Upgradation of PageRank Algorithm based upon Time Spent on Web Page and its Link Structure

Amit Kelotra
Assistant Professor
Department of Computer Science
SVKM`S NMIMS MPSTME Shirpur, Maharashtra

ABSTRACT
Page Ranking holds great importance in any information retrieval system. We are well aware of the fact that the World Wide Web boasts a vast array of pages. It becomes the duty of search engines to provide the most relevant web pages to the user. The PageRank is one approach to rank web pages. However, it lays more stress on link structure of a page. Hence, more parameters need to be accommodated in the already suggested algorithm. This will only make it more efficient. In this paper, a time-based approach is proposed as an extension to PageRank and is defined incrementally.

Keywords
PageRank, Link Structure, Time based Approach, Web Structure Mining

1. INTRODUCTION
The web remains to be the largest repository of data. Each day this repository witnesses a massive increase in the number of web pages. One cannot argue with the fact, that it has easily transformed itself to become one of the most popular and informative way of communication.

The random surfer uses a search engine to get specific information on the web. Search engines have thus, become an integral part of a user's surfing habit [1]. There are various hyperlinks on a web page which aid the user to move to a particular page until they have found what they are looking for.

These embedded links between related pages, data usage of a particular web page can be used by data mining techniques to help obtain relevant information and also assist in efficient browsing of web site. They could be referred to as trails to other web pages. The web graph structure is shown in figure 1 which contains hyperlinks as edges between pages and the various web pages as nodes. Hence, the link structure consists of data such as user's interests, web contents etc.

Web Structure Mining uses data mining techniques to help in resource finding, processing of information, retrieval of relevant pages and generalization [2]. The contemporary methods of this kind of mining lay weightage on the link structure of a web page to rank them. Search engines have deployed their own strategies to rank web pages. Several ranking algorithms have been proposed in literatures. They been thoroughly tested, with their strengths and limitations duly noted down and suggestions for their future enhancements have been proposed [3] [4]. This will only lead to development of a more reliable ranking algorithm. Some of the very famous approaches which use link structure analysis include Google PageRank formula [5] and Hypertext Topic Induced Search (HITS) algorithm [6]. PageRank uses in links to a page and out links from a page to rank its web pages whereas HITS depends upon hubs and authoritative framework.

Many researchers have come forward and proposed their own improvements to the existing PageRank formula since it focuses only on the link structure. In [7], Richardson et al. discussed a method to rank pages using content of pages and query term the user is looking for. In [8], Ben Choi et al. came out with three new factors to rank web pages, namely, keyword popularity, keyword to web page popularity, and web page popularity. In [9], Diligenti et al. proposed a unified probabilistic framework for ranking web page. In [10], Zhou Cailan used user clicks and click time of result pages as the major factors for his proposed algorithm.

In this paper, we introduce our own methodology to rank pages. We have used the existing PageRank formula and upgraded it by adding a time factor to it. This time factor will track the access duration of a user on a particular page and use its link structure to come forward with the page rank. One can also be notified about the user's interest of a particular web page by recording the time spent on it.

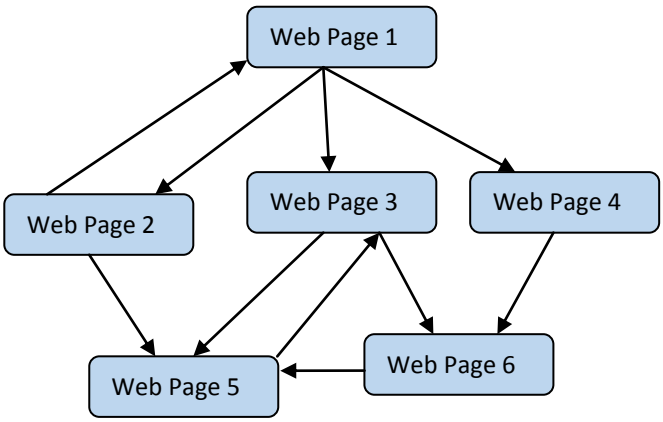

Fig. 1 A Web Graph

2. EXISTING PAGERANK ALGORITHM
Google search engine is the one which is most commonly used nowadays. It has deployed its own ranking formula known as, PageRank. Introduced by Larry Page and Sergey Brin, in 1998 [5], it is defined as a numeric value to determine the value of a web page. It takes into account all the incoming and outgoing links from a web site. The methodology presented behind this formula is as follows: Google considers that if a website has an incoming link from another web page, then that web page is casting a vote for it. Such incoming votes are summed up and a page rank is computed for that web site.

The PageRank algorithm is given by:

1. Calculate the page ranks of all pages by the following mathematical formula:

PR(A) = (1-d)/N + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

where,

PR(A) = PageRank of A

d = it is the damping factor whose values can be set between 0 and 1. However, it is usually taken as 0.85

N = number of nodes in web graph

PR(Ti) = it is the PageRank of pages Ti who have outbound links to page A

C(Ti)= it is the number of outbound links

that are present on pages Ti

This formula is iterative in nature. Each page is provided with a certain value at the beginning of computation and PageRanks of all subsequent pages are calculated in iterations.

We will now provide a simple example to get a better understanding of how PageRank is implemented.

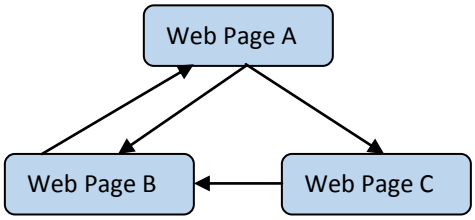

Fig 2. A simple Web Graph

Considering figure 2, we can clearly see three nodes, namely A,B and C. Node A has outgoing links to B and C, whereas, B has an outgoing link to A. Furthermore, C has an outgoing link to B. We have taken the damping factor, d as 0.85 for this implementation.

PR(A) = (1-0.85)/3 + 0.85*PR(B)

= 0.05 + 0.85*PR(B) (1)

PR(B) = (1-0.85)/3 + 0.85*((PR(A)/2)+PR(C))

= 0.05 + 0.85*((PR(A)/2)+PR(C)) (2)

PR(C) = (1-0.85)/3 + 0.85*(PR(A)/2)

= 0.05 + 0.85*(PR(A)/2) (3)

We have now obtained three linear equations. Solving (1), (2) and (3), we can have PageRanks as follows:

PR(A) = 686/1769 = 0.39

PR(B) = 703/1769 = 0.40

PR(C) = 380/1769 = 0.21

Thus, it is now safe to say that the Page Rank formula is completely a content independent algorithm. Moreover, it is not a whole number that is produced as a result, but a floating-point number is obtained. Proposed Improved PageRank Algorithm

3. PROPOSED IMPROVED PAGERANK ALGORITHM

3.1 Proposed Approach

In this paper, we propose an upgradation to the already existing PageRank formula. We know that the number of incoming or outgoing links from a particular page may change frequently. This can further lead to a change in user's interest for that web page. He might spend more time on that web page or can even lower his surfing time on it.

Therefore, one can know the amount of time spent by the user on that web page by studying this link structure.

3.2 Improved PageRank Algorithm

Each visit to a web page by a user is noted down. This will help in noting down user's behaviour and his interests while surfing. The time based approach helps to find the amount of time spent by a user on a web page.

We now present our PageRank formula which goes as follows:

PR(A) = (1-d)/N + d*((PR(T1)/C(T1))/T(A) + ... + (PR(Tn)/C(Tn))/T(A))+G

Here, the symbols have the following meaning:

- PR(A) = PageRank of A
- d = damping factor, usually taken as 0.85
- N = number of nodes in web graph
- PR(Ti) = it is the PageRank of pages Ti who have outbound links to page A
- C(Ti) = it is the number of outbound links that are present on pages Ti.
- T(i) = time spent on page
- G = it is calculated as:

  $(T(A)_{before} - T(A)_{now})/H*M*S$

  H = hours in a day

  M = minutes in an hour

  S= seconds in a minute

G is a factor which is introduced in order to computer PageRank within a day.

For its estimation, it becomes essential to keep note of all the visits that the user makes on a web page. This can further define user's interests and how much time he/she spends on a particular web page.

4. EXPERIMENT ANALYSIS

We have chosen ASP.Net as front end and SQL Server as the back end for the making of this application.

4.1 Evaluation of Time based Ranker

The IP address shown at the top of the interface (see Fig 3.) can help notify the user of a subsequent visit on a particular web page. If it's the very first visit on a web page, then the PageRank will be calculated according to the current time. However, if it is his second visit, then time for both the visits is taken into consideration.

The Visiting History of PageRank (see Fig 4.) can been by the user only for With Time Stamp calculations.

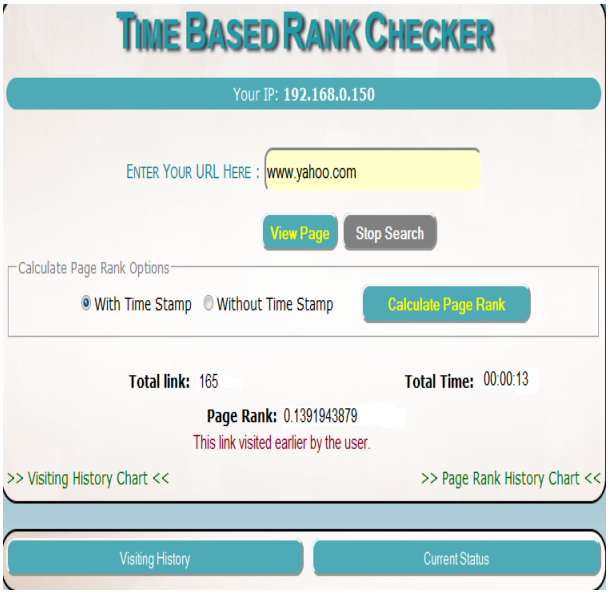

Fig 3. Interface to calculate PageRank

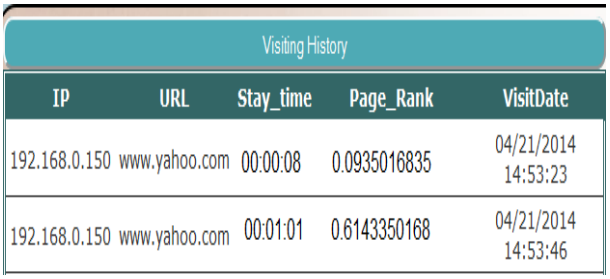

Fig 4. Visiting History of PageRank

4.2 Result Analysis and Discussion

Suppose, a user visits www.yahoo.com for 8 seconds on 21st April 2014, the PageRank comes out to be 0.0935016835.

During the user's next visit that is on the same day at a later time for 61 seconds there is an increase in the value of PageRank. It now changes to 0.6143350168. Hence, due to the user's increase in time spent on www.yahoo.com, the PageRank also suffers an increase in its value.

5. CONCLUSIONS AND FUTURE WORK

Time based Rank Checker clearly provides more efficient results than the one's provided by previous PageRank implementation. It helps one to understand the user's surfing habits. Another major advantage of this approach came out to be in the fact that if two pages, say A and B, have the same link structure and A is older than B. Google, in this case, will automatically assign A, a higher rank than B. However, our application places B ahead of A since B has been able to achieve that same link structure in less amount of time than A.

There is indeed some work that can be done in future so as to enhance this modified PageRank formula further since it has a limitation. The G factor that has been introduced in the new formula is considered for only 24 hours.

6. REFERENCES

[1] S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7 pp. 107-117, 1998.

[2] R. Kosala, H. Blockeel,, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol.2, No. 1, pp. 1-15, 2000.

[3] W. Xing, A. Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, pp. 305-314, 2004.

[4] N. Duhan, A. K. Sharma and K. Kumar Bhatia, "Page Ranking Algorithms: A Survey", In proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.

[5] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report, Stanford University, 1998.

[6] J. Kleinberg, "Authoritative Sources in a Hyper-Linked Environment", Journal of the ACM 46(5), pp. 604-632, 1999.

[7] M. Richardson, P. Domingos, "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank", Advances in Neural Information Processing Systems 14, pp. 1441-1448, Cambridge, MA: MIT Press, 2001.

[8] B. Choi, S. Tyagi, "Ranking Web Pages Relevant to Search Keywords", IADIS, International Conference WWW/Internet, pp. 200-205, 2009.

[9] M. Diligenti, M. Gori, M. Maggini, "Web Page Scoring Systems for Horizontal and Vertical Search", In Proceedings of the Eleventh International on World Wide Web, pp. 508-516, New York: ACM Press, 2002.

[10] Z. Cailan, C. Kai, L. Shasha, "Improved PageRank Algorithm Based on Feedback of User Clicks", Computer Science and Service System (CSSS), Nanjing, pp. 3949-3952, 2011.