Link Analysis using Time Series of Web Graphs

Lei Yang[1,2], Lei Qi[1,3], Yan-Ping Zhao[2], Bin Gao[1], Tie-Yan Liu[1]

[1]Microsoft Research Asia
Sigma Center, No. 49, Zhichun Road
Beijing, 100080, P. R. China
{bingao, tyliu}@microsoft.com

[2]Department of Computer Science
Beijing Institute of Technology
Beijing, 100081, P. R. China
jeffy_yanglei@hotmail.com
zhaoyp@bit.edu.cn

[3]Academic Talent Program
Tsinghua University
Beijing, 100084, P. R. China
qil04@mails.tsinghua.edu.cn

ABSTRACT

Link analysis is a key technology in contemporary web search engines. Most of the previous work on link analysis only used information from one snapshot of web graph. Since commercial search engines crawl the Web periodically, they will naturally obtain time series data of web graphs. The historical information contained in the series of web graphs can be used to improve the performance of link analysis. In this paper, we argue that page importance should be a dynamic quantity, and propose defining page importance as a function of both PageRank of the current web graph and accumulated historical page importance from previous web graphs. Specifically, a novel algorithm named TemporalRank is designed to compute the proposed page importance. We try to use a kinetic model to interpret this page importance and show that it can be regarded as the solution to an ordinary differential equation. Experiments on link analysis using web graph data in five snapshots show that the proposed algorithm can outperform PageRank in many measures, and can effectively filter out newly appeared link spam websites.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Performance

Keywords

Search engine, temporal information, page importance, link analysis, PageRank

1. INTRODUCTION

The Web has grown exponentially over the past decade. However, at the same time, the number of low quality websites is also increasing rapidly which brings many problems for users and search engines. To tackle this problem, employing an effective measure of page importance becomes very critical. Much work has been done on the problem of defining and calculating page importance by analyzing the link structure of the Web, such as HITS [8], PageRank [3][11], and the work in [2][6][7][9][10]. PageRank is a good measure of page importance. However, latest study [12] showed that PageRank can be easily cheated by some special techniques. Furthermore, it is not robust to link spam [4][5].

CIKM'07, November 6-8, 2007, Lisboa, Portugal.

To solve the above problem, we believe that effective measure for page importance should be defined on many snapshots of Web graphs over time. Spam pages usually have irregular patterns of link changes which make high-quality and spam pages quite different. There has been some work on adding historical information into link analysis. Berberich *et al* [1] developed the T-Rank algorithm, a link analysis method that takes into account the temporal aspects: freshness and activity of pages and links. Yu *et al* [13] proposed a time-weighted PageRank, in which the in-links of a page are weighted according to their timestamps. The advantage of the methods is that they extracted and utilized the temporal information, but they did not fundamentally change the framework of PageRank which uses one snapshot of web graph.

In this paper, we postulate that the calculation of page importance should not be a static process. Instead, one should define a dynamic process that computes page importance from two perspectives: the importance from the current web graph and the accumulated historical importance from previous web graphs. Therefore, conventional PageRank is embodied in the new importance score. We then develop an efficient algorithm named TemporalRank to compute the new page importance, and show that it can be interpreted using a kinetic model and solved using a differential equation. Experiments on a set of real-world web graphs show that this algorithm can be more effective in measuring page importance than traditional methods, and can detect and filter out link spam.

2. TEMPORALRANK

In this section, we propose a framework to calculate page importance based on a series of web graphs, and use a kinetic model to explain the relationship between the so-calculated page importance and PageRank in the physical viewpoint.

2.1 The Novel Framework

Based on the discussions in Section 1, we propose evaluating the page importance using both the current web graph and historical information contained in previous web graphs. We call the so-calculated page importance by TemporalRank. Denoting the TemporalRank of page i in the t-th web graph by $TR_t(i)$, $t = 1,\ldots,k$ we have the following formula.

$$TR_k = (1-\beta)H_k(i) + \beta PR_k(i) \quad (1)$$

Here $PR_k(i)$ is the PageRank of page i calculated from the current web graph G_k, corresponding to the current page importance; $H_k(i)$ is the accumulated historical importance information obtained

from the previous web graphs G_1, G_2, … , G_{k-1}; β ($0<\beta<1$) is a weighting parameter to balance how much we trust in the present and how much we trust in the history. The bigger β is, the more trust we have in the link information in the current snapshot.

The next step is to define $H_k(i)$. A simple approach is to linearly combine the PageRank scores of page i in the previous web graphs as follows.

$$H_k(i) = \sum_{t=1}^{k-1} \gamma_t PR_t(i) \tag{2}$$

Here γ_t, $t = 1,\ldots, k\text{-}1$ are decaying factors indicating how much the importance from each snapshot contributes to the overall page importance. In common, the earlier a snapshot is, the less amount of importance it should contribute to $H_k(i)$. Therefore, γ_t should be a group of decreasing factors from $t = k\text{-}1$ to $t = 1$.

Combining formulas (1) and (2), we have,

$$TR_k(i) = (1-\beta)\sum_{t=1}^{k-1} \gamma_t PR_t(i) + \beta PR_k(i) . \tag{3}$$

From formula (2), we can see that TemporalRank can be calculated by the linear combination of the PageRank scores from a series of web graphs. How to define proper weighting factors γ_t is a key problem in (3). We will elaborate on this in the next sub-section.

2.2 Kinetic Model

In this sub-section, we use a kinetic model to interpret the TemporalRank proposed in the previous sub-section, and then derive rational weighting parametersγ_t, $t = 1,\ldots, k\text{-}1$, from this model for the linear combination.

We suppose the page importance corresponds to the velocity of an object, driven by some kind of virtual force. If a page i gets a PageRank score $PR_t(i)$ in the current snapshot G_t, a corresponding driving force will be added to the virtual force $F_t(i)$ for this page. That is, the current PageRank score is a positive effect for the virtual force. At the same time, the *decay* of the TemporalRank score $TR_t(i)$ turns to be a resisting force which is a negative effect for this virtual force. In summary, we can use the following equation to model the above kinetic system.

$$F_t(i) = -\lambda TR_t(i) + \eta PR_t(i) \tag{4}$$

Here η ($\eta > 0$) is an enhancement constant describing the relation between the driving force and the $PR_t(i)$, and λ ($\lambda \geq 0$) is the decaying factor.

Actually equation (4) is common in many kinetic systems. For example, suppose there is a car running in the highway with the velocity v, then the resisting force R it suffers is,

$$R = -\lambda v . \tag{5}$$

Here λ is the damping factor for the resisting force, and the minus symbol means the direction of the resisting force is opposite to the velocity v. At the same time, the car has its driving force D, so its motion equation can be written as,

$$F = R + D, or\ F = -\lambda v + D . \tag{6}$$

Denoting the mass of the car as m, according to Newton's second law, we have

$$F = m\frac{dv}{dt} . \tag{7}$$

Based on the aforementioned analogy, we can get the relationship between TemporalRank and the virtual force as below.

$$F_t(i) = m(i)\frac{dTR_t(i)}{dt} . \tag{8}$$

Here $m(i)$ is some intrinsic quality of web page i with similar meaning to that of mass.

Combining the equations in (4) and (8), we get a first-order ordinary differential equation (ODE),

$$\frac{dTR_t(i)}{dt} + \frac{\lambda}{m(i)} TR_t(i) = \frac{\eta}{m(i)} PR_t(i) . \tag{9}$$

This ODE can be easily solved and its general solution is,

$$TR_k(i) = e^{-\frac{\lambda}{m(i)}k}\left[\int_0^k \frac{\eta}{m(i)} PR_t(i) e^{-\frac{\lambda}{m(i)}t} + C_0\right], \tag{10}$$

where C_0 is the integral constant.

Suppose that all the pages have the same initial page importance score at the beginning ($t = 0$). That is, $TR_0(i) = \frac{1}{N}$, where N is the number of web pages in the graph. Then the solution (10) turns to be,

$$TR_k(i) = \frac{1}{N} e^{-\frac{\lambda}{m(i)}k} + \frac{\eta}{m(i)} \int_0^k PR_t(i) e^{-\frac{\lambda}{m(i)}(k-t)} dt . \tag{11}$$

Since the temporal web graph data is discrete with respect to time, we reduce the above solution to its discrete form as the following.

$$TR_k(i) = \frac{1}{N} e^{-\frac{\lambda}{m(i)}k} + \frac{\eta}{m(i)} \sum_{t=1}^{k} PR_t(i) e^{-\frac{\lambda}{m(i)}(k-t)} . \tag{12}$$

Till now, we get a formula (12) to calculate the TemporalRank score of page i. It is easy to see that this formula is a linear combination of the PageRank scores of page i in a series of web graphs. If we write the solution in another form as below,

$$TR_k(i) = \frac{1}{N} e^{-\frac{\lambda}{m(i)}k} + \frac{\eta}{m(i)} \sum_{t=1}^{k-1} PR_t(i) e^{-\frac{\lambda}{m(i)}(k-t)} + \frac{\eta PR_k(i)}{m(i)} . \tag{13}$$

we will see that the first item on the right-hand side of formula (13) is a constant with respect to page i, representing the initial page importance; the second item is the linear combination of the PageRank scores of page i in previous web graphs, corresponding to $H_k(i)$; and the third item contains the PageRank of page i in the current web graph. This is consistent with what we have discussed in Section 2.1. Comparing (13) with (3), and ignoring the constant, we can get the combination coefficients as follows.

$$\gamma_t = \frac{\eta}{m(i) - \eta} e^{-\frac{\lambda}{m(i)}(k-t)} , \text{ and } \beta = \frac{\eta}{m(i)} . \tag{14}$$

From the above derivations, we can see that the kinetic model can well interpret the novel framework for TemporalRank proposed in

Section 2.1, and help us map the weighting coefficients γ_t and β to some more meaningful parameters λ, η, and $m(i)$.

According to the discussions in the previous subsections, we summarize the proposed TemporalRank algorithm in Table 1.

Table 1. The TemporalRank algorithm

Input: k successive web graphs $G_1, G_2, \dots, G_k$

Output: TemporalRank score vector TR_k for the k-th web graph

1) Get the transition matrices $A_1, A_2, \dots, A_k$ from web graphs $G_1, G_2, \dots, G_k$ and compute their PageRank vectors $PR_1, PR_2, \dots, PR_k$.
2) Set the self-decay constant λ, the enhancement constant η, and the intrinsic quality factor m for each page i existing in the k web graphs.
3) For each page i in the dataset, set the initial score $TR_0(i) = e^{-\frac{\lambda}{m}k}$ and $t = 0$.
   a. For $t = 1$ to $t = k-1$, add the historical importance score accumulatively, $TR_t(i) = TR_{t-1}(i) + \frac{\eta}{m} PR_t(i) e^{-\frac{\lambda}{m}(k-t)}$.
   b. Add the current importance score, $TR_k(i) = TR_{k-1}(i) + \frac{\eta}{m} PR_k(i)$.
   c. Output $TR_k(i)$.

3. EXPERIMENTAL RESULTS

3.1 Datasets

The dataset used in our experiments is a sub-graph sampled from a commercial search engine which contains five snapshots (G_1, G_2, … , G_5) crawled in the first half of year 2006, and each snapshot consists of around 30 million web pages and over 2000 million hyperlinks. Table 2 shows the basic information. The changes of the number of web pages between each two successive snapshots are summarized in Table 3.

Table 2. The number of web pages in each snapshot

Snapshot	# of web pages
G_1	31,464,052
G_2	31,062,569
G_3	30,361,077
G_4	29,510,456
G_5	28,955,134

Table 3. The changes of the number of web pages between each two successive snapshots

	# of new pages	# of disappeared pages
$G_1 - G_2$	2,952,105	3,353,588
$G_2 - G_3$	2,830,435	3,531,927
$G_3 - G_4$	2,838,130	3,688,751
$G_4 - G_5$	2,916,245	3,471,567

3.2 Evaluation Criteria

To evaluate the performance of the TemporalRank algorithm, we use the following three criteria: toolbar correlation (TC), click-through correlation (CC), and spam bucket distribution.

A part of users agreed to use the toolbar software offered by the commercial search engine which will log the URLs of the pages when these users surfed on the Web. There are 16,737,629 pages in the sampled sub-graphs that could be found in the toolbar log data, and we counted the click numbers of these pages as the ground truth for the evaluation. Consider the following two vectors: the ranking vector x with 16,737,629 elements outputted by the link analysis algorithm, and the toolbar click vector y which elements record the toolbar click numbers of the pages in the corresponding positions of x. The toolbar correlation (TC) is defined in (15) as the correlation between vector x and vector y, where M is the length of the vectors x and y.

$$TC = \frac{\frac{1}{M}\sum_{i=1}^{M} x_i y_i - (\frac{1}{M}\sum_{i=1}^{M} x_i)(\frac{1}{M}\sum_{i=1}^{M} y_i)}{\sqrt{\left[\frac{1}{M}\sum_{i=1}^{M} x_i^2 - (\frac{1}{M}\sum_{i=1}^{M} x_i)^2\right]\left[\frac{1}{M}\sum_{i=1}^{M} y_i^2 - (\frac{1}{M}\sum_{i=1}^{M} y_i)^2\right]}} \quad (15)$$

The search engine also logged the click-through data indicating which webpage was clicked after users submitting their queries to this search engine. There were 7,067,915 pages in the sub-graphs that could be found in the click-through data. Once again, their click numbers were used as the ground truth for evaluation. We use the same method with toolbar correlation to get the click-through correlation.

In common, a page will get more clicks if it is more important. Also for link analysis methods, the higher the ranking score of a page is, the more important this page is regarded by the link analysis method. Therefore, TC and CC can be good indicators for the performance of link analysis algorithms.

We also use spam bucket distribution in our experiments to validate this. With PageRank, a list of pages in the decent order according to their scores can then be generated. We divide the list into a group of buckets, in each of which there are a certain number of web pages. Given a set of labeled link spam pages, we can count the number of spam pages that fall in each bucket. Here we asked five well-trained experts to label a subset of the web graph and got 11,357 link spam pages. If a link analysis method has fewer spam pages in the top buckets than other methods, it will be regarded as a better approach.

3.3 Results

3.3.1 Evaluated by Correlations

To understand how these factors in our algorithm will affect the performance, we generate the following datasets as shown in Table 4 for our experiments.

Table 4. The datasets containing different numbers of graphs

Name	Contained web graphs
S1	G_1, G_2, G_3, G_4, G_5
S2	G_2, G_3, G_4, G_5
S3	G_3, G_4, G_5
S4	G_4, G_5
S5	G_5

We ran the TemporalRank algorithm with $m = 1$ on the five datasets S1 to S5 with different values of λ (0, 0.001, 0.01, 0.1, 1, 10), and plot the corresponding performance on TC and CC in Figure 1 and Figure 2. To be noted, since S5 only contains one single snapshot of Web graph and m = 1, the performance on this dataset is exactly the same with the PageRank algorithm, which is

independent of the decaying constant. From these figures, we can see that the performances of TemporalRank on datasets S1 to S4 are all better than that on S5 (corresponding to PageRank), no matter what value λ takes. This shows the superior of our proposed method to PageRank.

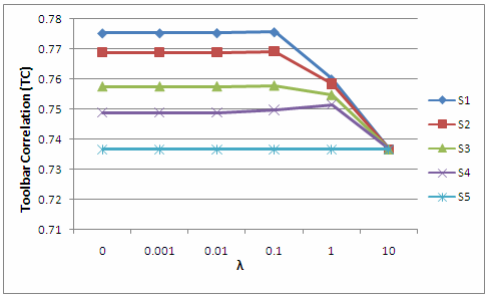

Figure 1. The performance on toolbar correlation

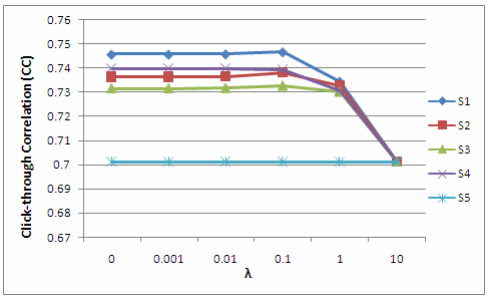

Figure 2. The performance on click-through correlation

3.3.2 Evaluated by Spam Bucket Distribution

As mentioned above, we then check the ranking results produced by the TemporalRank method, to see how those spam pages are ranked. Specifically, we plot the spam bucket distribution in Figure 3, where each bucket contains one million web pages (only the top 20 buckets are shown) to investigate how many of the 1,502 spam pages are there in these buckets. Once again, the result for the S5 dataset actually corresponds to the performance of PageRank since there is only one single web graph in this dataset. From this figure, we can see that the TemporalRank algorithm can depress the newly generated spam pages in a very effective manner as compare with PageRank. Furthermore, the more web graphs we use, the better performance we can achieve.

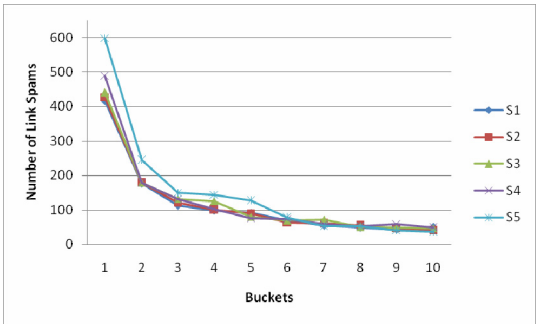

Figure 3. The spam buckets distributions

4. CONCLUSIONS

We propose a novel framework for link analysis on a series of web graphs, in which both the link structure in the current web graph and the historical importance information are considered as incentives for calculating page importance. We introduced a kinetic model to interpret this framework and develop an efficient algorithm within this framework. Experiments show that by leveraging the historical information, our proposed method can be much more effective than previous methods in ranking pages and removing link spam.

5. ACKNOWLEDGEMENTS

This work was performed when the first two authors were interns at Microsoft Research Asia. This research is supported by the National Science Foundation of China, the project code: 70471064, and Research Foundation of Beijing Institute of Technology, the project code: BIT-UBF-200308G10.

6. REFERENCES

[1] Berberich, K., Vazirgiannis, M., and Weikum, G.T-Rank: Time-aware Authority Ranking. In Algorithms and Models for the Web-Graph: Third International Workshop, WAW 2004, pages: 131–141, Springer-Verlag, 2004.

[2] Boldi, P., Santini, M., and Vigna, S. PageRank as a function of the damping factor. In Proceedings of the 14th International World Wide Web Conference, 2005.

[3] Brin, S., and Page, L. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International Wide Web Conference, Australia, 1998.

[4] Gyongyi, Z., and Garcia-Molina, H. Link spam alliances. Technical Report, Stanford University, 2005.

[5] Gyongyi, Z., and Garcia-Molina, H. Web spam Taxonomy. In the First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[6] Haveliwala, T. Topic-sensitive PageRank. In Proceedings of the International World Wide Web Conference, 2002.

[7] Haveliwala, T., Kamvar, S., and Jeh, G. An analytical comparison of approaches to personalizing PageRank. Technical Report, Stanford University, 2003.

[8] Kleinberg, J. Authoritative sources in a hyperlinked environment. In Journal of the ACM, 46(5):604-632, 1999.

[9] Langville, A., and Meyer, C. Deeper inside PageRank. Internet Mathematics 1(3):335-380, 2004.

[10] McSherry, F. A uniform approach to accelerated PageRank computation. In Proceedings of the 14th International World Wide Web Conference, 2005.

[11] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford University, Stanford, CA, 1998.

[12] Richardson, M., Prakash, A., and Brill, E. Beyond PageRank: Machine Learning for Static Ranking. In Proceedings of the Fifteenth International World Wide Web Conference, pages: 707-715, 2006.

[13] Yu, P.S., Li, X., and Liu, B. Adding the Temporal Dimension to Search - A Case Study in Publication Search. In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005.