International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

Prediction of Web Usages Using Apriori Algorithm

Thu Zar Htet
[1]dawthuzarhetisdept@gmail.com

***Abstract**:*

Web mining has become a necessity to use efficient information retrieval techniques to find and order the desired information. Although there exists quite some confusion about the Web mining, the most recognized approach is to categorize Web mining into three areas: Web content mining, Web structure mining, and Web usage mining. Web usage mining is one of the most popular web mining techniques in order to generate the web usage patterns which can be further exploited in better personalization, improving navigations, recommendations, and recognition of web sites and attracting more advertisements etc. Web Usage Mining is a great research area in discovering the interested patterns of user's usage data on the web. It focuses on the techniques that could predict user behavior while the user interacts with Web. Frequent pattern mining is an important knowledge discovery technique in data mining. Most frequent pattern mining has been designed with the traditional support-confidence framework that generates more interesting kinds of patterns. This system predicts the web usage using the Apriori algorithm which is also discovered the frequent patterns with support measure. In this system, Apriori algorithm is applied on NASA web log data in order to predict the usage patterns. This system accepts the NASA web log and preprocesses these data to improve data quality and produce the usages patterns. Finally, this system assesses the performance of Apriori algorithm. The approach used in this system, helps the website designers to improve their website usability.

Keywords

Web Usage Mining, Web Log Data, Apriori Algorithm.

1. Introduction

World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. The growth of web is tremendous as approximately one million pages are added daily. Due to these huge, unstructured and scattered amounts of data available on web, it is very tough for users to get relevant information in less time. To achieve this, improvement in design of web site, personalization of contents, prefetching and caching activities are done according to user's behavior analysis. The ability to know the patterns of users' habits and interests helps the operational strategies of enterprises. Various applications like e_commerce, personalization, web site designing, recommender systems are built efficiently by knowing users navigation through web.

Web mining is the application of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. The objects of Web mining are vast, heterogeneous and distributing documents. The logistic structure of Web is a graph structured by documents and hyperlinks, the mining results may be on Web contents or Web structures. Web mining is divided into three types. They are Web content mining, Web structure mining and Web usage mining. Web usage mining is one of the applications of data mining which is used to mine of log files to discover useful patterns which can be further exploited in better personalization, improving navigations, recommendations, and recognition of web sites and attracting more advertisements etc. Every time when a server of a website receives a request from web user and the usage data captures the identity or origin of web users along with their browsing behavior at a web site. User's activities can be captured into a special file called log file. There are various types of log: Server log, Proxy server log, Client/Browser log. These log files are used by web usage mining to analyze and discover useful patterns. Frequent pattern mining is an important knowledge discovery technique in data mining. Most frequent pattern mining has been designed with the traditional support-confidence framework that generates more interesting kinds of patterns. This specialized framework may use different types of interestingness measures, model negative rules, or use constraint-based frameworks to determine more relevant patterns. In the basic model of frequent patterns, a pattern (or an itemset) is considered frequent if it satisfies the user-defined minimum support (min-sup) constraint. The min-sup constraint controls the minimum number of transactions a pattern must cover in a database. Since only a single min-sup constraints used for the entire dataset, the model implicitly assumes that all items in a database have uniform frequencies.

However, this is often not the case in many real-world databases. In many real-world applications, some items appear very frequently in the data, while others rarely appear. It has to be noted that considering an item in a database as either frequent or rare is a subjective issue which depends on the user and/or application requirements. To discover the useful frequent pattern, Apriori algorithm is applied. Finally, this system assesses the performance of Apriori algorithm.

2. Related Work

Due to the rapid usage of World Wide Web, websites are the information provider to the Internet users. Storing and retrieving the information from the web is always a challenging task. Web mining, the term is defined as extract needed information to the users from the Web. The information provided by the Web is not only the exact information of user needs but also suggest the information associated to the exact one. The author in this paper [3] introduces the applications and the mining process of data mining tool (open source) Rapid miner.

They proposed work analyzes the usage of web pages (i.e. Browsing behavior of user) using two different clustering algorithms such as k-means, which is incorporated in the tool and Fuzzy c means(FCM) clustering using Rapid Miner. The results showed operational background of FCM clustering and k-means clustering algorithm based on the cluster centroid.

Web Usage Mining techniques are great area of research these days. Providing users what they are looking for in websites is the ultimate aim of web usage mining. In the approach of [8], the aim is fulfilled by using association rule mining technique on clustered data i.e. data applied clustering techniques first and then applied association rule technique for frequent accessed set of link. Basic Association Rule Mining has drawback of generation of irrelevant rules, generation of too many rules leading to contradictory prediction resulting in reduction of accuracy.

Minimum support and minimum confidence parameters can be set in such a way to eliminate false discoveries. But when minimum support is too small, every rule will get a chance to be true, leading to wrong result and when minimum support is too large, for small data set, wrong prediction may occur. Clustering frequent access patterns reduce dataset for Association Rule Mining and improve result accuracy and producing results of pattern discovery of web usage mining process effective.

Analyzing the web log files through web usage mining is very important to discover the similar behavior users of particular website. The paper [1] discussed how to find useful knowledge from web log file using some data mining technique like Association rule mining and clustering. First they preprocessed the web log file then applied association rule mining and clustering algorithm on web log file to discover usage pattern and same behavioral users. The approach used in this paper [1], helps the website designers to improve their website usability.

Continued growth of user number and size of shared content on Web sites cause the necessity of automatic adjusting content to users' needs. In the literature of Web Mining, such actions are referred to personalization and recommendation which led to improve the visibility of presented content. To perform adequacy actions which correspond to the expected users' needs, [5] utilized web server log files. Mining such data with accurate constraints can lead to the discovery of web user navigation patterns. Such knowledge is used by personalization and recommendation systems (PRS) due to performed actions against user behavior during a visit on the web portal. In this paper [5], they presented the system framework for mining web user navigation patterns in order to knowledge management and focused on constraints which are critical factors to evaluate the effectiveness of the implemented algorithm. On the other hand, these constraints can be perceived as knowledge validation criteria due to its adequacy. Thus only adequate knowledge can be added to existing in PRS knowledge base.

3. Web Mining

Web mining is the application of data mining techniques to extract uncovers relevant, hidden information on web. Web mining can be categorized into three classes based on content, structure and usage of web pages. Three areas of web mining are:

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

3.1 Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining generally uses basic data mining algorithms such as Association rule mining, Sequential rule mining, Clustering, Classification etc. for pattern discovery phase [6]. Due to high raise in number of transactions, Association rule mining is the most basic data mining technique to be used in web usage mining to find association between web pages. It refers to the set of pages that are accessed together in a single server session. This information can be useful to restructure the web site.

Web usage mining itself can be classified further depending on the kind of usage data considered web server data and application server data. User logs are collected by the web server and typically include IP address; page reference and access time are called web server data. Commercial application servers such as Web logic, Story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

3.2 Web Log Data

Web log data contains various parameters related to web server activity which are analyzed to extract useful information. Three main data sources are used to collect log data for web usage mining. Those are Server log, Proxy server log, and Client/ Browser log.

3.3 Log File Formats

Different web servers provide various format of log files such as Common log format, IIS standard/extended log format, Combined/Extended common log format, Log markup language (LogML), because of different setting parameters. Among them common log format are commonly used.

Common log format is a standard non-customized format (fixed no of attributes) suitable for http web sites. This type of log includes user’s IP address/hostname, rfcname, log name, date with time zone, page access method, PATH, http version, server response code and byte received. Extended log format is a customizable log file format which can add some additional attributes like referrer-url, http-user-agent and cookies. Combined/Extended common log format is the combination of common log format and extended log format but in this format multiple directories can be created for access logs.

3.4 NASA Log File Format

NASA log file format has the following parameters:

- IP address/hostname
- rfcname
- log name
- date with time zone
- Page access method
- PATH
- http version
- Status code
- Byte received

3.5 Prediction of Web Usages Using Apriori Algorithm

This system predict the web usage using the Apriori algorithm which is also discovered the frequent patterns with support measure.

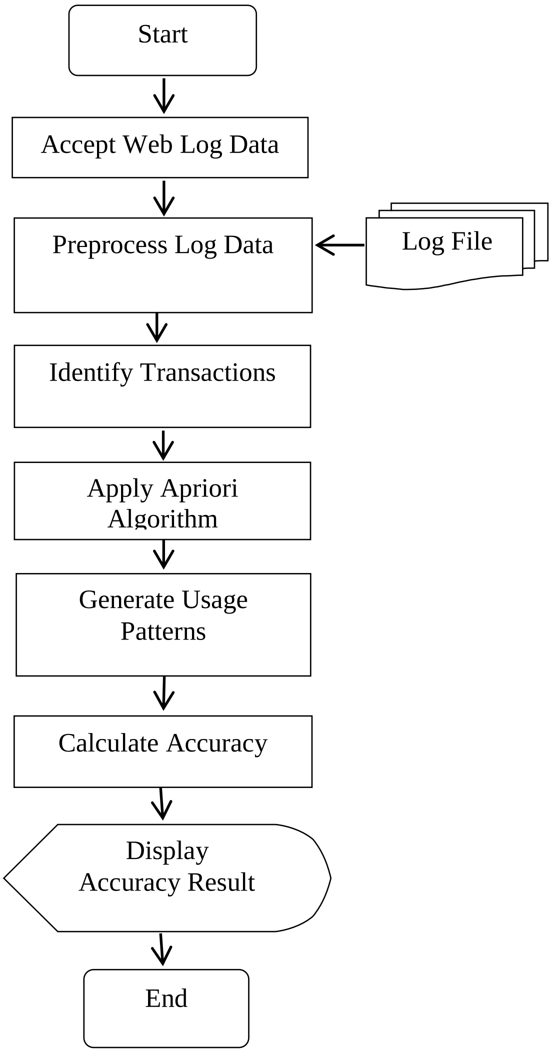

Figure 3.1 System Flow Diagram of Prediction of Web Usages Using Apriori

In this system, Apriori algorithm is applied on NASA web log data in order to predict the usage patterns. Firstly, this system accepts the NASA web log and preprocesses these data to remove the unnecessary data and to improve data quality. In preprocessing step include the data cleaning, user identification, session identification, transaction identification and formatting. After processing these phase, this system produces the usages patterns and then calculates the performance of Apriori algorithm and the system flow diagram is shown in Figure 3.1.

3.6 Data Preprocessing Process

Data preprocessing process performs web data into consistent data. There is a distinction between the explicit information and implicit information. Whereas data in the explicit information is collected in a proper form of the statistical analysis, beside the data needs a pre-processing process. This process can be done in several activities: Data Cleaning, User Identification, Session Identification and Transaction Identification. These activities are described in the following section.

3.6.1 Data Cleaning Process

Some information from web log files are redundant and irrelevant. So the removing is necessary in data cleaning task. The cleaning firstly removes the unsuccessful records in which the status code field recognizes the unsuccessful http request. It removes the entries with status codes that are not 200. Removing these kinds of requests is necessary because they are just increasing the size of log file and nothing to do with analysis of user's navigational behavior.

3.6.2 User Identification Process

There are several methods to identify unique user. User identification is one of the complicated tasks due to existence of local/external proxy servers, cache systems, cooperate firewalls and shared internet. IP address is used to identify the unique. IP address is logged into log file when a user hits a page and it can be used to identify different users. But in case of proxy server when many users request a particular page then web site server logged same IP address (Proxy server IP) into the log file. Practically different users are accessing that page. Caching also creates problem to identify unique user. Whenever a user tries to access previously accessed page, browser display pages from local cache and no entry are logged into the log file.

3.6.3 Session Identification Process

To find all page references made by user, session identification process is used. These two methods are also called as "proactive" and "reactive" methods. The session considers over the duration of user request to a particular website. When the time gap between two consecutive requests by the same user is greater than certain threshold then a new session is created. If the time between page requests exceeds a certain limit, it will assume that other user-session has started. This system takes 30 minutes threshold value. After preprocessing tasks, the useful log file used to identify the transaction.

3.6.4 Transaction Identification Process

According to the needs of the respective algorithms, transactions are applied to extract important information from the preprocessed data. The formatting of transactional data differs from the kind of algorithms that are used. The transactions are identified with serialization of numeric data and the related transactions are extracted.

3.7 Apriori Association Rule Mining

The Apriori is a computationally expensive algorithm because the frequent patterns discovered with the support measure. That is, although a pattern satisfies the user-defined minimum support threshold. It includes generating huge number of candidate patterns and multiple scans on the database.

3.7.1 Apriori Algorithm

The Apriori algorithm is as follows [2]:
Let I = {i1, i2, …, in} be a set of items, and DB be a database that consists of a set of transactions. Each transaction T contains a set of items such that T ⊆ I. Each transaction is associated with an identifier, called TID. Let X ⊆ I be a set of items, referred as an itemset or a pattern. A pattern that contains k items is a k-pattern. A transaction T is said to contain X if and only if X ⊆ T. The frequency (or support count) of a pattern X in DB, denoted as f(X), is the number of transactions in DB containing X. The support of X, denoted as S(X), is the ratio of its frequency to the DB size, i.e., S(X) = f(X)/|DB|. The pattern X is frequent if its support is no less than the user-defined minimum support (minsup) threshold. That is, X is said to be frequent if S(X) ≥ minsup.

3.8 Performance Measurement

To assess the performance of RSA algorithm, the measures of support and confidence are used [4].

3.8.1 Support

It measures the frequency of association, i.e. how many times the particular item has been occurred in a dataset.

$$Support = P(A \cap B) \qquad (3.1)$$

P (A∩B) is equal to the number of transactions containing both A and B/Total number of transactions.

3.8.2 Confidence

Confidence basically measures the strength of the association rules. It is defined as the fraction of the transactions that include both A and B to the total number of records that contain A. It determines how frequently item B occurs in the transaction that contains A. Confidence expresses the conditional probability of an item.

$$Confidence = P(A \mid B) = P(A \cap B) / P(A) \quad (3.2)$$

3.8.3 Predictive Accuracy

Predictive accuracy is also another way to measure interestingness of an association rule. The definition of predictive accuracy is: Let D be a data file with n number of records. If [a → b] is an Association Rule which is generated by a static process P then the predictive accuracy of [a →b] is c([a → b]) = Pn[n satisfies b |n satisfies a] where distribution of n is govern by the static process P and the Predictive Accuracy is the conditional probability of a → n and b → n.

4. Implementation

Step 1: This system accepts one hour NASA log data which has 1396 transactions with 150KB.

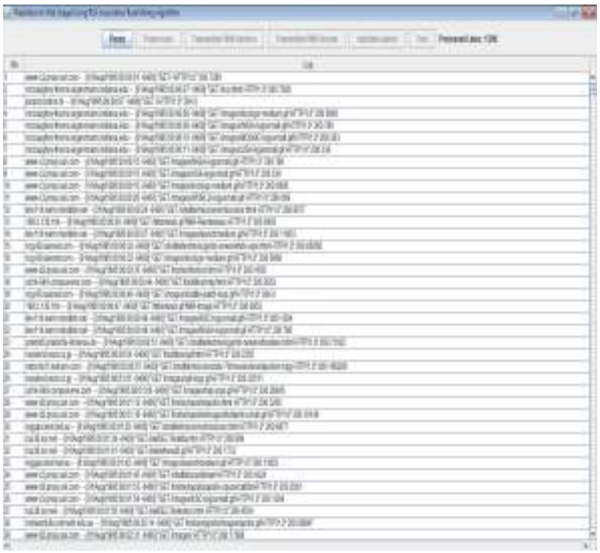

Figure 4.1: NASA log file format for an hour

Step 2: The log file format is a text-based format, the NASA log file also text- based. So, this file is changed into the useful format.

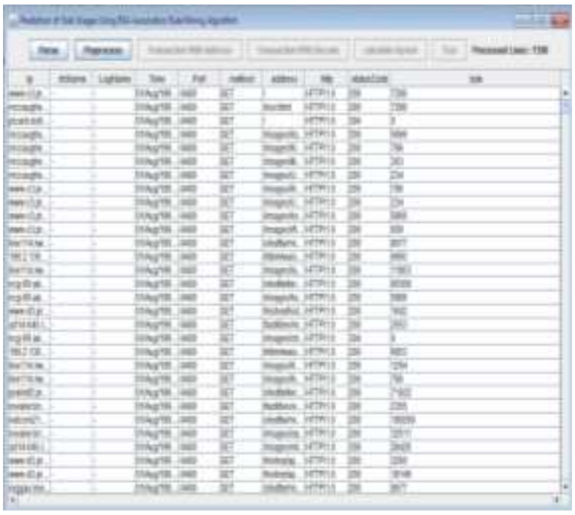

Figure 4.2: Parsing the data with relevant format

Step 3: Some information from web log files are redundant and irrelevant. So the removing is necessary in data cleaning task. The cleaning firstly removes the unsuccessful records in which the status code field recognizes the unsuccessful http request. It removes the entries with status codes that are not 200. Removing these kinds of requests is necessary because they are just increasing the size of log file and nothing to do with analysis of user's navigational behavior. Apply the preprocessing on web log files and store them into the database. The preprocess log file is shown in Table 4.3 which has 1257 transactions.

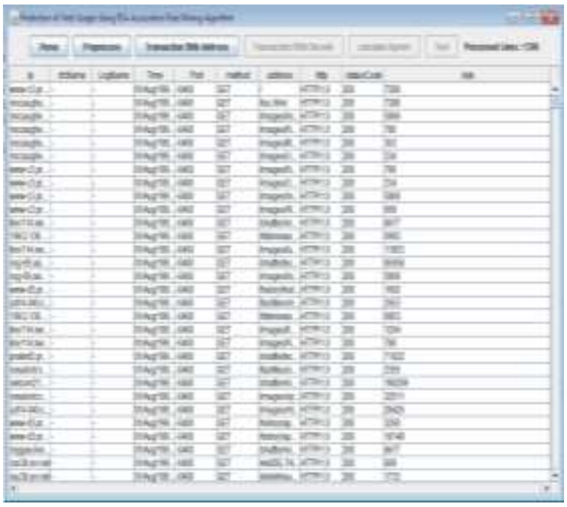

Figure 4.3: Preprocessed log file

Step 4: To find all page references made by user, session identification process is used. The session considers over the duration of user request to a particular website. When the time gap between two consecutive requests by the same user is greater than certain threshold then a new session is created. If the time between page requests exceeds a certain limit, it will assume that other user-session has started. This system takes 30 minutes threshold value. After preprocessing tasks, the useful log file used to identify the transaction which has 146 transactions.

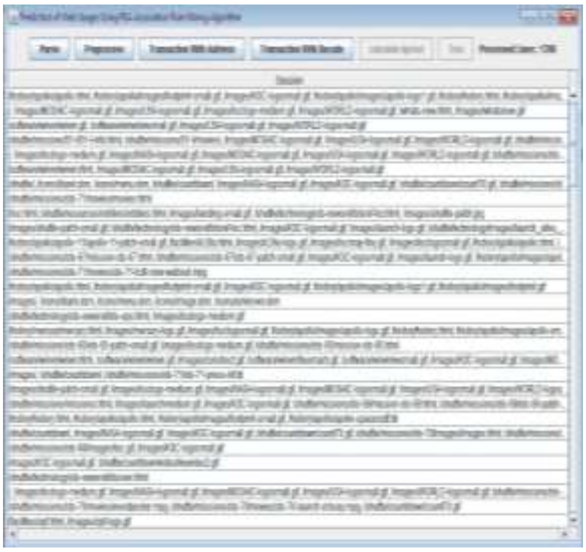

Figure 4.4: Transactions of Sessions

Step 5: This process involves determining frequent patterns. Association rules are used for prediction of next event or discovery of associated event. In the web data set, the transaction consists of the number of URL visits by the client, to the web site. To find the associated pattern Apriori Algorithm is used.

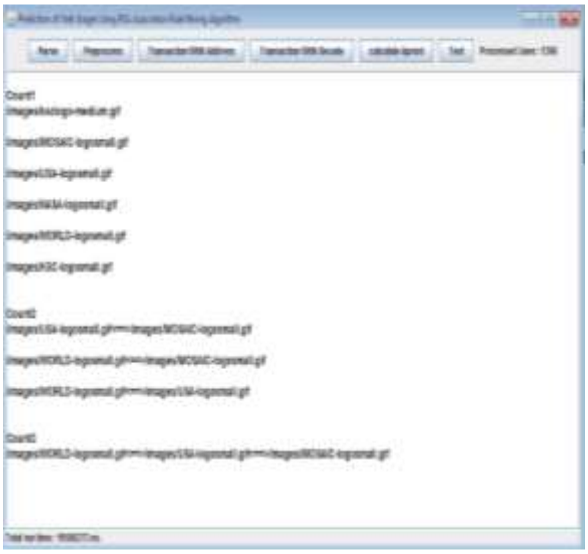

Figure 4.5: Rules produced by implementing with Apriori algorithm

Step 6: Finally, this system access the performance of Apriori algorithm.

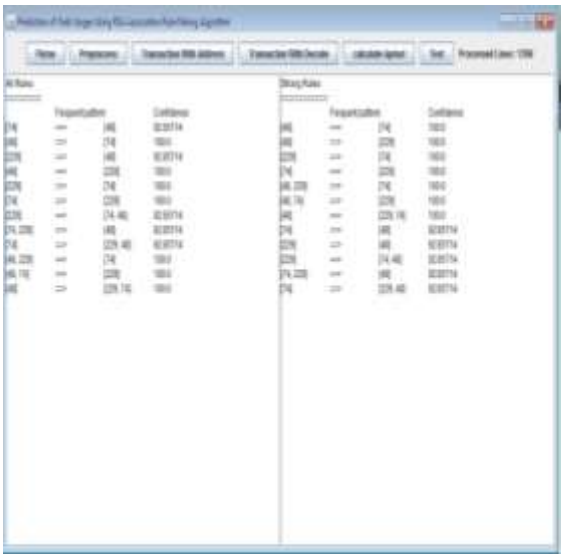

Figure 4.6: Correct rules with confidence

This algorithm implemented an hour NASA data, there are 1396 transactions. The accuracy is tested on 12 corrected rules, the minimum support count is 0.28 and relative minimum support count is 0.65. The values of confidence are illustrated in Table 4.4 and the average of confidence is 97.02%.

5. Conclusion

Web applications are increasing at an enormous speed and its users are increasing at exponential speed. The evolutionary changes in technology have made it possible to capture the user's essence and interactions with web applications. Web mining is one of the major and important fields of data mining. Data mining techniques are applied on contents, structures and on log files of web sites to achieve performance, web personalization and schema modifications of web sites. Web mining is divided into three categories such as Web Content Mining, Web Structure Mining and Web Usage Mining.

In web usage mining (WUM) or web log mining, user's behavior or interests are revealed by applying data mining techniques on web log file. Web log file is saved as text (.txt) file. Due to large amount of "irrelevant information" in the web log, the original log file cannot be directly used in the web usage mining (WUM) procedure.

Web Usage Mining is a great research area in discovering the interested patterns of user's usage data on the web. In this system, implementation of a system is pattern discovery using association rules which introduce the process of web log mining, and show how to find frequent pattern from the web log data in order to obtain useful information about the user's navigation behavior when the user browses or makes transactions on the web site.

Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on the threshold value called support, identifies the frequent item sets. In Apriori algorithm, the frequent patterns discovered with the support measure. So, it is a computationally expensive. That is, although a pattern satisfies the user-defined minimum support threshold. The approach used in this system, helps the website designers to improve their website usability.

6. Acknowledgements

I would like to express my special thanks to **all my teachers** who gave me their time and guidance, and all my friends who helped in the task of developing this paper. Finally, I would like especially to thank **my parents** for their continuous support and encouragement throughout my whole life.

7. References

[1] A. M. Parekh, A. S. Patel, S. J. Parmar, and Prof. V. R. Patel, “Web usage Mining: Frequent Pattern Generation using Association Rule Mining and Clustering”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 4, Issue 4, April 2015.

[2] H. Yun, D. Ha, B. Hwang, and K. H. Ryu, “Mining association rules on significant rare data using support”, J. Syst. Softw., 67:181–191, September 2003.

[3] M. Santhanakumar, and C. Christopher Columbus, “Web Usage Based Analysis of Web Pages Using Rapid Miner”, E-ISSN: 2224-2872, Vol. 14, 2015.

[4] M. Sharma, J. Choudhary, and G. Sharma, “Analysis of the Performance of Various Algorithms and Interestingness Measures in Association Rule Mining”, 2012.

[5] P. Weichbroth, M. Owoc, and M. Pleszkun, “Web User Navigation Patterns Discovery from WWW Server Log Files”, Proceedings of the Federated Conference on Computer Science and Information Systems, ISBN 978-83-60810-51-4, Page(s): 1171-1176.

[6] R. Garage, and P.K. Mishra, “Web Usage Mining: A Survey”, International Journal of Computer Applications (0975 – 8887), Vol. 97, No. 18, July 2014.

[7] R. U. Kiran, and M. Kitsuregawa, “Towards Efficient Discovery of Frequent Patterns with Relative Support”.

[8] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, “Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery”, International Journal of Data Mining Techniques and Applications, ISSN: 2278-2419, Vol. 2, Issue 1, June 2013. http://iirpublications.com.