I.J. Modern Education and Computer Science, **2019, 10, 41-46**
Published Online October 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2019.10.05

Recent and Frequent Informative Pages from Web Logs by Weighted Association Rule Mining

Dr. SP. Malarvizhi
Associate Professor, Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India.
Email: spmalarvizhi1973@srivasaviengg.ac.in

Received: 17 July 2019; Accepted: 26 August 2019; Published: 08 October 2019

Abstract—Web Usage Mining provides efficient ways of mining the web logs for knowing the user's behavioral patterns. Existing literature have discussed about mining frequent pages of web logs by different means. Instead of mining all the frequently visited pages, if the criterion for mining frequent pages is based on a weighted setting then the compilation time and storage space would reduce. Hence in the proposed work, mining is performed by assigning weights to web pages based on two criteria. One is the time dwelled by a visitor on a particular page and the other is based on recent access of those pages. The proposed Weighted Window Tree (WWT) method performs Weighted Association Rule mining (WARM) for discovering the recently accessed frequent pages from web logs where the user has dwelled for more time and hence proves that these pages are more informative. WARM's significance is in page weight assignment for targeting essential pages which has an advantage of mining lesser quality rules.

Index Terms—Web logs, Web Mining, Page Weight Estimation, Weighted Minimum Support, WARM, WWT.

I. Introduction

The Literature shows that Data Mining is a field which gains a rapid growth in recent days. Association Rule Mining (ARM) of this field plays a vital role in research [1]. Frequent itemset mining uses ARM algorithms to get the association amongst items based on user defined support and confidence [2]. Existing literature say that from the foremost frequent itemset mining algorithms like Apriori and FP-growth, many algorithms have so far been evolved. ARM gains application in business management and marketing.

In case of WARM, every individual item is assigned a weight based on its importance and hence priority is given for target itemsets for selection rather than occurrence rate [3,4,5,6]. Motive behind WARM is to mine lesser number of quality based rules which are more informative.

Web log mining also known as web usage mining , a category of web mining is the most useful way of mining the textual log of web servers to enhance the website services. These servers carry the user's interactions with the web [7, 8].

This paper provides a method of WARM for mining the more informative pages from web logs by a technique called WWT where weight is assigned based on time dwelled by the visitor on a page and the recent access. Log is divided into n windows and weights are provided for each window. Last window is the recently accessed one and carries more weight. Along with the window weight the time dwelled on a page also adds to the priority of a target page to be mined.

Rest of the paper is arranged as follows. Section 2 appraises related works of Frequent Patterns and WARM for Web log. Section 3 explains about the proposed system of how to preprocess the web logs, weight assignment techniques, WWT structure and WARM. Section 4 bears the experimental evaluation. Section 5 offers conclusion.

II. Related Work

To obtain the frequent pages from web logs and to provide worthy information about the users FP-growth algorithm is used [9].

Web site structure and web server's performance can be improved by mining the frequent pages of the web logs to cater the needs of the web users [10].

A measure called w-support uses link based models to consider the transaction's quality than preassigned weights [6].

ARM does not take the weights of the items into consideration and assumes all items are equally important, whereas WARM reveals the importance of the items to the users by assigning a weight value to each item [11].

An efficient method is used for mining weighted association rules from large datasets in a single scan by means of a data structure called weighted tree [3].

Wei Wang et al [4] proposed an effective method for WARM. In this method a numerical value is assigned for every item and mining is performed on a particular weight domain. F.Tao et al [5] discusses about weighted setting for mining in a transactional dataset and how to discover important binary relationships using WARM.

Frequently visited pages that are recently used shows user's habit and current interest. These pages may be mined by WARM techniques and can be made available in the cache of the server to make the web access speedy [12]. Here the web log is divided into several windows

and weight is assigned for each window. The latest window which carries the recently accessed pages possesses more weight.

To deliver the frequent pages that acquire the interestingness of the users an enhanced algorithm is proposed by Yiling Yang et al. Content-link ratio and the inter-linked degree of the page group are the two arguments made use of in support calculation [13].

Weight of each item is derived by using HITS model and WARM is performed [14]. This model produces quality oriented rules of lesser number to improve the accuracy of classification.

Vinod Kumar et al [15] have contributed a strategy which aims in discovering the frequent pagesets from weblogs. It focuses on fuzzy utility based webpage sets mining from weblog databases. It involves downward closure property.

K.Dharmarajan et al [16] have provided valuable information about users' interest to obtain frequent access patterns using FP growth algorithm from the weblog data.

Proposed system performs enhanced WARM for obtaining the frequent informative pages from textual web logs of servers. WWT method used in the system involves tree data structure. Database has to be divided into several windows and weights have to be assigned to the windows with the latest window having more weightage and then WWT is constructed.

III. Proposed System

Existing literature discusses about mining the frequent pagesets and association rules based on user defined minimum support and confidence. Proposed system aims at achieving quality rules from web logs using WARM. Weights for web pages visited by users are assigned based on factors like frequent access, time dwelled by visitors on web pages and how far the pages are recently used. Weighted Window Tree proposed here uses WARM to discover such recently used frequent pages from web log. Fig.1 shows the work flow process involved in the system.

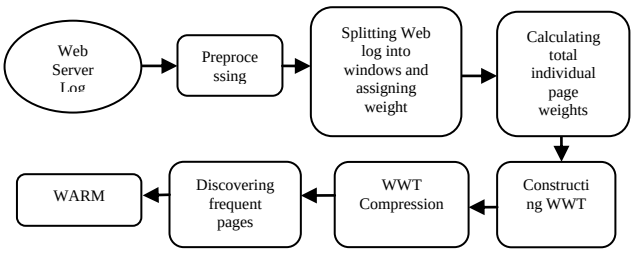

Fig.1. Work Flow

A. Web Log Preprocessing

Web log is preprocessed for removing the duplicates, images, invalid and irrelevant data by cleaning. Data preprocessing is a more difficult task, but it serves reliability and data integrity necessary for frequent pattern discovery. Preprocessing takes about 80% of the total effort used up for mining [17]. Preprocessed web log then consists of attributes like IP address, Session ID, URLs of the pages visited by the user in that session ID, Requested date and time of each page and time dwelled on each page.

The arrangement of data in the web log will be like a relational database model. The details of the pages requested by the users are stored consecutively in the web log. For each page requested all the attributes are updated. The log stored during a particular period from which the frequent informative pages are to be mined forms the back end relational database. Table 1 shows a part of the web log of an educational institution (EDI). Every requested page is reorganized as a separate entry with a distinct session ID. In the table session ID 3256789 under IP Address 71.82.20.69 has recorded three pages.

Table 1. Partial web log of edi data set

Sl.. No.	IP Addr.	Session ID	Page URL	Requested Date & Time	Dwelling Time (Min.)
1.	71.82.20.69	3256789	/ece.html	06/May/2013 03:12:32	4.25
2.	71.82.20.69	3256789	/civil.html	06/May/2013 03:16:47	5.36
3.	71.82.20.69	3256789	/mark.html	06/May/2013 03:22:09	1.00
4.	73.165.66.213	3256790	/admission.html	06/May/2013 03:23:09	3.06
5.	73.165.66.213	3256790	/plac.html	06/May/2013 03:26:13	6.43

B. Windows and Weights

Frequent pages may have regularly occurred in the previous stage of a particular duration of a web log than in the later stage. In order to categorize most recently and least recently accessed frequent web pages WWT arrangement is introduced.

The log for a particular duration is first divided into N number of windows, probably of equal size. While splitting windows, care should be taken to see that the pages of same session ID are not getting divided into two different windows and belong to a same window. To resolve this, division of windows is made based on total number of sessions available in the log divided by number of windows needed. This gives equal number of sessions S per window except for the latest window sometimes, which carries remainder number of sessions got by the division.

Windows are given index numbers from bottom to top as 1 to N. Last window at the bottom of the log consisting of recently accessed web pages has index as 1 and is given a highest weight (hWeight) lying between 0 to 1, $0<hWeight< 1$. Window 2, just before window 1 from bottom will have a weight less than that of window 1. Equation (1) is made use of to provide the weight for windows. It is understandable that window Index i < window index j and from equation (1) it is obvious that

$weight_i < weight_j$, i.e. when window index raises, window weight reduces [12]. Sessions of each window are renumbered separately from 1 to S as shown in Table.3.

$$weight_i = (hWeight)^i \quad (1)$$

Where i is the window index from 1 to N. After assigning weights for individual windows, each individual page URL in the windows gets assigned with weight equal to the weight of the window. Hence a particular URL, lying in the last window would have gained more weight than the same URL lying on other windows, which clearly shows the importance of the pages recently accessed.

Same URL appearing in different windows may have different page dwelling time. The total URL weight of every individual URL is then found based on the occurrence rate of the page by using equation (2), which adds up all the products of individual window weight and dwelling time T of one particular URL lying in several windows. If URL $\in$ i^{th} window and j_{th} session of i_{th} window then,

$$W_{URL} = \sum_{i=1}^{N} \sum_{j=1}^{S} (weight_i * T_{ij}) \quad (2)$$

In equation (2) N is the total number of windows, S is total number of sessions in i^{th} window, $weight_i$ is the weight of i^{th} window found in equation (1) which assigns the same weight to all the URLs in that window irrespective of sessions and T_{ij} is the total dwelling time of a page in j^{th} session of i^{th} window. Number of products of ($weight_i * T_{ij}$) incurred is the count of number of occurrences of the particular page URL. To calculate the average page weight or pageset (set of pages) weight, equation (3) is used.

$$W_{ps} = \frac{1}{n_{ps}} \sum_{k=1}^{m} (W_{URL})_k \quad (3)$$

In equation (3) W_{ps} is the weight of a pageset ps and it is the ratio of sum of total weights of individual pages belonging to that page set to the number of visitors of that pageset n_{ps} (number of session IDs in which the pageset is occurring) and m is the number of pages in the pageset. W_{URL} is found using equation (2) for those records of ps having non zero entries for all the pages of the pageset. For 1-pageset $n_{ps}=n_p$ and $W_{ps}=W_p$ i.e. the individual page weights and hence m=1.

C. Weighted Window Tree (WWT) and WARM

WWT method consists of the following steps.

(i) Constructing WWT.
(ii) Compressing the tree.
(iii) Mining recently and frequently visited pagesets.
(iv) WARM.

Table 2 shows a partial sample data log with dwelling time of the pages for few session IDs. Log is divided into windows and window weights are allotted. A relational database as shown in table 3 is constructed for the data in table 2 which aids in constructing the tree. Pages are numbered as p1, p2, p3,… from the last record of the last window (index 1) of the web log dataset towards the top window of a web log considered for a duration. When an URL gets repeated it is given the same page number.

Table 2. Partial sample data log

Visitors (or) Session ID	Page dwelling time (min.)				j	Window Index (i)	Window Weight
	p1	p2	p3	p4			
1	8	3	0	0	1	2	0.25
2	0	2	4	0	2		
3	4	0	4	0	3		
4	6	2	5	0	1	1	0.5
5	0	7	0	15	2		

Table 3. Sample relational database

Visitors (or) Session ID	Window weight * Page dwelling time				j	Window Index (i)	Window Weight
	p1	p2	p3	p4			
1	2	0.75	0	0	1	2	0.25
2	0	0.5	1	0	2		
3	1	0	1	0	3		
4	3	1	2.5	0	1	1	0.5
5	0	3.5	0	7.5	2		

The entries under the pages of the sample relation are the product of dwelling time of the particular page by the particular visitor and the window weight in which the page is lying. Lowest window is the latest window and is given index 1. Weight of first window is assumed 0.5 (between 0 to 1). Using equation (1) weight of 2^{nd} window is calculated as 0.25.

1) Constructing Weighted Window Tree

By one scan of the data base with divided windows and weight, the Weighted Window Tree as shown in fig.2 can be constructed. Ellipses of the tree are called as page nodes, which consists of Page URL and rectangles are called as SID (Session ID) nodes which consist of SID and weight. There are 2 pointers for every page node. One pointer is directed towards the next page node and the other towards the successive SID nodes consisting of that Page URL [3].

.

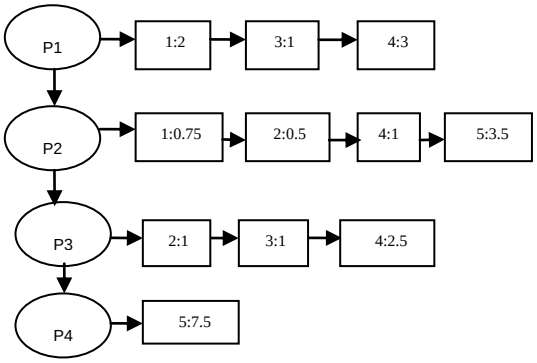

Fig.2. WWT for the sample database

Once the tree is constructed, it has to be compressed to make it prepared for mining using user defined weighted minimum support (w_{ms}) to obtain the recent and frequent informative pages.

2) Tree compression

Frequent pagesets are mined based on user defined weighted minimum support W_{ms}. Let W_{ms}=1.5 for the sample relation. If individual page weight $W_p \geq W_{ms,}$ then the page is said to be frequent. By equation (2) and (3) the average page weights of the individual pages are calculated as shown below.

$$W_{p1} = (2+1+3)/3 = 2$$
$$W_{p2} = (0.75+0.5+1+3.5)/4 = 1.44$$
$$W_{p3} = (1+1+2.5)/3 = 1.5$$
$$W_{p4} = (7.5)/1=7.5.$$

Pages p1, p3 and p4 are alone frequent-1 pages, as they have weights more than W_{ms}. Hence the page node p2 and its attached SID nodes are removed from the tree since it is infrequent and the pointer is made to point directly to p3 from p1. This tree after reducing the infrequent page nodes and its relevant branches of weight nodes is called as compressed tree.

3) Mining recently and frequently visited pagesets

WWT method covers both user's habit and interest. Habit is one which is regularly used (frequently used pages) and interest is one which changes according to time (recently used pages). High w_{ms} leads to mining of less number of more recently used frequent pages, which shows probably users timely interest and low w_{ms} leads to mining of more number of less recently used frequent pages, which shows users regular habits. More recently used pages can be kept in server's cache to speed up web access.

In literature WARM does not satisfy the downward closure property. It is not necessary that all the subsets of a frequent pageset should be frequent, because the logic of frequent pageset handles weighted support. But the proposed WWT method overrides this. For an m-pageset to be frequent, all the individual pages in the non zero records of m pageset have to be individually frequent.

All the frequent-1 pages are put in a set named {F1} and the non empty subsets of it are obtained except the 1 element subsets and power set. For our example {F1}={p1,p3,p4} and such non empty subsets of it are {p1,p3}, {p3,p4} and {p1,p4}. The pageset weight W_{ps} of all the non empty subsets obtained from {F1} is calculated using equation (3) and they are checked for whether frequent or not. For example for ps={p1,p3}, $W_{p1p3}=(1/n_{p1p3})(W_{p1}+W_{p3})$. n_{p1p3} is the number of visitors who have visited both p1 and p3. It is found by the intersection of the session IDs of the visitors of p1 and p3. Hence {1,3,4}∩{2,3,4} gives {3,4} i.e. only the non zero entry records of all the pages of the pageset {p1,p3}. Two visitors (3^{rd} and 4^{th} visitor) have visited p1 and p3, out of a total of five visitors. Therefore W_{p1p3} = (1+1+3+2.5)/2 = 3.75>1.5. Individual weights of p1 and p3 for 3^{rd} and 4^{th} visitor's records are W_{p1}=(1+3)/2=2>1.5 and W_{p3}=(1+2.5)/2=1.75>1.5. Hence pageset {p1,p3} is a frequent pageset according to downward closure property. Similarly all other subsets have to be checked.

4) WARM

Weighted association rules are those strong rules obtained from the frequent pagesets by prescribing a user defined minimum weighted confidence C_{min}. Let 'x' be one of the frequent pagesets and's' a subset of x, then if $W_x/W_s \geq c_{min,}$ then a rule of the form s ⟹ x-s is obtained.

Let us consider for example the frequent pageset x={p1,p3} and s={p1}, the subset of x. If C_{min} = 1.5, then since W_{p1p3}/Wp_1 = 3.75/2 = 1.875 >1.5, p1 ⟹ p3 is a strong association rule.

IV. Experimental Evaluation

For evaluating the performance of the proposed method, experiments are performed on two different datasets. One is an EDI dataset (Educational Institution) and the other is msnbc dataset available in UCI machine learning repository from Internet Information Server (IIS) logs for msnbc.com. Comparison is made between weighted tree [3] and our proposed WWT methods, in terms of speed and space for various W_{ms}. Both the data are extracted for one full day duration. Speed is calculated in terms of CPU execution time by including stubs in the program.

Experiments were performed on an Intel Core I5, 3.2 GHz processor machine with 2GB RAM and 500 GB hard disk with Windows XP platform. WWT algorithm is implemented in Java.

The proposed WWT method shows an enhanced performance with less execution time and less space than the Weighted Tree method and the experimental results are presented below from figure 3 to figure 6 for both datasets. It means it produces less number of frequent pagesets and hence takes less time (speed) and space.

Comparison of the results for execution time (Speed) of both Weighted Tree (WT) and Weighted Window Tree (WWT) methods are shown in fig.3 and fig.4 and that for comparison of number of pagesets (Space) generated by both the methods are given in fig.5 and fig.6. It is seen that, as w_{ms} increases the execution time and space

reduces more for WWT than for WT. WWT is also well scalable when input data set size increases.

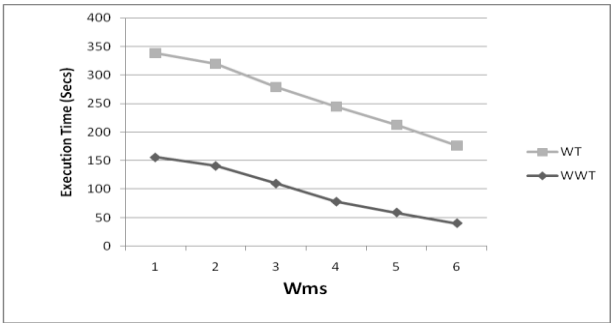

Fig.3. Comparison of Execution time for WT and WWT for msnbc dataset

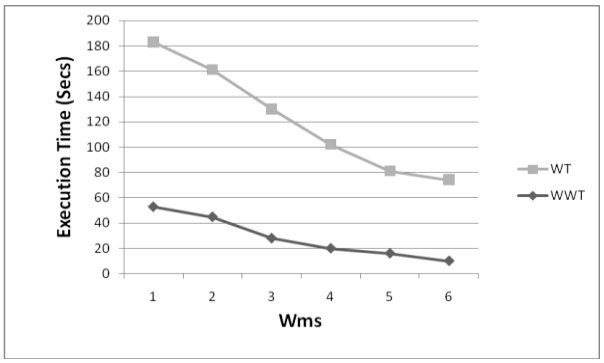

Fig.4. Comparison of Execution time for WT and WWT for EDI dataset

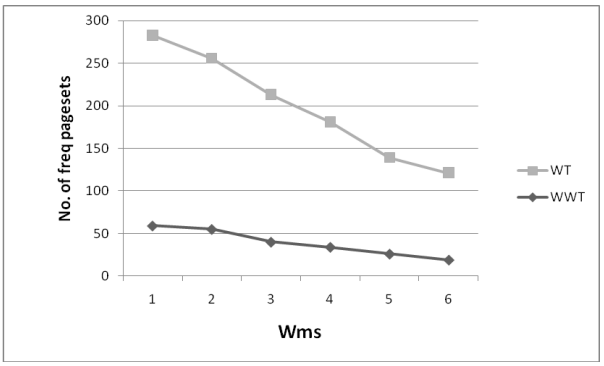

Fig.5. Comparison of Number of pagesets generated by WT and WWT for msnbc dataset

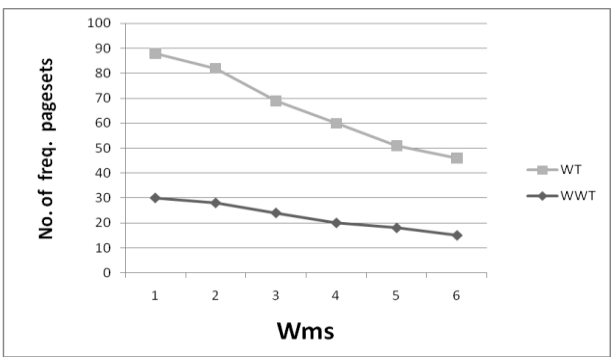

Fig.6. Comparison of Number of pagesets generated by WT and WWT for EDI dataset

Fig.3 implies that WWT has its ececution time reduced nearly by 50% than WT for msnbc dataset. Fig.4 implies that the execution time of WWT for EDI dataset has reduced one fourth than WT. This clearly reflects that the removal of insignificant pagesets by WWT reduces the execution time. Fig.5 and fig.6. reveal that number of frequent pagesets mined using WWT considerably reduces because of window weightage technique.

V. Conclusion

Method for mining recent and frequent informative pages from web logs based on window weights and dwelling time of the pages is discussed. This utilizes Weighted Window Tree arrangement for Weighted Association Rule Mining and it is seen more efficient than Weighted Tree from experimental evaluation by means of speed and space. The system covers less recently used frequent pages from earlier stages and more recently used frequent pages from later stages.

This method finds its main application in mining the web logs of educational institutions, to observe the surfing behaviour of the students. By mining the frequent pages using WWT, most significant pages and websites can be identified and useful informative pages can be accounted for taking future decisions. This technique can be used for mining any organizational weblog to know the employees browsing behaviour.

The limitation of WWT lies in availing at a better weight allotting scheme which can reduce the execution time and number of pages mined even more better than WWT.

References

[1] Qiankun Zhao, Sourav S. Bhowmic, "Association Rule Mining: A Survey" Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116 , 2003.

[2] Han. J and M. Kamber(2004), "Data Mining Concepts and Techniques": San Francisco, CA:. Morgan Kaufmann Publishers.

[3] Preetham kumar and Ananthanarayana V S, "Discovery of Weighted Association Rules Mining", 978-1-4244-5586-7/10/$26.00 C 2010 IEEE, volume 5, pp.718 to 722.

[4] W. Wang, J. Yang and P. Yu, "Efficient mining of weighted association rules (WAR)", Proc. of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 270-274, 2000.

[5] F.Tao, F.Murtagh, M.Farid, "Weighted Association Rule Mining using Weighted Support and Significance framework", SIGKDD 2003.

[6] Ke Sun and Fengshan Bai, "Mining weighted Association Rules without Preassigned Weights", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 4, pp.489-495, April 2008.

[7] Hengshan Wang, Cheng Yang and Hua Zeng, "Design and Implementation of a Web Usage Mining Model Based on Fpgrowth and Prefixspan", Communications of the IIMA, 2006 Volume 6 Issue2, pp.71 to 86.

[8] V.Chitraa and Dr. Antony Selvadoss Davamani, "A Survey on Preprocessing Methods for web Usage Data", (IJCSIS)

International Journal of Computer Science and Information Security,Vol. 7, No. 3, pp.78-83, 2010.
[9] Rahul Mishra and Abha Choubey, " Discovery of Frequent Patterns from web log Data by using FP Growth algorithm for web Usage Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 9, pp.311-318, Sep 2012.
[10] Renáta Iváncsy and István Vajk, "Frequent Pattern Mining in web Log Data", Acta Polytechnica Hungarica Vol. 3, No. 1, 77-90, 2006.
[11] P.Velvadivu and Dr.K.Duraisamy, "An Optimized Weighted Association Rule Mining on Dynamic Content", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 5, pp.16-19, March 2010.
[12] Abhinav Srivastava, Abhijit Bhosale, and Shamik Sural, "Speeding Up web Access Using Weighted Association Rules", S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 660–665, 2005. _c Springer-Verlag Berlin Heidelberg 2005.
[13] Yiling Yang, Xudong Guan, Jinyuan You,"Enhanced Algorithm for Mining the Frequently Visited Page Groups", Shanghai Jiaotong University, China.
[14] S.P.Syed Ibrahim and K.R.Chandran, "compact weighted class association rule mining using information gain", International Journal of Data Mining & knowledge Management Process (IJDKP) Vol.1, No.6, November 2011.
[15] Vinod Kumar and Ramjeevan Singh Thakur, "High Fuzzy Utility Strategy based Webpage Sets Mining from Weblog Database", International Journal of Intelligent Engineering and Systems, Vol.11, No.1, pp.191-200, 2018.
[16] K.Dharmarajan and Dr.M.A.Dorairangaswamy, "Web Usage Mining: Improve the User Navigation Pattern using FP-Growth algorithm", Elysium Journal of Engineering Research and Management, Vol.3, Issue 4, August 2016.
[17] Liu Kewen, "Analysis of preprocessing methods for web Usage Data", 2012 International conference on measurement, Information and Control (MIC), School of Computer and Information Engineering, Harbin University of Commerce, China.

Author's Profile

Dr.SP.Malarvizhi received the BE degree in Electrical and Electronics Engineering from Annamalai University, India in 1994, ME degree in Computer Science and Engineering from Anna University of Technology, Coimbatore, India in 2009 and received the Ph.D. degree in Data Mining from Anna University Chennai in 2016. She is currently working as an Associate Professor in CSE department at Sri Vasavi Engineering College, Andhra Pradesh since 2017. She has participated and published papers in many National and Internal Conferences and also published 7 papers in National and International journals. Her research interests are Data Mining, Big Data and Machine Learning.

How to cite this paper: SP. Malarvizhi, " Recent and Frequent Informative Pages from Web Logs by Weighted Association Rule Mining", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.10, pp. 41-46, 2019.DOI: 10.5815/ijmecs.2019.10.05

