure the importance of web pages. The most famous link analysis algorithm is PageRank algorithm. However, previous researches in recent years have found that there exists an inherent bias against newly created pages in PageRank. In the previous work, a new ranking algorithm called DRank has been proposed to solve this issue. It utilizes the cluster phenomenon of PageRank in a directory to predict the possible importance of pages in the future and to diminish the inherent bias of search engines to new pages. In this paper, we modify the original DRank algorithm to complement the weaker part of DRank which could fail while the number of pages in directory is not enough. In our experiments, the augmented algorithm, i.e., DRank+ algorithm, obtains more accuracy in predicting the importance score of pages at next time stage than the original DRank algorithm. DRank+ not only alleviates the bias of newly created pages successfully but also reaches more accuracy than Page Quality and original DRank in predicting the importance of newly created pages.

1 INTRODUCTION

Since the rising of the World Wide Web, more and more people change their behavior seeking for information from local scale to Web scale. The Web users are relying on search engines day by day. The whole search results, which are related to keyword queries, we may ask that how many pages the Web user would view. The report in Pewinternet [1] indicates that 62% of search engine users click on a search result within the first page of results. The 90% of search engine users click on a result within the first three pages of search results. In other words, if a page is not ranked in the top of the search results, it is hard to be clicked by the users. Also, link analysis algorithms have been widely used in recent years. Most of existing search engines adopt the link analysis algorithm to measure the importance of Web pages. It mainly considers the link structure between pages and uses the link relation to evaluate the ranking score of a page.

A new algorithm was proposed by Page and Brin called PageRank (Brin and Page, 1998) to measure the importance of Web pages. PageRank defines the importance of Web pages and helps a search engine to select high quality pages more efficiently. However, PageRank algorithm still suffers a problem: biased ranking to newly pages. Since the newly pages often do not receive enough in-links to show its real importance in the initial time stage. The work in (Cho and Roy, 2004) showed how to evolve the importance of the newly pages after taking a long time. The bias is unfired to the newly pages and thus the search results will be undependable.

The main goal in this paper is to alleviate the inherent bias in search engines to the newly pages. We design a new algorithm to speed up the convergence of the importance score of a page. Toward this goal, we first observe the intrinsic feature of World Wide Web through simple

[1] http://www.pewinternet.org/pdfs/PIP_Data_Memo_Searchengines.pdf

Kao H., Liu C., Tsai Y., Shih C. and Tse-Ming T. (2008).
DRANK+: A DIRECTORY BASED PAGERANK PREDICTION METHOD FOR FAST PAGERANK CONVERGENCE.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 175-180
DOI: 10.5220/0001521701750180

experiments. The Web is highly dynamic, the birth rate of new pages is fast and old pages disappeared soon. Besides, we found that the pages in the same host are extremely link to each other. These results imply that the link structure in the same host may be quite similar and the directories vice versa. If we sort the PageRank of Web pages in a directory and draw each dot of PageRank, we can observe that the PageRank values would form several clusters. The pages in the same cluster, which have similar link structure and thus the numbers of in-link are similarly. Thus, the similarly linking relation pages would cause the similar PageRank and would be a cluster after sorting the PageRank of pages.

In our previous work, (Kao and Lin, 2007) predicts the “true importance score” of pages in the future that based on the clustering feature of PageRank in a directory. The PageRank of a page at different previous time stages is growth in the cluster. Thus, the prediction of PageRank at next time stage could be the average PageRank of the cluster, which this page belongs to. In this paper, we modify the original prediction algorithm to give a more precise prediction. In our experiments, we show that the augmented prediction algorithm can reduce the relative error of prediction effectively under the cases, which the original method can be not covered.

2 RELATED WORK

There have been many researchers who investigated the Web search engines for a long history. The Information Retrieval (IR) community had proposed many outstanding algorithms to match documents for a given query. They analyze the content of the documents to find the best matching results. Authors in (Salton and McGill, 1983) provide an overview of these traditional works.

A number of researchers have investigated the link structure of the Web and discovered how to utilize it to improve the search results. Also they have proposed various ranking metrics. Major search engines are used the PageRank algorithm. Works in (Abiteboul et al., 2003) (Kamvar et al., 2003) provided the different ways to improve PageRank computation. Authors in (Haveliwala 2002) study how to personalize PageRank by giving different weights to pages. Work in (Xing and Ghorbani, 2004) shows that we can get a better search results by considering another weighting function to link. Authors in (Jiang et al., 2004) found that dividing the Web into different blocks and assigning different weights to different blocks based on some principles can achieve a better performance of PageRank search results. Authors in (Xue et al., 2005) discover the inherent property of the Web and then propose a novel ranking method called Hierarchical Rank to re-estimate the PageRank of pages. Authors in (Yates et al., 2002) propose a new method to calculate page importance by considering the last modified time and thus it could treat newly created pages equitably. Authors in (Eiron and McCurley, 2003) (Kumar et al., 2000) also investigate the properties of the Web.

2.1 Page Quality

Cho et al. (Cho et al., 2005) proposed a new point of view to explain the meaning of PageRank. They believe that the users determine the PageRank score of Web pages. The quality estimator is listed in formula (1):

$$\hat{Q}(p,t) = I(p,t) + P(p,t) = \left(\frac{n}{r}\right)\left(\frac{dP(p,t)/dt}{P(p,t)}\right) + P(p,t) \qquad (1)$$

where $\hat{Q}(p,t)$ means the page quality of page *p* at time *t*, $I(p,t)$ and $P(p,t)$ represent the increasing popularity and popularity of page *p* at time *t*, respectively. Moreover, n is the total number of Web users and r is normalization constant. In practice, however, we cannot obtain the time derivative immediately, but only can be approximated through the increase of PageRank at different time points. In other words, we utilize the PageRank score at discrete time points to reach the goal of anticipation. Formula (1) is modified as the following:

$$\hat{Q}(p,t_i) = \frac{n}{r}\left(\frac{\Delta PR(p,t_i)/\Delta t_i}{PR(p,t_i)}\right) + PR(p,t_i) \qquad (2)$$

where $PR(p,t_i)$ is the PageRank of *p* at time *t*,

$$\Delta PR(p,t_i) = PR(p,t_i) - PR(p,t_{i-1})$$

and $\Delta t_i = t_i - t_{i-1}$.

In their model, the quality of pages is composed of the “increasing popularity” and the “page popularity at current time”. Then the quality of a page in the long run comes to a stable value. Figure 1 shows the evolution of page popularity. [2] The

[2] This plot is cited from http://www.seoresearcher.com/popularity-ranking-faults.htm, also introduced in (Cho and Roy, 2004). The original source is experimentally observed in the site popularity evolution data collected by web ranking companies.

horizontal axis is the increasing of the time and the vertical axis is the value of page popularity. We can observe that a page generally goes through three stages before it becomes mature. After the maturely stage, the page popularity stabilizes at a certain value. The basic idea of page quality is to alleviate the inherent bias of PageRank algorithm.

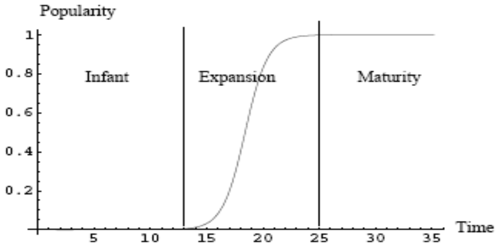

Figure 1: Time evolution of page popularity.

3 THE PROPOSED METHOD

It is important to know that how much time it takes for a page to become a popular page (assume it is a high quality page) and returned by search engines at the top of search results. Cho et al. (Cho and Roy, 2004) analyze two Web user models, i.e., the random-surfer model" and "the search dominant model". The former means that users browse the Web pages directly. The later is opposite to the former, the users only use search engines to browse the Web pages. They found that the search dominant model takes about 66 times slower than the random-surfer model for a page to become popular. This shows that the search engines indeed dominate the users' browsing behavior.

We calculated the PageRank values of pages in our dataset. Figure 2 and Figure 3 show the ranking distribution of pages in different directories. In a small directory, such as Figure 2, there is often a 2-cluster graph and the lower cluster is usually long and straight. In a large directory, such as Figure 3, we can observe three (or four) clusters in this graph.

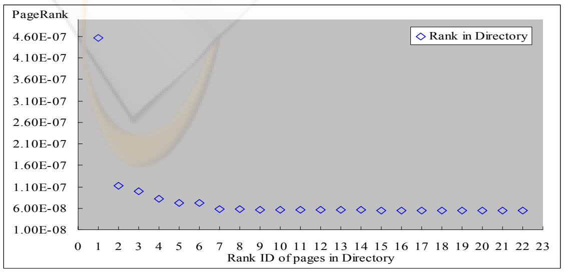

Figure 2: Ranking distribution of pages in a small directory.

It is expectable that the clustering of PageRank values in a directory since pages within the same directory may have similar contents. Thus, the in-link numbers of them are alike, because the pages in the same cluster are referred to the pages with similar contents.

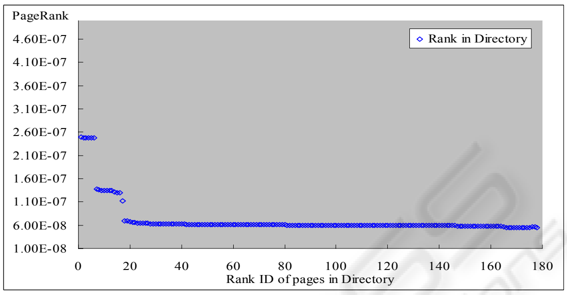

Figure 3: Ranking distribution of pages in a large directory.

3.1 Augmented Directory Feature based Ranking

The DRank algorithm in our previous work (Kao and Lin, 2007) can reduce the inherent bias of PageRank algorithm on the newly pages. It calculates the variation of PageRank of a page in the directory, and it utilizes the PageRank of Web pages obtained at different time stages. It observes that the PageRank of pages are close to within cluster of pages. The algorithm of DRank is stated below.

Assume that P_i is a newly page located within a directory. In order to predict the DRank value of *Pi* at next time stage, we divide the process of prediction into two steps: (1) we calculate the Page Quality of P_i at time T_2, and then (2) check if there is exist any clusters close to Page Quality of P_i. If it is true, we set DRank value of P_i to the average PageRank value of the nearest cluster, or we set DRank value of P_i to the Page Quality of P_i. The cluster extraction is based on the clusters in directories at time stage T_2. Since the web is highly dynamic, the link structure of directories might be quite different at time stage T_3. The variations of link structure in directories would affect our prediction accuracy.

Since the Page Quality involves in the part of prediction results in DRank algorithm. It could be controversial, while we compare the prediction accuracy between the Page Quality and the DRank. We modify the original DRank algorithm to a new one. The DRank can completely get rid of the part of prediction results, which are equal to the prediction of Page Quality. We call the original DRank algorithm as $DRank_0$ and the modified one as DRank. The detailed algorithm is as the following:

1. Firstly, we cut the ladders and search for clusters in the range between $PR_2 - \alpha\times\Delta$ and $PR_2 + \alpha\times\Delta$.
2. Secondly, check if there is exist any adoptable clusters close to PR_2 of P_i
   A. If true, set DRank value of P_i to the average PageRank value of the nearest cluster.
   B. Else, set DRank value of P_i for PR_2 of P_i.

It is different from $DRank_0$ that DRank starts the search scope at PR_2 while $DRank_0$ starts at Q_2. In our experiment results, we will show that the relative prediction error of DRank will be better than $DRank_0$.

We define $\Delta = | PR_2 - PR_1 |$ where PR_i means the PageRank of page P_i at time stage T_i and α is a constant parameter which determines the cluster search scope in the range of α×Δ. Figure 4 illustrated the process of cluster extraction. These two solid red lines illustrated the prediction range of our previous alogirthm. If there is a cluster in the scope of α×Δ, DRank will predict the average PageRank of the cluster in the scope of α×Δ. In this figure, N is another constant parameter to measure if the cluster in the range of α×Δ can be adopted or not. S is a dynamic parameter used to cut the ladders dynamically. In our experiments, we set α from 1 to 5 and N from 0.1 to 0.5.

Our cluster extraction method is referred to one of the well-known cluster algorithm: Statistical Information Grid (STING (Wang et al., 1997)) in Grid-Based Methods. STING is a hierarchical statistical information grid based approach for spatial data mining. Although the detail process is a little different from the STING, we refer to its layer cutting for our cutting ladder method and its "confidence interval" for determining an adoptable cluster.

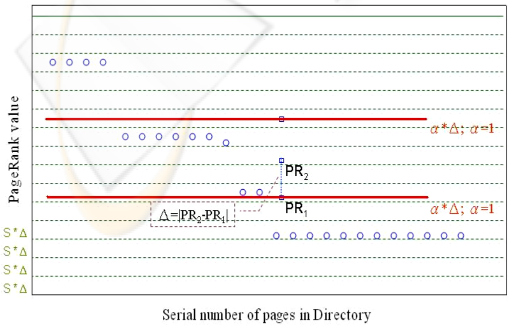

Figure 4: An example of the nearest cluster extraction.

Figure 5 shows an example of DRank algorithm. There are three clusters in this distribution. The target page is close to the center cluster at time stage T2. However, the upper cluster is closer to Q_2 and thus DRank could assign the average PageRank of upper cluster as the predictive PageRank at time stage T_3.

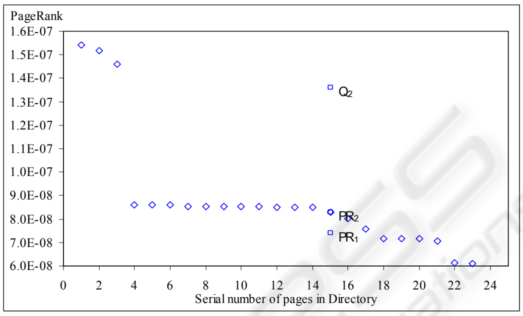

Figure 5: An example of DRank algorithm.

4 EXPERIMENT

In this section, we verify our proposed method and test the performance by some experiments performed on our dataset. We follow the crawling policy set in (Cho et al., 2005) to download the Web pages. We select about 35 different entry pages for the web crawlers. The entry pages are selected from the Google Directory (http://directory.google.com/). We downloaded pages from each entry pages until the maximum of 200,000 pages or we could not reach any more pages. We crawled three snapshots in three months. For each snapshot, the total pages are between 4.6 million and 4.85 million. Out of 4.85 million pages, there are 1.43 million pages were common in all three snapshots. Out of 1.43 million pages, there are only 0.33 million pages were PageRank growing pages (PR_2>PR_1). It is hard to define the newly pages, so we choose the PageRank growing pages to predict. An intuitive notion is that we consider the PageRank growing pages as the newly pages.

The prediction accuracy of the future PageRank is based on the average relative "error". We compare our methods with Page Quality and DRank. The standard formula of relative error is as the following:

$$Err(p)\,of\ t_i = \begin{cases} \left|\dfrac{PageRank(p,t_i) - Q(p,t_{i-1})}{PageRank(p,t_i)}\right| & for\ PageQuality(p,t_i) \\ \left|\dfrac{PageRank(p,t_i) - DRank_0(p,t_{i-1})}{PageRank(p,t_i)}\right| & for\ DRank_0(p,t_i) \\ \left|\dfrac{PageRank(p,t_i) - DRank(p,t_{i-1})}{PageRank(p,t_i)}\right| & for\ DRank(p,t_i) \end{cases} \qquad (3)$$

The err(p) function is a standard formula of prediction performance cited from (Cho et al., 2005).

4.1 Experiment Results

Since our dataset is raw. It could be hard to compare the prediction accuracy of DRank$_0$ with DRank. We analyze our dataset and partition it into several subsets, which have the same features. We define the growing pages as $PR_3 \geq PR_2$ and the downside pages as PR$_3$<PR$_2$.

In Table 1, there is only 21.1% ratio of pages are growing pages. Most of pages, their PageRank will descend at time stage T_3. Additionally, since our method is mainly focus on the directory feature of pages in a directory, the directory size is an important factor impacting on our method. Thus, we partition the dataset into three small datasets as shown in Table 2.

Table 1: Ratio of growing and downside pages at T$_3$.

	Page	Ratio
Growing pages	69,924	21.1%
Downside pages	261,810	78.9%

Table 2: The partition of our data set.

Directory	Director	Page number	Ratio
Small	1~10	195,469	58.9
Middle	11~100	92,898	28.0
Large	> 100	43,367	13.1

There are six small datasets selected from the initial dataset. They are the growing pages, the downside pages, the random selected pages, and the pages in small size directory, the middle size directory and large size directory. For each dataset, we randomly choose 1000 pages to predict their PageRank at time stage T_3. In DRank$_0$ algorithm, since Page Quality would involve in part of prediction results. Here, we set the parameter n/r of Page Quality to 0.0000005, which is the best setting in DRank$_0$ algorithm (Kao and Lin, 2007).

In Figure 6, the prediction error in Page Quality and DRank$_0$ is good. However, our DRank algorithm still can perform better. We believe that if a page grows stably, its PageRank will not also grow rapidly at next time stage. That is why DRank can obtain better prediction accuracy than DRank$_0$.

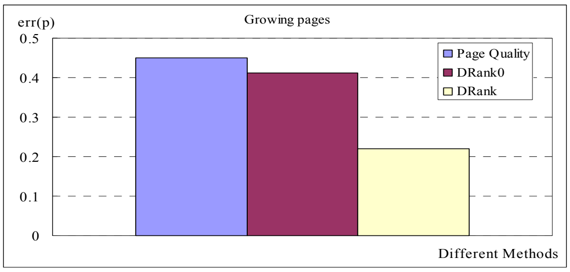

Figure 6: Err(p) in the growing pages.

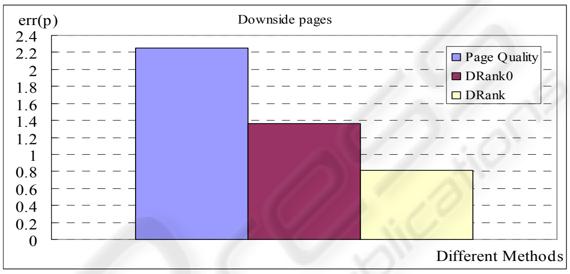

Figure 7: Err(p) in the downside pages.

In Figure 7, the relative err(p) of DRank is even better than DRank$_0$. In DRank$_0$, Page Quality would involve in part of prediction results. However, Page Quality can only perform well in growing pages. In downside pages, the distance between PR$_3$ and Q$_2$ is larger than PR$_2$ and Q$_2$. Thus, the part of DRank prediction results equaling Q$_2$ will increase the relative error.

5 CONCLUSIONS

Considering the cluster phenomenon in a directory and the link structure of Web, we can measure and predict the true importance of a page. We propose a modified directory feature based on ranking algorithm to reduce the relative error while predicting the PageRank at the next time stage. To summarize briefly, the proposed augmented DRank algorithm, i.e., DRank+, can significantly improve the prediction accuracy. The Drank+ can give a more accurate prediction under cases while original DRank can not perform well. It can reduce the bias on the newly pages effectively with predicting their convergent PageRank value.

ACKNOWLEDGEMENTS

The authors are supported in part by National Science Council Project No. NSC 95-2221-E-006-229-MY2, Taiwan, Republic of China and the Innovative Digitech-Enabled Applications & Services Institute (IDEAS) of Institute for Information Industry, Taiwan, Republic of China.

REFERENCES

Abiteboul, S., Preda, M., and Cobna, G., 2003. Adaptive on-line page importance computation. In *Proceedings of the International World-Wide Web Conference*.

Brin, S. and Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW Conference*.

Cho, J., Roy, S. and Adams, R. E., 2005. Page Quality: In Search of an Unbiased Web Ranking. In *Proc. of the SIGMOD Conference*.

Cho, J. and Roy, S., 2004. Impact of Search Engines on Page Popularity. In *Proceedings of the International World-Wide Web Conference*.

Eiron, N. and McCurley, K. S., 2003. Locality, Hierarchy, and Bidirectionality on the Web. In *Workshop on Web Algorithms and Models*.

Haveliwala, T. H., 2002. Topic-sensitive pagerank. In *Proceedings of the International World-Wide Web Conference*.

Jiang, X. M., Xue, G. R., Zeng, H. J., Chen, Z., Song, W.-G. and Ma, W.-Y., 2004. Exploiting PageRank Analysis at Different Block Level. In *Proceedings of Conference of WISE*.

Kamvar, S., Haveliwala, T., Manning, C., and Golub, G., 2003. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the International World-Wide Web Conference*.

Kumar, R., Raghavan, P., Rajagopalan, S. and Sivakumar, D., 2000. Stochastic models for the Web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*.

Kao, H.-Y. and Lin, S.-F., 2007. A Fast PageRank Convergence Method based on the Cluster Prediction. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*.

Salton, G. and McGill, M. J., 1983. Introduction to modern information retrieval. McGraw-Hill.

Wang, W., Yang, J., Muntz, R., 1997. STING: A Statistical Information Grid Approach to Spatial Data Ming. In *Proceedings of the 23rd VLDB Conference*.

Xing, W. and Ghorbani, A., 2004. Weighted PageRank Algorithm. In *Proceedings of the 2nd Annual Conference on Communication Networks and Services Research*. 305-314.

Xue, G.-R, Yang, Q., Zeng, H.-J., Yu, Y., Chen, Z., 2005. Exploiting the Hierarchical Structure for Link Analysis. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yates, R. B., Castillo, C. and Jean, F. S., 2002. Web Dynamics, Structure, and Page Quality. In *Proceedings of SPIRE Conference*.