Web Recommendation Framework based on Association Rules Coverage to be Applied for Site Modification

M. Maged M. Deghaidy
Egyptian Armed Forces
Cairo, Egypt

Khaled Mahmoud Badran
Egyptian Armed Forces
Cairo, Egypt

Gouda Ismail Mohamed
Egyptian Armed Forces
Cairo, Egypt

ABSTRACT

This paper introduces a Web Recommendation Framework based on the usage history to be applied for Site Modification as one of the applications of Web Usage Mining (WUM) that is applicable for online business and marketing applications. The framework focuses on the three main interdependent tasks for performing WUM which are Preprocessing, Pattern Discovery and Pattern Analysis. In Preprocessing, we remove all irrelevant users' requests from the web server log file leading to log reduction followed by users' identification then sessions' identification. We take into consideration the ambiguity found in some researches concerning the HTTP common methods. In Pattern Discovery, we extract Association Rules that can be used to relate pages that are most often referenced together in a single server session and may not be directly connected to one another via hyperlinks. In Pattern Analysis, we analyze the set of generated rules independent of the website's topology to extract valid set of rules that achieves highest coverage for the dataset. Our experimental results confirmed that calculating association rules coverage in our case study can lead to the best rules to be provided as recommendations that can help Web designers to restructure their Web site, Web applications or even portals to better serve Web customers.

General Terms

Web Mining, Web Usage Mining

Keywords

Web Mining, Web Usage Mining, Web Recommendation, Preprocessing, Association Rules, Site Modification

1. INTRODUCTION

Web is a distributed hypertext-based information system. It is a globally interconnected network of hypermedia information and considered to be one way to utilize the infrastructure of the Internet/Intranet based on Hypertext Transfer Protocol (HTTP) which is a stateless protocol. With the huge amount of information available online, the Web is a fertile area for data mining research. Web mining can be defined as the discovery and analysis of useful information from the Web data so it is considered to be an application of data mining to large web data repositories that can be divided into three categories [1, 2]:

•Web Content Mining: Extracting useful information from the contents of Web documents.

•Web Structure Mining: Discovering structure information from the Web.

•Web Usage Mining: Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities.

In Web Usage Mining (WUM), data can be collected at: Server side, Client-side or Proxy servers. Server Side Web Data resides in Log files which are considered to be the best source for WUM. The data in the logs may be: unstructured, incomplete, noisy and Inconsistent, so there are three main tasks for performing WUM: Preprocessing, Pattern Discovery and Pattern Analysis [3].

Preprocessing is the most difficult task in the WUM process that is nearly 80% of mining efforts often spend to improve the quality of data [4, 5]. The typical Web server logs contain useful information that can be used in WUM (e.g. IP address, request time, HTTP method, URL of the requested files, HTTP version, return codes, the number of bytes transferred, the Referrer's URL and agents[6]. Indeed, there are irrelevant data needed to clean (e.g. accessorial resources, robots' requests and error requests). Preprocessing consists of four sub-phases: data cleaning, user identification, session identification and path completion [7].

In pattern discovery there are several kinds of mining activities that can be applied on web data after preprocessing such as: Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns and Dependency Modeling. These mining activities cover wide range of application areas such as: Personalization, System Improvement, Site Modification, Business Intelligence and Usage Characterization [3]. Association rules generation as a mining activity can be used to relate pages that are most often referenced together in a single server session. These pages may not be directly connected to one another via hyperlinks, so such rules can help Web designers to restructure their Web site.

In pattern analysis, the most common form is based on knowledge query mechanism such as SQL, or loading usage data into a data cube in order to perform OLAP operations (e.g. slice and dice, drill down, roll up, and pivot). Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data extracted from the chosen pattern discovery task. In association rules analysis, researchers depend in their studies on rules extraction process using different algorithms, other studies focus on rules reduction through introducing schemes for pruning irrelevant rules depending on either pre knowledge with the website's topology or depending on the interestingness measures (e.g. confidence, conviction, lift or leverage) to select the best rules according to user defined threshold, others depend in their work on comparing algorithms performance, some researchers introduce algorithms modification and others adopt semantic information using ontologies introducing semantically enhanced navigational patterns to effectively generate useful recommendations.

This paper is divided into 6 sections. Section 1 introduction to WUM, Section 2 introduces related work to this paper. Section 3 presents the proposed framework. Section 4 describes the dataset and the software development tools. Section 5 introduces the experimental results finally, section 6 gives the conclusion.

2. RELATED WORK

Web log files are the best source to predict user's behavior. Along with the useful information, the raw log files also contain entries for unnecessary details like image access, failed entries etc. which are of no use from the perspective of the WUM. Therefore, it becomes necessary to get rid of this irrelevant information.

Marathe Dagadu Mitharam [8] have described the web usage mining process focusing on the preprocessing phase through describing data fusion and cleaning, user identification, sessionization and path completion.

V.Chitraa and Antony Selvadoss Thanamani [9] stated that in the preprocessing, the data cleaning process includes removal of irrelevant records of graphics, videos, the format information, the records with the failed HTTP status code. In the proposed method the records accessed by robots are also cleaned.

L. Balaji and Dr. Y.S.S.R. Murthy [10] proposed a novel approach using universally accepted formatting language XML. In their approach text based log files are converted into XML format using parsers. Once a log file is in XML format, we can retrieve the required information such as user and session identification and the paths that are frequently accessed in an easy manner.

Mr. Akshay Upadhyay and Mr. Balram Purswani [11] tried to give a clear understanding of the data preparation process and pattern discovery process. They provide algorithms for data preparation, user identification and session identification.

Association rules as a mining activity can be used to relate pages that are most often referenced together in a single server session. Aside from association rules finding, schemes for pruning irrelevant rules must be implemented to identify and exploit those rules that are truly interesting to the user.

Maja Dimitrijević and Zita Bošnjak [12] conducted a comprehensive analysis of web usage association rules found in a website of an educational institution. In their experiments they proposed and implemented a set of schemes for pruning irrelevant rules.

Pinar Senkul and Suleyman Salin [13] developed a framework for integrating semantic information into Web navigation pattern generation process, where frequent navigational patterns are composed of ontology instances instead of Web page addresses.

R. Suguna and D. Sharmila [14] proposed association rule mining algorithm to find the user's interest for better web personalization and web recommendation. They proposed the technique for web page recommendation based on web usage mining.

Ashok Kumar D and Loraine Charlet Annie M.C. [15] proposed a new K-Apriori Algorithm to perform frequent itemset mining in an efficient manner. Apriori algorithm is used for K clusters and the frequent itemsets are generated from which Association rules are derived.

Mr. Rahul Mishra and Ms. Abha Choubey [2] used the FP-growth algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user's interest and found that the FP-growth algorithm is one of the fastest approaches for frequent item set mining.

In this paper we introduce a framework focusing on the three main interdependent tasks for performing WUM which are Preprocessing, Pattern Discovery and Pattern Analysis. In Preprocessing, we take into consideration studying the HTTP common methods (e.g. Options, Get, Head, Post, Put, Delete, Trace and Connect) as discussed by the World Wide Web Consortium (W3C). These methods can be used to differentiate between real user requests and administrative requests used for testing and diagnosing purposes. This research helps to remove the ambiguity found in such methods as some researchers depend on the succeeded requests with GET method [7, 16-18], other researchers exclude one or more of the other methods as Put, Head or Connect methods [5, 19-21]. However, according to the development of the websites or web applications, real users' requests are those requests using GET method or POST method in case there is no significant form parameters to be transferred to another page. Indeed, by excluding other methods we can improve the preprocessing phase. In Pattern Discovery, as the user requests are residing in the log file chronologically, we randomly divide the preprocessed dataset into two equal datasets then we generate sets of Association Rules from both datasets. In Pattern Analysis, first we measure the similarity in both of the datasets depending on the similarity in both sets of generated rules, second we introduce an algorithm to calculate the coverage of the rules generated from a dataset on the patterns of the other leading to the best, valid and reduced set of rules that achieves the highest coverage that can provide recommendations for administrators to restructure their websites. Our experimental results confirmed that the analysis of association rules coverage in our case study can lead to actions that can help Web designers to restructure their Web sites, Web applications or even portals to better serve Web customers.

3. PROPOSED FRAMEWORK

In this section, we present our methodology for building our proposed framework taking into consideration the three main interdependent tasks for performing WUM which are Preprocessing, Pattern Discovery and Pattern Analysis.

3.1 Preprocessing

After studying the web log file and its contents, first, loading web log file into RDBMS: we load the unstructured log file into RDBMS according to the delimiters separating the data fields to be in a structured format

Second, Page Aliasing: Extract all visited documents from the web log file to be imported in a separate table, followed by pages aliasing for each visited document, so the documents will appear as P1, P2, P3, etc. for simplification

Third, Create new fields (RecId, UserId, Timestamp, SessionId, etc...): For each transaction representing a hit in the log file, RecId to be the primary key, UserId to identify user, Timestamp for each transaction and SessionId to identify a user's sessions

Fourth, User Identification: Identifying UserId according IP address and the Agent which is carrying on some information about the user's browser, version and operating system

Fifth, Date Format Adjustment: Validating date/time data type to extract the timestamp of each transaction according to the format of the used RDBMS

Sixth, Session Identification: Break down the requests belonging to each user into sessions according to Time Oriented way depending on time stamps assuming the difference between the first request and the last one to be < =30 minutes

Seventh, Data Cleaning: Removing all those irrelevant entries including: Common HTTP methods except GET and POST methods, Accessorial Resources embedded in web documents, Robots' Requests and Error Requests

Eighth, Data Aggregation into sessions: Database view is extracted to aggregate transactions by grouping hits into sessions, each session having: UserId, No. of Hits, Total Time Interval, Total bytes Transferred, Click Stream and Session Referrer

Finally, Excluding sessions where Hits < 2: Exclude all those sessions having no. of hits < 2 to filter out suitable sessions for association rules mining. The Preprocessing task is applied as depicted in Figure 1.

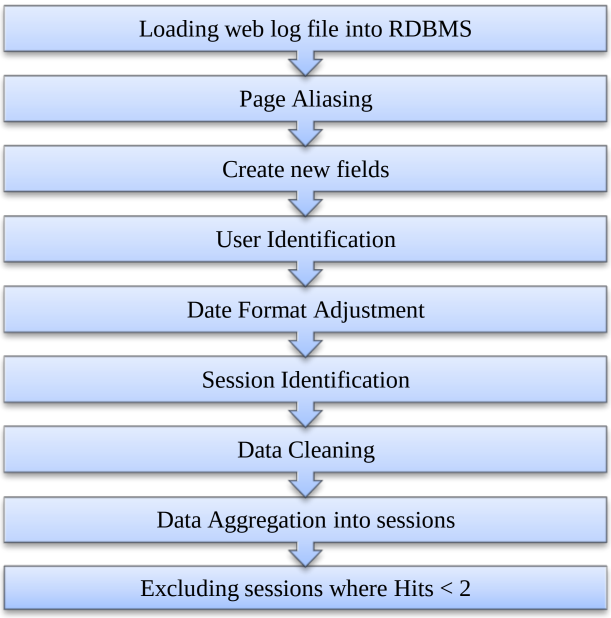

Figure 1: Preprocessing Task

3.2 Pattern Discovery

The pattern discovery task is applied based on Association Rules that can be used to relate pages that are most often referenced together in a single server session.

First, programmatically divide the preprocessed log randomly into two datasets DS1 and DS2

Second, using WEKA software to build a knowledge flow based on Apriori algorithm which is widely used algorithm in association rules mining.

Third, programmatically generate ARFF files for DS1 and DS2 based on the click stream field in each dataset. Each ARFF file has a header containing a list of all attributes followed by patterns of transactions.

Finally, Association Rules Extraction: Use the generated ARFF files as input for the Knowledge Flow as shown in Figure 2. Extract n sets with same no. of association rules at different threshold values for support and confidence. The Pattern Discovery task is applied as depicted in Figure 3.

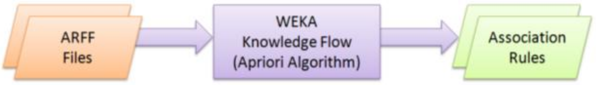

Figure 2: Association Rules Generation

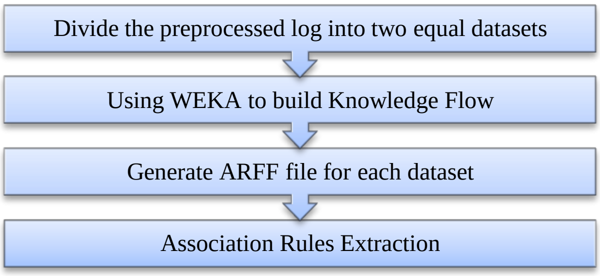
Divide the preprocessed log into two equal datasets
Using WEKA to build Knowledge Flow
Generate ARFF file for each dataset
Association Rules Extraction

Figure 3: Pattern Discovery task

The support supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set, while the confidence of a rule is defined as:

$$conf(X \rightarrow Y) = supp(X \cup Y) / supp(X) \qquad (1)$$

The confidence value is one of the widely used interestingness measures to identify the strength of the rule. It is one of the advantages of using WEKA that it can generate different interestingness measures with each rule.

3.3 Pattern Analysis

In pattern analysis, the main target here is first, to measure the similarity between association rules generated from DS1 and DS2, second, calculate the coverage of the rules generated from DS1 on the patterns of DS2 leading to the best, valid and reduced set of rules that achieves the highest coverage excluding those irrelevant ones. We load the extracted sets of generated rules into the RDBMS to facilitate comparing both sets of rules with each other using set of tailored SQL queries. Applying visualization techniques, such as graphing to visually present the association rules generated from both datasets.

First, programmatically load the generated sets of association rules from both of the datasets into RDBMS.

Second, measure similarities between association rules generated from DS1 and DS2.

Third, calculate the coverage percentage and the distinct coverage percentage for each set of the generated rules from DS1 with the click stream patterns of DS2.

Fourth, we represent the sets of rules into graph to visually show the best and valid set of rules generated from both of the datasets that achieves highest coverage.

Finally, Recommendation: Introducing the best and valid set of association rules generated from both datasets ordered in a descending order according to confidence. The Pattern Analysis task is applied as depicted in Figure 4.

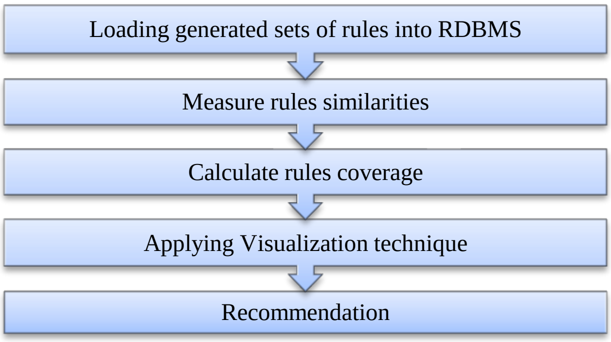
Loading generated sets of rules into RDBMS
Measure rules similarities
Calculate rules coverage
Applying Visualization technique
Recommendation

Figure 4: Pattern Analysis task

Rules Coverage Percentage is calculated according to:

$$Rules\ Cov. = \frac{\text{Patterns No. Covered in DS2}}{\text{DS2 Patterns No.}} \quad (2)$$

Rules Distinct Coverage Percentage is calculated according to:

$$Dist. Rules\ Cov. = \frac{\text{Dist. Patterns No. Covered in DS2}}{\text{DS2 Distinct Patterns No.}} \quad (3)$$

4. DATASETSET DESCRIPTION

In our proposed framework for experimental purposes, we have used the log file that has been used by Maja Dimitrijević and Zita Bošnjak [12]. We have used the log file published online with their results. They used WumPrep which is a part of Open Source Project HypKnowSys consisting of a set of Perl scripts. The log file is containing web requests on the official website of an educational institute carrying on 5999 web requests on an arbitrarily selected day that was not mentioned if there were any special activities or events at the institution on that day which could have led to any unusual behavior of the website's users. According to the Web Usage Mining Taxonomy Dimensions Table 1 shows the five major dimensions applied to our research.

Table 1: Taxonomy Dimensions

Data Sources	Input Data	Users	Websites	Application
Single Site, Multi Users	Server-Side (Web Server Log)	390	1	Association Rules

5. EXPERIMENTAL RESULTS

In our approach, data cleaning which is the most important stage in preprocessing the web log file is based on removing all irrelevant entries. Table 2 shows the different types of irrelevant entries, while Table 3 shows the excluded HTTP common methods.

Table 2: Irrelevant Entries

Total Requests	Get/Post Methods	Excluded Methods	Accessorial Resources	Robots' Requests	Error Requests
5999	5881	118	3646	16	1302

Table 3: Irrelevant HTTP common methods

Excluded Methods	Options	Head	Connect	Put	Delete	Trace	Others
118	85	10	0	0	0	0	23

There exist 118 requests (1.97%) out of the 5999 requests to be excluded for the reason of not being real users' requests.

Our preprocessing approach introduces better results for being more reduced by 1.65% compared with the preprocessing results mentioned in the related work. Aside from data cleaning, our approach shows better results in the session identification leading to 6.10% reduction in the number of sessions compared with the related work. Moreover, by excluding sessions where hits<2 to filter out suitable sessions for association rules mining, this introduces more log reduction. Table 4 shows the enhancements in the Preprocessing results.

Table 4: Preprocessing Results

Phases	Related Work	Reduced	Our Approach	Reduced By	Diff.
Data Cleaning	2122	-64.63%	2087	-65.21%	-1.65%
Users Identification	N/A	N/A	390	N/A	N/A
Sessions Identification	426	N/A	400	-6.10%	N/A
Excluding Sessions (hits<2)		N/A	292	-31.46%	N/A

In Pattern Discovery, we randomly divide the preprocessed log into two equal datasets each of 146 users' sessions then programmatically generate the ARFF files from both datasets patterns. The ARFF files are used as input for the built knowledge Flow using WEKA software. Table 5 shows that each dataset is valid for pattern discovery and shows the similarities in generated rules to validate the dataset division process are not biased.

Table 5: Generated Rules Similarities

Exp.	No. of Rules	Similar Rules	Percentage	Equal Conf.
1	25	0	0	0
2	50	41	82	4
3	75	44	58.7	4
4	100	72	72	5
5	125	87	69.6	6
6	150	101	67.3	6
7	175	112	64	6
8	200	128	64	6

We notice that in experiment 1, when comparing the 25 rules generated from both datasets there is no any similar rules, while in experiment 2, when comparing 50 rules leads to 41 similar rules discovered exactly from each of the datasets and achieving the highest similarity percentage among our experiments with value 82%. Moreover, having 4 rules exactly discovered with exactly the same confidence value. In Pattern Analysis, we notice that the number of similar rules extracted from both of the datasets is directly proportional with the number of extracted rules. Consequently, we can calculate the average similarity percentage after excluding the first result with 0%. We found the average similarity 68.23% which is an indicator to what extent of both of the datasets DS1 and DS2 are similar. The generated rules similarities are depicted in Figure 5.

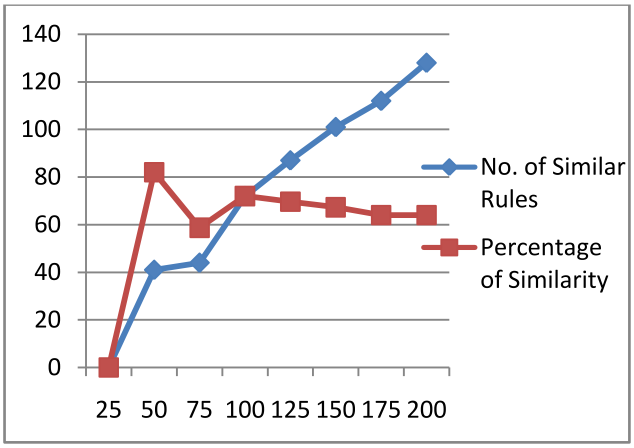

Figure 5: Generated Rules Similarity

Moreover, an experiment is applied 8 times, using the sets of rules generated from DS1 to check their coverage percentage and distinct coverage percentage on the second dataset DS2. The results can be shown in Table 6.

Table 6: Rules Coverage Results

Exp.	No. of Rules	Coverage.	Cov. %	Distinct Cov.	Distinct Cov. %
1	25	57	39%	52	41%
2	50	58	40%	53	41%
3	75	82	56%	74	58%
4	100	83	57%	75	59%
5	125	85	58%	77	60%
6	150	104	71%	96	75%
7	175	106	73%	98	77%
8	200	113	77%	105	82%

Figure 6 visually present the coverage percentage and distinct coverage percentage of the generated association rules.

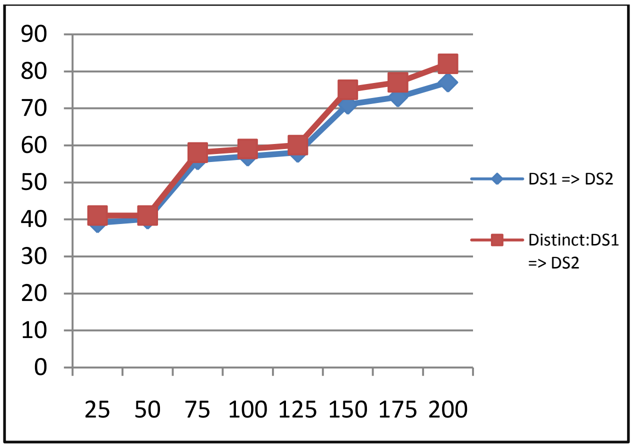

Figure 6: Rules Coverage Percentage

The 8th run shows 200 rules generated from DS1 and achieving the highest coverage on DS2. These rules are categorized into categories according to confidence values as shown in Figure 7. We notice that the best category is the category including those 36 association rules at the highest Confidence value which is 1. Consequently, those rules achieving highest Confidence value will be provided as a recommendation for the administrator to be taken into consideration in site modification.

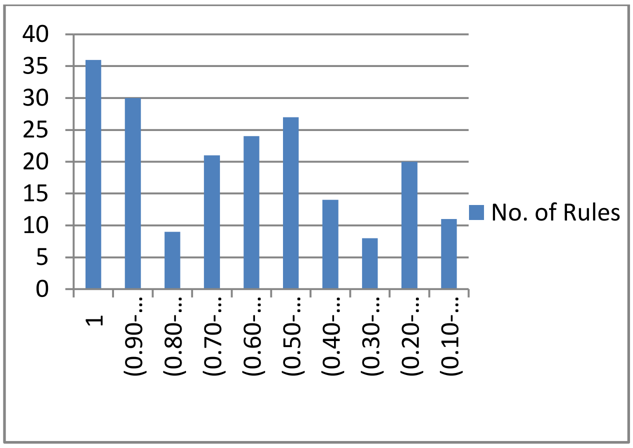

Figure 7: Rules Coverage Percentage

6. CONCLUSION

Our research introduces a Web Recommendation Framework for Site Modification as one of the applications of Web Usage Mining (WUM) that is applicable for online business and marketing applications. The framework is designed as a recommendation framework consistent with the Collaborative filtering approach. It is based on usage patterns discovery and analysis describing users' behaviors and predicting what users will like based on their similarity to other users without relying on machine analyzable content. Therefore, it is capable of accurately recommending items without requiring an understanding of the item itself. The framework focuses on the three main interdependent tasks for performing WUM which are Preprocessing, Pattern Discovery and Pattern Analysis, taking into consideration removing the ambiguity found in some researches concerning the HTTP common methods. Moreover, the framework introduces a set of association rules that achieves highest coverage for the dataset independent of the website's topology. Our experimental results confirmed that calculating association rules coverage in our case study can lead to the best association rules as recommendations that can help Web designers to restructure their Web based applications to better serve Web customers.

7. REFERENCES

[1] M. A. Bayir, "A new reactive method for processing web usage data," M.Sc., Computer Engineering, Middle East Technical University, 2006.

[2] M. R. Mishra and M. A. Choubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining," *International Journal*, vol. 2, pp. 311-318, 2012.

[3] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, pp. 12-23, 2000.

[4] D. Padmabhushana and D. Srikanth, "Predicting Software Bugs Using Web Log Analysis Techniques and Naïve Bayesian Technique," *International Journal of Computer Trends and Technology*, vol. 3, pp. 185-191, 2012.

[5] M. Seema and M. P. Makkar, "An Approach to Improve the Web Performance By Prefetching the Frequently

Access Pages," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* vol. 1, pp. 625-634, 2012.

[6] T. Revathi, M. M. Rao, C. S. Sasanka, K. J. Kumar, and B. U. Kiran, "An Enhanced Pre-Processing Research Framework for Web Log Data," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, pp. 358-363, 2012.

[7] S. Anand and R. Rani Aggarwal, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions," *International Journal of Computer Applications,* vol. 48, pp. 13-18, 2012.

[8] M. D. Mitharam, "Preprocessing in Web Usage mining," *International Journal of Scientific & Engineering research,* vol. 3, pp. 1-7, 2012.

[9] V. Chitraa and A. S. Thanamani, "Web Log Data Cleaning For Enhancing Mining Process," *International Journal of Communication and Computer Technologies,* vol. 01, pp. 49-55, 2012.

[10] L. Balaji and Y. Murthy, "An Effective Web Usage Mining," *International Journal of Electronics Communication and Computer Engineering,* vol. 3, pp. 281-286, 2012.

[11] M. A. Upadhyay and M. B. Purswani, "Web Usage Mining has Pattern Discovery," *International Journal of Scientific and Research Publications,* vol. 3, pp. 1-4, 2013.

[12] M. Dimitrijević, Z. Bošnjak, and S. Subotica, "Discovering interesting association rules in the web log usage data," *Interdisciplinary Journal of Information, Knowledge, and Management,* vol. 5, pp. 191-207, 2010.

[13] P. Senkul and S. Salin, "Improving pattern quality in web usage mining by using semantic information," *Knowledge and information systems,* vol. 30, pp. 527-541, 2012.

[14] R. Suguna and D. Sharmila, "Association Rule Mining for Web Recommendation," *International Journal on Computer Science and Engineering,* vol. 4, pp. 1686-1690, 2012.

[15] A. Kumar and L. Charlet Annie MC, "Web Log Mining using K-Apriori Algorithm," *International Journal of Computer Applications,* vol. 41, pp. 16-20, 2012.

[16] V. Verma, A. Verma, and S. Bhatia, "Comprehensive Analysis of Web Log Files for Mining," *International Journal of Computer Science Issues(IJCSI),* vol. 8, pp. 199-202, 2011.

[17] P. Patil and U. Patil, "Preprocessing of web server log file for web mining," *World Journal of Science and Technology,* vol. 2, pp. 14-18, 2012.

[18] B. C. Palmer, "Web Usage Mining: Application To An Online Educational Digital Library Service," Ph.D., Instructional Technology & Learning Sciences, Utah State University, 2012.

[19] R. Suguna and D. Sharmila, "User Interest Based Web Usage Mining using a Modified Bird Flocking Algorithm," *European Journal of Scientific Research,* vol. 86, pp. 218-231, 2012.

[20] A. S. Lalani, "Data mining of web access logs," M.Sc., Computer Science Department, Royal Melbourne Institute of Technology, 2003.

[21] F. Khalil, J. Li, and H. Wang, "Integrating recommendation models for improved web page prediction accuracy," in *Proceedings of the thirty-first Australasian conference on Computer science-Volume 74*, 2008, pp. 91-100.