The Anatomy of a Large-Scale Hyper Textual Web Page Ranking

Abha Joshi[1], Avani Jadeja[2]

[1]Hasmukh Goswami College of Engineering, Ahmedabad, India

[2]Hasmukh Goswami College of Engineering, Ahmedabad, India
Student of Master of Computer Science and Engineering, HGCE, Vahelal

Abstract: ***The World Wide Web is the collection of large amount of data or information sources in the form of the web pages as websites and due to dynamic nature of the web plenty of webpages are deleted and added newly. When finding the web for exacting topics, users typically obtain unrelated and unnecessary information due to a waste in user time and accessing time of the search engine. So it is required to pact this problem, user's benefits and needs from their performance have become more and more significant. Page Ranking is an important component for information retrieval system. It is also used to measure the importance and behavior of webpages. Numbers of page ranking algorithms Page Rank (PR), Weighted Page Rank (WPR) are generally used in web structure mining. This paper proposed a new page rank algorithm which use inlink, outlink, number of the times user visits the link of webpages and mean value of the page rank. The relevancy of webpages returned is more and reducing number of iterations to reach convergence point.***

Keyword: Page Rank, Web Mining, Weighted Page Rank, Visit count, Page ranking algorithm, inlinks, outlinks, weighted page rank, Normalization

1. Introduction

Now a days, the WWW is the most popular and interaction medium to publicize information or data. The Web is vast, different and lively. In other hand Web contains infinite amount of information source and offer an access to it at anywhere, any place at any time. Nearly the entire people use internet for obtains information. But many time, they gets lots of irrelevant and extraneous file. For gets information from the Web, the Web mining techniques are used.

As at present WWW is the ubiquitous information collection for understanding indication. The following challenges [1] in Web Mining are:

1) Web is enormous.
2) Web pages are partially structured.
3) Web information stands to be miscellany in meaning.
4) Degree of quality of the in sequence extracted.

Process of Web Mining

The whole process of extracting information from Web data [5] is shown in Figure 1.

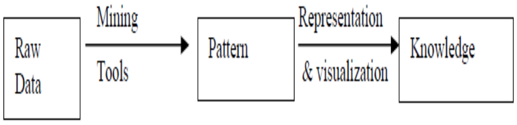

Figure 1: Web Mining Process

The various steps are explained as follows [5]:

(1) Resource finding: It is the task of retrieving intended web documents.
(2) Information selection and pre-processing: Automatically selecting and pre- processing specific from information retrieved Web resources.
(3) Generalization: Automatically discovers general patterns at individual Web site as well as multiple sites.
(4) Analysis: Validation and interpretation of the mined patterns.

Web Mining [2] is the use of data mining techniques to automatically discover and extract Information from web documents and services. The World Wide Web, www or web is becoming a complex universe. Naturally, deriving Something valuable out of it is targeted use of web mining.

Three sub categories [2]:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Web Content Mining:

Web Content Mining refers to the discovery of useful information from the web content, here Content refers to Text, Audio Video etc. that numerous websites are holding. Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

Example of Web Content Mining is:

Typical Google or Yahoo or Microsoft Bing search that we do, and the resultant links listing page we get is an example of content mining. The process of extracting useful information from the web content happens here.

Web Usage Mining:

Web usage mining process involves the log time of pages. The world's largest portal like yahoo, msn etc., needs a lot of insights from the behavior of their users‟ web visits.

Without this usage reports, it will be difficult to structure their monetization efforts. Usage mining has direct impact on businesses.

Web Structure Mining:

The aim of the Web Structure Mining [3] is to generate the structural abstract about the Web site and Web page. It tries to determine the link structure of the hyperlinks at the inter document level. Basic foundation on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and spawn the information like similarity and relationship between different Web sites. This type of mining can be carried out at the document level or at the hyperlink level. It is important to appreciate the Web data structure for Information Retrieval.

2. Related Work

WWW is huge amount of depository of interlinked hypertext credentials right to use through the Internet. Web may have like text information, images, video data, and other multimedia data. Search Engine gives lots of results and applies Web mining techniques to order the results [4]. The sorted arrange of search outcome is obtained by using a number of particular algorithms or method called Page ranking algorithms. There are two types of Page Ranking algorithms; PageRank and Weighted Page Rank. They are the generally used algorithm or method in Web Structure Mining. The algorithm procedures the value of the pages by analyzing the amount of inlinked and outlinked pages.

Page rank of a web page is a weighted number to represent the relative importance of the page based on the number of inbound and outbound links. Inbound links are links from outside pointing to a page [5]. Outbound links are links from a page to any other pages. Page rank algorithm considers a web site is more important if many other web pages are pointing to it.

The original PageRank algorithm is [5]:

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

Where, 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 and d is a damping factor which can be set between 0 and 1.

HITS (Hyper-link Induced Topic Search)

Hyperlink Induced Topic Search (HITS) algorithm ranks the web page by processing in links and out links of the webpages. In links and Out links identifies two different forms of Web pages called hubs and authorities. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many good hub pages on the same subject.

The following are the constraints of HITS algorithm:

- Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.
- Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights. Automatically generated links: HITS gives equal importance for automatically generated links which may not produce relevant topics for the user query. To resolve this problem, a probabilistic analogue of the HITS (PHITS) algorithm is proposed.
- A probabilistic explanation of relationship of term document is provided by PHITS. It is able to identify authoritative document as claimed by the author. PHITS gives better results as compared to original HITS algorithm. Other difference between PHITS and standard HITS is that PHITS can estimate the probabilities of authorities compared to standard HITS algorithm, which can provide only the scalar magnitude of authority.
- Efficiency: HITS algorithm is not efficient in real time.

Weighted PageRank Algorithm

Wenpu Xing and Ali Ghorbani proposed a Weighted PageRank (WPR) algorithm which is an extension of the PageRank algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as W^{in}(m, n) and W^{out}(m, n) respectively.

W^{in}(m, n) is the eight of link(m, n) calculated based on the number of incoming links of page n and the number of incoming links of all reference pages of page m.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p=R(m)} I_p}$$

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p=R(m)} O_p}$$

where In and Ip are the number of incoming links of page n and page p respectively. R(m) denotes the reference page list of page m. W^{out}(m, n) is the weight of link(m,n) calculated based on the number of outgoing links of page n and the number of outgoing links of all reference pages of m.

Where On and Op are the number of outgoing links of page n and p respectively. The formula, as proposed by Wenpu et al,
for the WPR which is a modification of the Page Rank formula.

$$\mathrm{WPR(n)} = (1-\mathrm{d}) + \mathrm{d}\sum_{m\in B(n)} \mathrm{WPR(m)}\, W^{in}_{(m,n)}\ W^{out}_{(m,n)}$$

3. Proposed Algorithm

In this paper a new page ranking is presented which include weights of in links and out links based on the popularity of links in defined ratio. Popularity means number of in links and out lixnks to that link of page. The algorithm also considers the number of time the user visits the in links of any webpages. It is also include normalization in page rank algorithm.

The $W^{in}_{(v,u)}$and $W^{out}_{(v,u)}$ are used to record the popularity (weight) of the in links and out links based on the in links and out links of that link. The mathematically equation of weight are given as follows.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p=R(v)} I_p} \text{ (a)}$$

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p=R(v)} O_p} \text{ (b)}$$

The value of this equation is used for calculating PR.

The equation for the proposed algorithm is as follows.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{(V_u * x * W^{in}_{(v,u)} + y * W^{out}_{(v,u)})PR(v)}{TL(v)} \quad (1)$$

Where PR(u) and PR(v) are ranking of the web pages u and v respectively , d is the dampening factor , V_u is the number of visits of link which points from **v** to u , TL(v) is the total number of visits of all links present on v , B(u) are the pages which points to webpage u , $W^{in}_{(v,u)}$ is the weight of inlinks of connecting page v and u , $W^{out}_{(v,u)}$ is the weight of out links of connecting page v and u.

Algorithm: how actually page rank works as follows.

Step 1: Take the link structure of the retrieved webpages from crawler.
Step 2: Obtain the webgraph from the link structure of the retrieved webpages.
Step 3: Assign 1 as initial ranking to all the webpages.
Step 4: Calculate the weights of inlinks and outlinks using equation (a) and (b).
Step 5: Apply the proposed algorithm as in the equation (1).
Step 6: Calculate mean value of all page rank by following formula:
Summation of page rank of all webpages / number of webpages
Step 7: Then Normalize the page rank of each page.
Norm PR(u) = PR(u) / mean value
Where norm PR(u) is Normalized page rank of page u and PR(u) is page rank of page u.
Step 8: Assign PR(u) = Norm(u).
Step 9: Iteratively repeat process until ranks of all webpages are stable i.e. same in two consecutive iteration.
This algorithm reduce the problem of theme drift which is present on every link structure based algorithms .This results retrieved are efficient and relevant as per user's query.

4. Experimental Analysis

For experimental purpose, we have taken four pages in our database which are crawled by crawler and the figure shows the web graph and links among the four web pages:

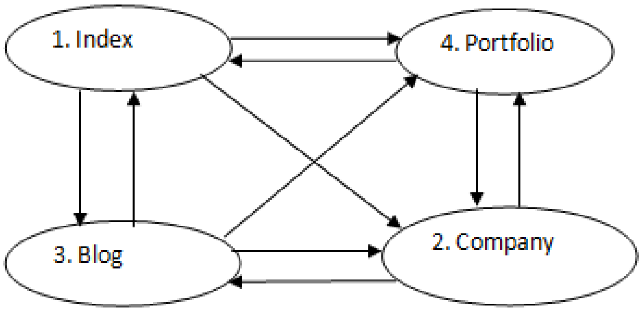

Figure 2: Sample Web Graph

We have applied original pagerank algorithm and our proposed algorithm on this web graph and search one sentence. All four pages are retrieved in search results but according to algorithms the rank value of each page is different and hence the page which is top of the search results by original pagerank algorithm is different by proposed algorithm which is shown in below table:

Table 1: Comparison between the rank values by original pagerank and our proposed algorithm

Id	*Webpages*	*Pagerank Algorithm*	*Our Proposed algorithm*
1	Index	1.1239	0.1680
2	Company	0.8758	0.3586
3	Blog	1.1239	0.1619
4	Portfolio	0.8758	0.2155

5. Conclusion

In this paper we have analyzed various pageranking algorithms for getting efficient and relevant search results as per user's query. We have implemented basic pagerank algorithm. Then we have understood that the algorithms have the main challenge of theme drift. In our proposed algorithm we use web structure mining and Normalization Technique for calculating pagerank values of webpages. After comparing of algorithms we conclude that the our proposed pagerank algorithm provides the better results than the standard page ranking algorithm in terms of the better relevancy and ranking based on the non visited webpages on the basis of the outlink and provides more efficient and relevant search results as per user's query than original pagerank algorithm because when we will get maximum comparison strings in one webpage as per query by user and we will get best relevant results as per user's query.

References

[1] M. G. da Gomes Jr. and Z.Gong, "Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference on Information Acquisition, 2005.
[2] N. V. Pardakhe , Prof. R. R. Keole "Analysis of Various Web Page Ranking Algorithms in Web Structure Mining" ISSN (Print) : 2319 – 5940 ISSN (Online) : 2278 -1021 International Journal of Advanced Research in Computer and Communication Engineering Vol.2 , Issue 12,December 2013
[3] Neelam Tyagi, Simple Sharma "Comparative study of various Page Ranking Algorithms in Web Structure

Mining (WSM)" ISSN: 2278-3075, Volume-1, Issue-1, June 2012, IJITEE.

[4] POOJA SHARMA, DEEPAK TYAGI and DEEPAK TYAGI "Weighted Page Content Rank for Ordering Web Search Result", International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7301-7310

[5] Cooley, R., Mobasher, B., and Srivastava, J. "Web mining: Information and pattern discovery on the World Wide Web". In proceedings of the 9th IEEE International conference on Tools with Artificial Intelligence (ICTAI' 97), Newposrt Beach, CA, 1997.

[6] Ashish Jain, Rajeev Sharma, Gireesh Dixit and Varsha Tomar"Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages " 2013 International Conference on Communication Systems and Network Technologies

[7] Ranveer Singh and Dilip Kumar Sharma," Enhanced-RATIORANK: Enhancing Impact of Inlinks and Outlinks" IEEE Conference on Information and Communication Technologies 2013.

[8] Zhou Cailan and Chen Kai,Li Shasha," Improved PageRank Algorithm Based on Feedback of User Clicks" IEEE 2011.

[9] Wei Huang and Bin Li, "An Improved Method for the Computation of PageRank" International Conference on Mechatronic Science, Electric Engineering and Computer August 19-22, 2011, Jilin, China.

[10] Mohamed-K HUSSEIN Mohamed-H MOUSA "An Effective Web Mining Algorithm using Link Analysis"(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (3) , 2010, 190-197

[11] Shruti Aggarwal, Gurpreet Kaur "Improving the Efficiency of Weighted Page Content Rank Algorithm using Clustering Method" International Journal of Computer Science & Communication Networks, Vol 3(4),231-239