www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 03 Issue 07 July, 2014 Page No. 6990-6995

Enhanced Page Rank Algorithm Using Time Factor

Monica Sehgal, Priya

B. S. Anangpuria Institute of Technology and Management

Email Id: monica.sehgal326@gmail.com

B. S. Anangpuria Institute of Technology and Management

Email Id: priyagupta2k6@gmail.com

Abstract

The Explosive development and the broad availability of the web has prompted surge of exploration movement in the zone of data recovery on the World Wide Web. Positioning has dependably been a vital part of any data recovery framework. If there should be an occurrence of web hunt its vitality gets basic. Because of the span of the web, it is basic to have positioning capacities that catch the client needs. To this end the Web offers a rich setting of data which is communicated through the hyperlinks. This paper presents the idea of page rank utilizing another Formula "Upgraded Page rank utilizing Time Component" which has careless limit as contrasted with Conventional Page Rank Algorithm.

Keywords: Ranking, Page Rank, Backlinks, Hyperlinks

1. Introduction

Searching on the World Wide Web is the second most frequent operation on the Web after e-mail. Therefore, it is important to have tools that perform search efficiently and effectively. Recently, much research has been devoted to creating better search engines for the Web. Even though there is a rich literature in the area of information retrieval[2], the Web, due to its size, and the diversity of the users that perform search, poses new challenges and problems.

In Google[7], the downloading of web pages (crawling) is done by various distributed crawlers. The URL server sends lists of URLs to be fetched to the crawlers. The fetched web pages are sent to the storeserver. The storeserver then compresses and stores the web pages into a repository. Each web page has a docID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer reads the repository, uncompresses the documents, and parses them. Each document is converted into a set of word occurrences called hits.

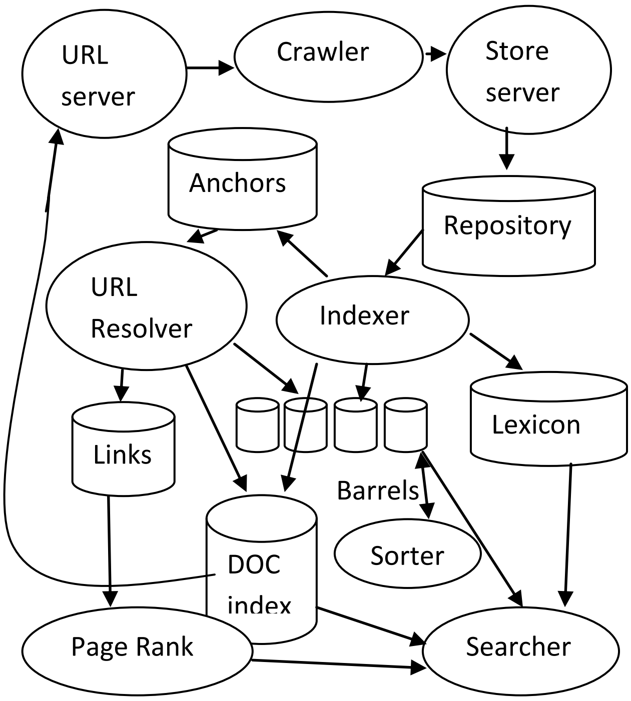


Figure 1 Shows the Google Architecture

The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link.

The URL resolver reads the anchors file and converts relative URLs into absolute URLs and in turn into docID. URL resolver puts anchor text into forward index, associated with the docID that the anchor points to and also generates a database of links which are pairs of docID. The links database is used to compute Page Ranks for all the documents[1][2].

1. Traditional Page Rank Algorithm

A. Basic Page Rank: Bringing order to the Web

Page rank (PR) is a quality metric invented by Google's owners Larry Page and Sergey Brin[5,6]. The Value 0 to 10 determines a page's importance, reliability and authority on the web according to Google[7]. This metric does not, however, directly affect a website's search engine ranking. A website with a PR 2 could be found on the first page of search results while a website with a PR 6 for the same keyword may appear on the second page of search results.

B. Algorithm

Page Rank is a "vote", by all the other pages on the web, about how important a page is. A link to a page counts as a vote of support. If there's no link there's no support (but it's only an abstention from voting rather than a vote against the page).

PageRank[5] calculated the probability that someone randomly clicking on links will arrive at a certain page. The more inbound links the page has from other popular pages, the more likely it is that someone will end up there purely by chance. Of course, if the user keeps clicking forever, they'll eventually reach every page, but most people stop surfing after a while. To capture this, PageRank also uses damping factor of 0.85, indicating that there is an 85% chance that a user will continue clicking on links at each page.

A probability is expressed as a numeric value between 0 and 1. A 0.6 probability is commonly expressed as a "60% chance" of something happening. Hence, a PageRank of 0.6 means there is a 60% chance that a person clicking on a random link will be directed to the document with the 0.6 Page Rank.

Now we refer PageRank as "PR"

$$PR(A) = (1 - d) + d\,(PR(T1)/(C(T1) + \ldots\ldots.)(PR(Tn)/C(Tn) + Prev_{rank}(A))$$

Here,

- PR(A) is the PageRank of Page A
- PR(Ti) is the PageRank of Pages Ti which link to Page A
- C(Ti) is the number of outbound links on Page Ti
- D is a damping factor which can be set between 0 and 1 (usually set to 0.85)
- $\text{Prev}_{\text{rank}}$(A) is the older PageRank of Page A. (if exists)

A page's PageRank = 0.15 + 0.85 * (a "share" of the PageRank of every page that links to it)

"share" = the linking page's PageRank divided by the number of outbound links on the page.

C. *How Page Rank Works?*

Page Rank calculation is based on the graph of web where each webpage is like a node and each hyperlink is like an edge. Consider an example, Pages B, C and D all links to A, and they already have their PageRank calculated. B also links to three other pages and C links to four other pages, D only links to A.

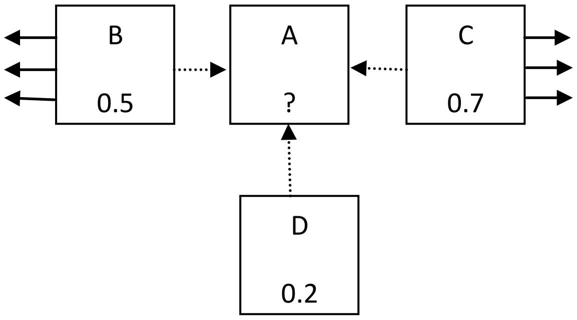


Figure 2: Linking of Four Pages A, B, C and D

To get A's PageRank, take the PageRank(PR) of each of the pages that links to A divided by the total number of links on that page, then multiplying this by a damping factor[3] of 0.85 and add a minimum value of 0.15. The calculation of PR(A) is :

PR(A) = 0.15 + 0.85 ((PR(B) / links(B) + PR(C) / links(C) + PR(D) / links(D))

$$
\begin{aligned}
&= 0.15 + 0.85 * (0.5/4 + 0.7/5 + 0.2/1) \\
&= 0.15 + 0.85 * (0.125 + 0.14 + 0.2) \\
&= 0.15 + 0.85 * 0.465 \\
&= 0.54525
\end{aligned}
$$

2. Proposed Work

A. *Enhanced PageRank Algorithm – An Introduction*

In the classic PageRank Algorithm, the main factor to measure page's PR value is the number of links pointing to that page. Hence, a page that exists on the network longer, it is more likely to hold more links from web pages. According to the algorithm, the PR value of old web pages are higher than new web page, the old web page is seemingly more important than the new page. But this is not true, because people expect to see new information. Therefore, it is necessary to introduce the time factor to revise the PageRank algorithm.

The Enhanced PageRank Algorithm utilizes the time factor to increase the accuracy of the web page ranking. Due to the methodology used in this algorithm, it can be assumed to be a combination of content and link structure. The results of this algorithm are very satisfactory and in agreement with the applied theory for developing the algorithm.

B. *Algorithm*

Usually, the PageRank of the Website increases whenever a user clicks on the Website. Whether that website is beneficial to that particular user or not, a single click will change the PageRank of that Website. It has been seen many times that users usually clicks their Websites in order to increase their PageRank. In order to remove this limitation, we now consider the concept of time. Now the PageRank will not change at the single click but the total time which the user spends on that particular website will also be considered. Now, suppose if the user will spends time > 14 sec (assumed) on the particular website then only the PageRank will change otherwise the previous one is considered.

Now we refer PageRank as "PR"

$$
\begin{aligned}
PR(A) = (1-d) + \ d * (&rank \ / \ (backlinks) \\
&+ 0.15 + 0.85 \ (\ (PR(T1) \\
&/ \ (C(T1)) + \cdots .... \ (PR(Tn) \\
&/ \ (C(Tn)) * counter \\
&+ Prevrank \ (A)) \ \ ))
\end{aligned}
$$

Here,

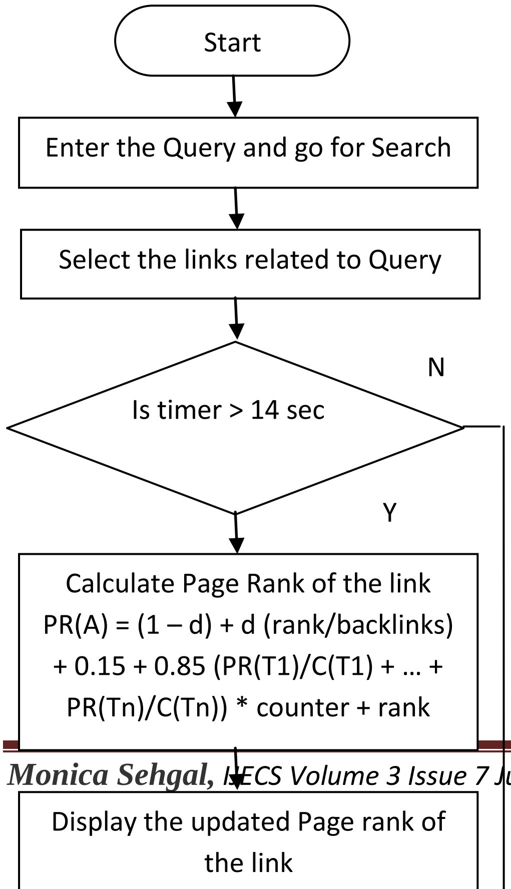
- PR(A) is the PageRank of Page A
- PR(Ti) is the PageRank of pages Ti which link to page A
- C(Ti) is the number of outbound links on page Ti
- D is a damping factor which can be set between 0 and 1.
- Counter denotes the time that a user spends on the website.

- Prevrank(A) is the older PageRank of page A.
- Backlinks are links that are directed towards your Website.

Another factor Backlinks are also known as Inbound links (IBL’s). The number of backlinks is an indication of the popularity or importance of that website. Backlinks are important for SEO because some search engines, especially Google, will give more credit to websites that have good number of quality backlinks, and consider those websites more relevant than others in their results pages for a search query.

A search Engine considers the content of the sites to determine the QUALITY of a link. When inbound links to your site come from other sites, and those sites have content related to your site, these inbound links are considered more relevant to your site. If inbound links are found on sites with unrelated content, they are considered less relevant. The higher the relevance of inbound links, the greater their quality.

C. *Flowchart for Proposed Page Rank Algorithm*

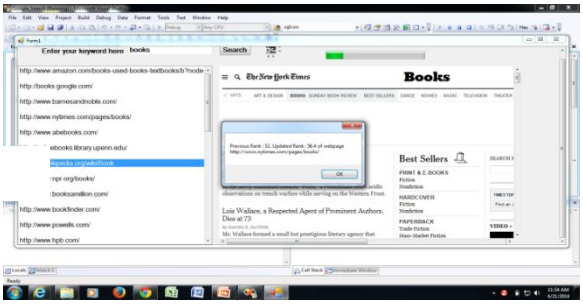

Start
Enter the Query and go for Search
Select the links related to Query
Is timer > 14 sec
N
Y
Calculate Page Rank of the link
PR(A) = (1 – d) + d (rank/backlinks)
+ 0.15 + 0.85 (PR(T1)/C(T1) + … +
PR(Tn)/C(Tn)) * counter + rank
Display the updated Page rank of
the link

3. Experimental / Simulation Results

The execution is performed on 2.40ghz Intel Core(tm) i3 CPU with 2.00 GB RAM, running Windows 7 Professional. Dotnet programming dialect is utilized; since it is an Object Oriented Language and has Security bundles. The Microsoft Visual Studio 2008 advancement framework is a suite of improvement devices intended to support programming designers whether they are tenderfoots or prepared experts – face complex difficulties and make inventive results. Visual Studio's part is to enhance the methodology of improvement and make the work of accomplishing leaps forward simpler and all the more fulfilling. In the Implementation, C# programming dialect is utilized. C# is a straightforward, cutting edge, universally useful; article situated programming dialect created by Microsoft inside its .NET activity headed by Anders Hejlsberg. The latest edition is C# 5.0 discharged on August 15, 2014.

A. *Implementation Details*

The proposed Algorithm utilizes two elements. One is time and other is Backlinks. These two components are utilized with the essential page rank recipe. We have actualized the proposed algorithm as a window based undertaking.


Enter your keyword here books

http://www.amazon.com/books-used-books-textbooks/b?node

http://books.google.com/

http://www.barnesandnoble.com/

http://www.nytimes.com/pages/books/

http://www.abebooks.com/

ebooks.library.upenn.edu/

ikipedia.org/wiki/Book

npr.org/books/

booksamillion.com/

http://www.bookfinder.com/

http://www.powells.com/

http://www.hpb.com/

Previous Rank : 52, Updated Rank : 56.4 of webpage http://www.nytimes.com/pages/books/

Figure 3 Snapshot of Proposed Work

B. Result Analysis and Discussion

The page rank recorded for the simulation of conventional PAGE RANK and the ENHANCED PAGE RANK Algorithms is tabulated in Table 1.

Let's take an example of one URL like http:// www.nytimes.com/pages/books/...

PageRank using Conventional PageRank Algo	PageRank using Enhanced PageRank Algo
52	56.4

Table 1 Shows the Calculated PageRank using CONVENTIONAL PAGE RANK and ENHANCED PAGE RANK ALgo

The Result demonstrates that the Enhanced Page Rank Algorithm is preferable option over the Conventional Page Rank Algorithm. And it also increases the performance of the proposed page rank algorithm.

4. Conclusion

In this paper, an upgraded page rank calculation utilizing time and backlinks as an alternate element has been proposed. Relative investigation of the computational aspects of the proposed plan with the past work means that the proposed page rank is a superior option to the formerly presented calculation as seen from the possibility of time intricacy and the computational investment funds. Later on, the analysts can further investigate more on the page rank calculation to build its execution.

REFERENCES

[1] Wenpu Xing and Ali Ghorbani," Weighted Page rank Algorithm", In Proceeding of the 2nd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

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

[3] Mridula Batra and Sachin Sharma, "Comparative Study of Page rank Algorithm With Different Ranking Algorithms Adopted By Search Engine For Website Ranking", IJCTA, Vol(4), Jan-Feb 2013.

[4] S.Brin and L.Page, "The Anatomy of a Large Scale Hypertextual Web Seracg Engine", Computer Networks and ISDN Systems, Vol 30, Issue 1-7 , 1998.

[5] L.Page, S.Brin, R. Motwani, and T. Winograd, " The Pagerank Citation Ranking: Bringing order to the web". Technical Report, Stanford Digital Libraries SIDL-Wp-1999-0120,1999.

[6] C. Ridings and M. Shishigin, "Pagerank Uncovered", Technical Report, 2002.

[7] Tamanna Bhatia, " Link Analysis Algorithm for Web Mining", IJCST Vol 2, Issue 2, June 2011.

[8] Rekha Jain, Dr.G.N Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications (0975 – 8887) , vol 13, Jan 2011.

[9] Neelam Tyagi, Simple Sharma," Comparative study of various Page Ranking Algorithms in Web Structure Mining", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol 1, Issue 1, June 2012.

[10] R. Kosala and H. Blockeel, " Web Mining Research : A survey", In ACM SIGKDD Explorations, 2(1), PP. 1 – 15, 2000.

[11] S. Chakrabarti et al., "Mining the Web's Link Structure", Computer, 32(8): 60 – 67, 1999.

[12] http://WWW.webrankinfo.com/english/seo-news/topic-16388.htm. January 2006, Increased Google index size.

[13] Naresh Barsagade, "Web Usage Mining And Pattern Discovery: A Survey Paper", CSE 8331, Dec. 8, 2003.

[14] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997.

[15] A. Broder, R. kumar, F Maghoul, P. Raghavan, S.Rajagopalan, R.Stata, A.Tomkins, J. Wiener, "Graph Structure in the Web", Computer Networks: The International Journal of Computer and telecommunications Networking, Vol 33, Issue 1-6 2000.

[16] J. Kleinberg, R. kumar, P. Raghavan, P. Rajagopalan and A. Tompkins, "Web as a graph: Measurements, Models and methods", Proceedings of the international Conference Combinatorics and Computing, 18, 1999.

[17] A.M. Zareh Bidoki and N. Yazdani, "Distance Rank: An intelligent ranking algorithm for web pages" information Processing and Management, Vol 44, No. 2, PP. 877 -892, 2008.

[18] Lihui Chen and wai Lian Chue, " Using Web Structure and summarization techniques for Web Content Mining", Information Processing and Management, Vol 41, PP. 1225 – 1242, 2005.

[19] Kavita D. Satokar and Prof. S. Z. Gawali, " Web Search Result Personalization using Web Mining", International Journal of Computer Applications, Vol 2, No. 5, PP. 29 – 32, June 2010.

[20] P.Boldi, M.Santini, S.Vigna, "PageRank as a Function of the Damping Factor", Proceedings of the 14th World Wide Web Conerence, 2005.