International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

# Evolution of Web Usage Mining in Page Rank Algorithms

**Priyanka Bauddha[1], Thirunavukkarasu[2]**

[1]M. Tech, Galgotias University, School of Computing Science & Engineering, Greater Noida (U.P.) India.

[2]Assistant Professor, Galgotias University, School of Computing Science & Engineering, Greater Noida (U.P.), India

**Abstract:** ***Web Usage Mining is the most active research area for ranking of web pages. Due to abundance of information on web Web Usage Mining is used to extract the behavior of users and to improve the ranking of web pages. Various page rank algorithms have been developed which use Web Usage Mining to calculate value of page ranks. In this paper, surveys of Page Rank algorithms with Web Structure Mining and Web Usage Mining have been performed.***

**Keywords:** inlinks, outlinks, visit of links, web usage mining, log analysis, web mining

## 1. Introduction

Due to enormous amount of information on the web in the form of text, image, video, audio, etc. , it is very difficult to find relevant information for a user. Various ranking algorithms based on web mining have been developed to rank the web pages so that relevant pages are displayed at the top.

Web Usage mining is the application of web mining techniques on log data collected from web server logs. Various Page Ranking algorithms have been developed based on web usage mining. This paper provides a survey of web usage mining based ranking algorithm.

This paper provides a survey of web usage mining comprises of various page rank algorithms with number of visit of links using server logs. Section 2 describes the web usage mining technique in detail. Section 3 describes the various algorithms like PageRank algorithm, Weighted PageRank algorithm, PageRank algorithm using visit of links (VOL), Weighted PageRank algorithm using VOL, Enhanced Weighted PageRank algorithm using VOL. Various algorithms have been compared in Section 4. Section 5 provides conclusion and future work.

## 2. Web Usage Mining Technique

As there are large web data repositories so it's a tough task to get a useful information .So various techniques are applied to get useful knowledge. There are four stages of web usage mining [2]:

### 2.1. Data Collection

Data are collected from different sources like server side, client side and proxy servers. Server Side data collection basically collects the client requests and stored in the weblogs of server. Client Side data collection covers both the caching and session identification problems. They stored the browsing behaviour. Proxy Level collects the data from intermediate server between browsers and web servers.

### 2.2. Data Preprocessing

The heterogeneous and unstructured information available on the web is preprocessed by some tasks like merging and cleaning, user and session identification, etc. Data Cleaning is a process of removing irrelevant items to improve the quality of data. User Identification is done to get information about the user by using his IP address. Session Identification is a set of pages which are visited by the same user within the duration of particular visit to website. Path Completion is used when pages are missed after constructing transactions due to proxy servers so by the help of log record back button is used until the page has been reached.

### 2.3. Pattern Discovery

Association Rule Discovery techniques are applied to transaction databases by using Apriori algorithm to discover biggest frequent item set. Clustering technique is used to group the users who have similar browsing behaviour.

### 2.4. Pattern Analysis

The Pattern analysis is the last stage of web usage mining technique. Pattern Analysis is used to filter out uninteresting patterns from the set found in pattern discovery phase.

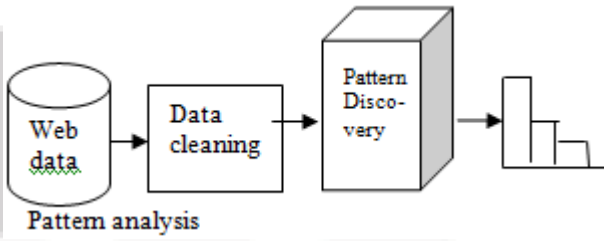


**Figure 1:** Web Usage Mining [2]

## 3. Page Ranking Algorithms

### 3.1. PageRank Algorithm

**Brin and Page [9]** introduced Page Rank Algorithm first time by using page link structure for the calculation of page ranks at Stanford University. Page Rank Algorithm is the heart of Google search engine. The principle of Page Rank Algorithm is that if a document has important links towards

it then the outlinks of that page are also important. The page rank of a document is calculated as given in equation 1:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \quad (1)$$

$N_v$ = total number of outlinks of web page p
c= factor of normalization.
Original algorithm was modified as all users don't follow direct links on WWW. The modified formula is given in equation 2:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \quad (2)$$

Here,
d= dampering factor which signifies that the probability of using direct links and its value varies between 0 and 1.

### 3.2. Weighted PageRank Algorithm

**Wenpu Xing and Ghorbani [8]** proposed Weighted Page Rank Algorithm by extending Page Rank Algorithm. The principle of this algorithm is that page rank of a web page is distributed among its incoming and outgoing pages in proportional to its popularity. The popularity of a web page from its incoming links is calculated as given in equation 3:

$$W^{in}(v,u) = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad (3)$$

where,
$I_u$= Number of inlinks of page u.
$I_p$= Number of outlinks of page p.
R(v)= set of web pages pointed by v.

The popularity of a web page from its outgoing links is calculated as given in equation 4:

$$W^{out}(v,u) = \frac{O_u}{\sum_{p \in R(v)} O_p} \quad (4)$$

where,
$O_u$= number of outlinks of page u.
$O_p$= number of outlinks of page p.

The weighted page rank is calculated as given in equation 5:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}(v,u) W^{out}(v,u) \quad (5)$$

### 3.3. Page Ranking Algorithm using Visit Of Link

In basic Page Rank Algorithm the p, is divided between inlinks and outlinks. **Gyanendra Kumar[3]** introduced a new algorithm for the calculation rank of web pages. The original PageRank algorithm gives the page rank value among its outlinks equally but this algorithm assigns proportion to the number of visit of links. The formula is given in equation 6:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} L_u \frac{PR(v)}{TL(v)} \quad (6)$$

Here,
$L_u$= the number of visit of links which are pointing page u from v.
TL(v) denotes total number of visit of all links.
PR(u)= page rank of page u
PR(v)= page rank of page v
d denotes the dampering factor.
B(u)= set of web pages pointing to u.

### 3.4. Weighted Page Rank Algorithm Using VOL

**Neelam Tyagi and Simple Sharma[4]** proposed a new algorithm called Weighted PageRank algorithm using Visit Of Links(VOL) to combine Visit Of Links with Weighted PageRank. In this algorithm the value assign more rank to outgoing links which are mostly visited by the users and have higher popularity inlinks. Calculate $W^{in}(v, u)$ for each node by applying following formula in equation 7:

$$W^{in}(v,u) = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad (7)$$

where,
$I_u$= Number of inlinks of page u.
$I_p$= Number of outlinks of page p.
R(v)= set of web pages pointed by v.

Weighted Page Rank with VOL formula

$$WPR_{VOL}(u) = (1-d) + d \sum_{v \in B(u)} \frac{L_u WPR_{VOL}(v) W^{in}(v,u)}{TL(v)} \quad (8)$$

where,
$WPR_{VOL}(u)$= Page rank of page u
$WPR_{VOL}(v)$= Page rank of page v
d= dampering factor
Lu= the number of visit of links which are pointing page u from v.
TL(v)= total number of visit of all links.
B(u)= set of web pages pointing to u.
$W^{in}_{VOL}(v,u)$= number of visits of inlinks of page u and the number of inlinks of all referencesof page v.

### 3.5. Enhanced Page Rank Algorithm Using Visit of Links (VOL)

**Sonal Tuteja[5]** proposed enhancement in weighted page rank algorithm using VOL. The algorithm considers that the popularity from the number of visits of inlinks and the popularity of the number of outlinks are used to calculate the value of weighted page rank which is not used in weighted page rank with VOL algorithm. The weight of inlink using VOL is calculated by equation 9:

$$W^{in}_{VOL}(v,u) = \frac{I_{u(VOL)}}{\sum_{p \in R(v)} I_{p(VOL)}} \quad (9)$$

Where,
$W^{in}_{VOL}(v,u)$= number of visits of inlinks of page u and the number of inlinks of all referencesof page v.
$I_{u(VOL)}$= incoming visits of links of page u.
$I_{P(VOL)}$= incoming visit of links of page p.
The weight of outlinks using VOL is calculated by equation 10:

$$W^{out}_{VOL}(v,u) = \frac{O_{u(VOL)}}{\sum_{p \in R(v)} O_{p(VOL)}} \quad (10)$$

Where,
$W^{out}_{VOL}(v,u)$= weight of link(v,u) or number of visits of outlinks of page u and number of visits of all references of page v.

$O_u(VOL)$= outgoing visits of links of page u.
$O_p(VOL)$= outgoing visits of links of page p.
Then, the final formula given in equation 11 as:

$$EWPR_{VOL}(u) = (1-d) + d \sum_{v \in B(u)} WPR_{VOL} W^{in}_{VOL}(v,u) W^{out}_{VOL}(v,u) \quad (11)$$

Where,
d = Dampering factor,
B(u)= set of pages point to u.
$EWPR_{VOL}(u)$ =Rank score of page u.
$EWPR_{VOL}(v)$= Rank score of page v.
$W^{in}_{VOL}(v,u)$= popularity of number of visits of inlinks.
$W^{out}_{VOL}(v,u)$= popularity of number of visits of outlinks.

## 4. Comparison Between Various PageRank Algorithms

This section compares the various page ranking algorithms on the basis of web mining technique used by the algorithm, input parameters, importance and their relevancy for users. Table 1 shows the comparison of ranking algorithms PageRank algorithm and Weighted PageRank algorithm with PageRank algorithm using visit of links(VOL), Weighted PageRank algorithm using VOL, and Enhanced Weighted PageRank algorithm using VOL.

**Table1:** Comparison between various Page Ranking Algorithms

| Algorithm | Page Rank Algorithm | Weighted Page Rank Algorithm | PageRank with VOL | Weighted PageRank with VOL | Enhanced Weighted Page Rank with VOL |
|---|---|---|---|---|---|
| Web mining technique used | Web structure mining | Web Structure Mining | Web structure mining, web usage mining | Web structure mining, web usage mining | Web structure mining, web usage mining |
| Input Par-ameters | Backlinks | Backlinks, Forward links | Backlinks and VOL | Backlinks and VOL | Backlinks, forward links and VOL |
| Importance | More | More | More | More | More |
| Relevancy | Less | Less | More | More | More |

## 5. Conclusion

Search engines are using Page Ranking algorithms to improve ranking of pages so that the users wouldn't have to spend lot of time to search relevant pages. Various discussed algorithms have their own benefits on searching. Various other factors like time spending on a link, web server logs, most recently used links, etc. can also be used to calculate the ranking of web pages.

## References

[1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", Department of Computer Science and Engineering, University of Minnesota, Minneapolis, ACM Jan 2000.

[2] V.Chitraa, Dr.Antony Selvdoss Davamani,"A survey on Preprocessing Methods for Web Usage Data",Department of Computer Science,CMS college of Science and Commerece, Tamil Nadu,India,IJCSIS vol. 7,No. 3, 2010.

[3] Gyanendra Kumar, Neelam Duhan, A. K. Sharma, "Page Ranking Based on Number of Visits of Links of Web Page", Department of Computer Engineering, YMCA University of Science & Technology, Faridabad, India, ICCCT 2011.

[4] Neelam tyagi, Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[5] Sonal Tuteja,"Enhancement in Weighted Page Rank Algorithm Using VOL", Software Engineering, Delhi Technological University, India,IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume X, Issue X (Sep. - Oct. 2013), PP 01-00.

[6] Rodney Fuller, Johannes J.Graaff,"Measuring User Motivation from Server Log Files",Microsoft Usability:Dsigning for web.

[7] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Science Department, Stanford University, Stanford, CA 94305.

[8] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Faculty of Computer Science, University of New Brunswick,Fredericton, NB, E3B 5A3, Canada.

[9] Auth Dell Zhang, Yisheng Dong, "A novel Web usage mining approach for search engines", Computer Networks 39 (2002) 303–310or Profile

## Author Profile

**Priyanka Bauddha** pursuing M. Tech. in Software Engineering from Galgotias University. She completed B.Tech. in Information Technology from HCST, Mathura in 2012 Area of interest is Data Mining.

**Thirunavukkarasu K.,** is Assistant Professor at Galgotias University, Greater Noida, Delhi-NCR. He is pursuing PhD in CSE in the research area of Spatial Database. He has been a student of Madras University, Bharathiar University, and Anna University, Chennai, India. He has more than 14 years of experience in Teaching and 3 years in software industry. He has taught for APIIT, (affiliated to Staffordshire University, UK) at Panipat, Vijaya College, Surana College and KKECS College, Bangalore University, Bangalore and worked as Software Engineer for I2 Technology, UK at Bangalore. He has involved in various academic activities like BoE member and Assistant Custodian for PG-Unit, Bangalore University, Bangalore. He has wide research interests that include Knowledge Engineering, Data Mining, and Databases Technology. He is a Member of IEEE, CSI and Life Member of ISTE. He has 9 certificates from IMS, IBM and trained 150 students on IBM DB2 certificates. He was the organizing Secretary for an International Conference ICACCT 2010. He has conducted various workshops, short term and summer courses. He has published 9 papers in international and 3 papers in national level.