Volume 2, Issue 9, September 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining

Mr. Rahul Mishra
Computer Science & CSVTU
India

Ms. Abha Choubey
Computer Science & CSVTU
India

Abstract*— Web usage mining refers to the automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with web resources on one or more web sites. It consists of three phases which are data Preprocessing, pattern discovery and pattern analysis. In the pattern discovery phase, frequent pattern discovery algorithms applied on raw data. In the pattern analysis phase interesting knowledge is extracted from frequent patterns and these results are used for website modification. In this paper we are using the FP-growth algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user's interest.

Keywords*— Web usage Mining, FP-Growth algorithm, Apriori algorithm, Web Server log data, Pattern Discovery.

I. INTRODUCTION

With the growth of explosive Internet information, the data mined from the web are useful information, which has gradually become a very important research .Generally, according to the different subjects, Web mining can be divided into three categories .Web content mining, Web structure mining and Web usage mining. Web log mining is a way of Web usage mining. Web access patterns mined from Web logs are interesting and useful knowledge in practice. Examples of applications of such knowledge include improving designs of web sites, analyzing system performance to understand user's reaction and motivation, build adaptive web sites [1].

Every day, new information, products and services are being offered by providers on the World Wide Web. At the same time, the number of consumers and the diversity of their interests increase. As a result, providers are seeking ways to infer the customers' interests and to adapt their web sites to make the content of interest more easily accessible. Frequent Pattern mining is a promising approach in support of this goal. Assuming that past navigation behaviour is an indicator of the users' interests, and then the records of this behaviour, kept in the form of the web-server logs, can be mined to infer what the users are interested in. On that basis, recommendations can be dynamically generated, to help new web-site visitors find the information of interest faster.

A user session is all of the page references made by a user during a single visit to a site. A transaction differs from a user session in that the size of a transaction can range from a single page reference to the entire page references in a user session, depending on the criteria used to identify transactions. Once user transactions or sessions have been identified, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst, such as path analysis, discovery of association rules and sequential patterns and clustering, and classification.

II. RELATED WORK

Huiping Peng in 2010 stated the use of FP-growth algorithm for processing the web log records, obtaining a set of frequent access patterns, then using the combination of browse interestingness and site topology interestingness of association rules for web mining [1]. Analysis of web usage mining by using Web Log analyzer tool, "Web Log Expert" was carried out by Sanjay Kumar Malik et al. in 2010. They focused on the development of Ontology for an intelligent or efficient web and it's relation with web usage mining. Finally, they also summarize some other research challenges towards an intelligent machine and web environment [2]. Hao Yan et al. in 2010 proposed a two-step K-means clustering algorithm to search user groups in realistic data collected from WAN. They gave some useful practical conclusions to facilitate design of targeting and recommending applications [3].Jiawei Han et al. in 2004 proposed a novel frequent-pattern tree structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth [4]. Rakesh Agrawal and Ramakrishan Srikant in 1994 consider the problem of discovering association rules between items in a large database of sales transactions. They present two new algorithms for solving this problem that are fundamentally different from the known algorithms. They also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid [5].Mohd helmy Abd Wahab et al. in 2008 describes the pre-processing techniques on IIS Web Server Logs ranging from the raw log file until before mining process can be performed [6]. C.P Sumathi et al. in 2011 presented an overview of the

various steps involved in the preprocessing stage [7].Renata Ivancsy et al. in 2006 investigated three pattern mining approaches from the web usage mining point of view [8].Vaibhav Kant singh et al. in 2008 shows how the different approaches achieve the objective of frequent mining . They also look for hardware approach of cache coherence to improve efficiency of the above process [9]. Dr. R.Krishnamoorthi and K.R Suneetha in 2009 has done the in-depth analysis of Web Log Data of NASA website to find information about a web site, top errors, potential visitors of the site etc. which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining[10].

III. WEB USAGE MINING

Web Usage Mining can be used to make search relevant by determining frequent access behavior for users, needed links can be identified to improve the overall performance of future accesses. Web Usage mining has been defined as the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of Web based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Web Usage Mining may be applied to data such as contained in logs files. A log file contains information related to the user queries on a website. Web usage mining may be used to improve the website structure or giving recommendations to visitors [2].

The aim in web usage mining is to discover and retrieve useful and interesting patterns from a large dataset. In web mining, this dataset is the huge web data. Web data contains different kinds of information, including, web structure data, web log data, and user profiles data. Web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure or usage data is used in the mining process. Web usage mining has various application areas such as web pre-fetching, link prediction, site reorganization and web personalization. Most important phases of web usage mining is discovering useful patterns from web log data by using pattern discovery techniques such as Apriori, FP-Growth algorithm[4].

IV. WEB USAGE MINING PROCESS

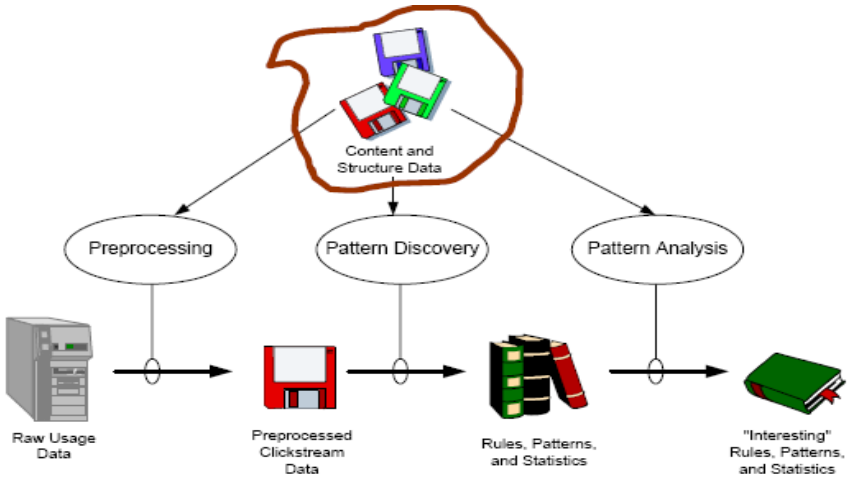

Fig.1 Web Usage Mining Process

A. Web Server Log Data

The web plays an important role and medium for extracting useful information. There is a need for data log to track any transaction of the communications.Log file data can offer valuable information insight into web site usage. It characterizes the activity of many users over a potentially long period of time.Server logs can be used to provide some technical information regarding server load, successful requests, as well as assisting in marketing and site development and management activities. Below is an example of common transfer log collected. This data is collected from NASA web server log.

EXAMPLE: 199.72.81.55 - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245
unicomp6.unicomp.net - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
199.120.110.21 - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 200 4085
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 304 0
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/video/livevideo.gif HTTP/1.0"2000
205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/countdown.html HTTP/1.0" 200 3985

The server log consists of several attributes. The attributes are as follows:-

1) *Date:* The date from Greenwich Mean Time (GMT x 100) is recorded for each hit. The date format is YYYY-MM-DD. The example above shows that the transaction was recorded at 1995-07-01.

2) *Time:* Time of transactions. The time format is HH:MM: SS. The example from above shows that the transaction time was recorded at 00:00:01.

3) Client IP Address: Client IP is the number of computer who access or request the site.

4) *User Authentication:* Some web sites are set up with a security feature that requires a user to enter username and password. Once a user logs on to a Website, that user's "username" is logged in the log file.

5) *Server IP Address:* Server IP is a static IP provided by Internet Service Provider. This IP will be a reference for access the information from the server.

6) *Server Port:* Server Port is a port used for data transmission. Usually, the port used is port 80.

7) *Server Method (HTTP Request):* The word request refers to an image, movie, sound, pdf, .txt, HTML file and more. The above example indicates that folder.gif was the item accessed.

8) *URI:* URI is path from the host. It represents the structure of the websites. For examples:/tutor/images/icons/fold.gif.

9) *Agent Log:* The Agent Log provides data on a user's browser, browser version, and operating system. This is the significant information, as the type of browser and operating system determines what a user is able to access on a site.

B. Pattern Discovery and Pattern analysis

The three main stages of web usage mining are data preprocessing, pattern discovery and pattern analysis. Data preprocessing involves removal of unnecessary data. Pattern discovery data mining techniques are used in order to extract patterns of usage from Web data. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the web access log. Pattern Analysis is the final stage of the Web usage mining. The aim of this process is to extract the interesting rules or patterns from the output of the pattern discovery process by eliminating the irrelative rules or patterns [10].

V. PROPOSED ALGORITHM

We are proposing FP-Growth algorithm for web usage mining, since no real time server available so we tested our algorithm on available log files on HTTP requests to the NASA Kennedy Space Center WWW server in Florida. The log was collected from 00:00:00 July 1, 1995 through 23:59:59 July 31, 1995, a total of 31 days. Now to extract the information such as requested files and most frequently accessed files, first we need to analyze the log file, below some entries of log file names are show:

TABLE I
INDEXING DONE TO REPRESENT STRINGS INTO SPECIFIC NUMBERS

INDEX ID	FILE NAMES
1.	shuttle/countdown/index.html
2.	KSC.html
3.	shuttle/missions/missions.html
4.	shuttle/missions/sts-71/images/images.html
5.	shuttle/missions/sts-71/movies/movies.html
6.	shuttle/countdown/liftoff.html
7.	shuttle/missions/sts-71/mission-sts-71.html
8.	shuttle/missions/sts-70/mission-sts-70.html
9.	shuttle/countdown/countdown.html
10.	history/Apollo/Apollo.html

TABLE I shows indexing operation applied for the purpose of data preprocessing. Here the strings are represented by unique index id. it shows the indexing for the requested files. We have done the same indexing operation for the gif files available in the web server log data and then applying the FP-growth algorithm to obtain various results such as most frequently visited pages, Top downloaded Pages from the web site, Top downloaded gif files and most frequently downloaded gif files from the web server log data.

A. Data Preprocessing

This operation is defined as filtering or pre-processing of data. Since the mathematical operations cannot be performed on strings therefore the strings are represented by specific numbers which is called Indexing. Hence we consider each file name by a unique index id. Next we apply the frequent pattern FP-Growth algorithm on the log files.

B. The analysis of frequent patterns from the web log data

After data preprocessing, we apply the following conditions. The following is a formal statement of the problem: Let L= {i1, i2…im} be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that T is a subset of L. Associated with each transaction, we say that a transaction T contains X, a set of some items in L. For example, if various users repeatedly access the same series of pages, a corresponding series of log entries will appear in the log file, and this series can be considered as an access pattern.

We have studied the performance of the FP-growth method in comparison with the basic apriori algorithms in large databases. Then we have find out frequent patterns from web log data by using the FP-growth method, our performance study shows that the method mines both short and long patterns efficiently in large databases, The FP-growth algorithm is one of the fastest approaches for frequent item set mining. The FP-growth algorithm uses the FP-tree data structure to achieve a condensed representation of the database transaction and employees a divide-and conquer approach to decompose the mining problem into a set of smaller problems.

VI. COMPARISON BETWEEN FP-GROWTH ALGORITHM AND APRIORI ALGORITHM

Apriori algorithm searches for large itemsets during its initial database pass and use its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent itemsets and those below are called small itemsets. The algorithm is based on the large itemset property which states: Any subset of a large itemset is large and any subset of frequent item set must be frequent.

The FP-growth method is efficient and scalable for mining both long and short frequent patterns and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent-pattern mining methods.

The Apriori heuristic achieves good performance gained by (possibly significantly) reducing the size of candidate sets. However, in situations with a large number of frequent patterns, long patterns, an Apriori algorithm suffer [4].

A. Execution Time Comparisons

The execution time comparison experiment is performed on datasets with 80K, 50K and 30K records; In these graphs the response times of both the algorithms increases as the support threshold is reduced.

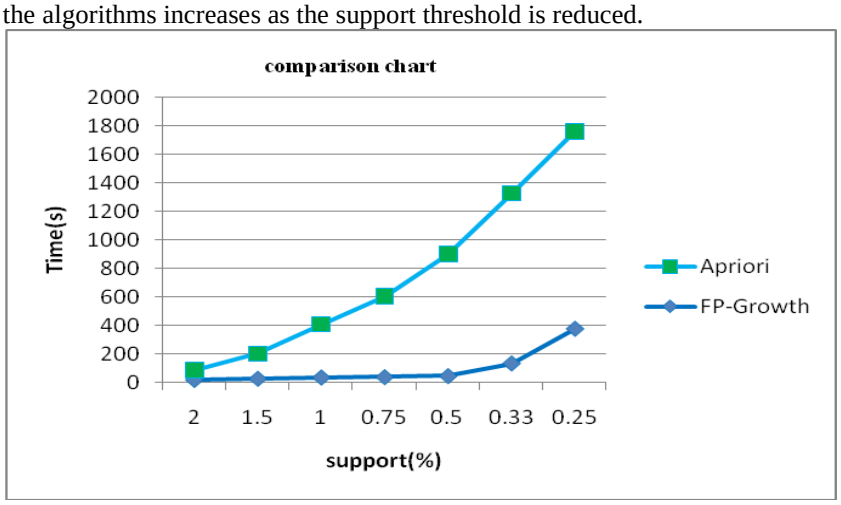

Fig.2 The above graph shows the comparison done using 80K Database

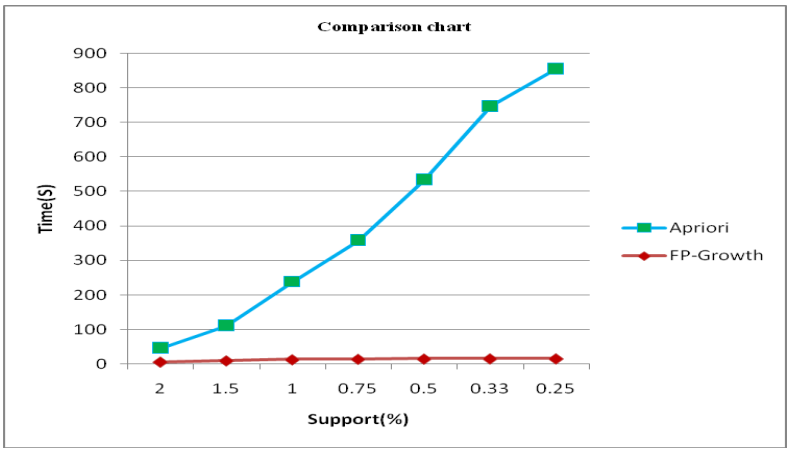

Fig.3 The above graph shows the comparison done using 50K Database

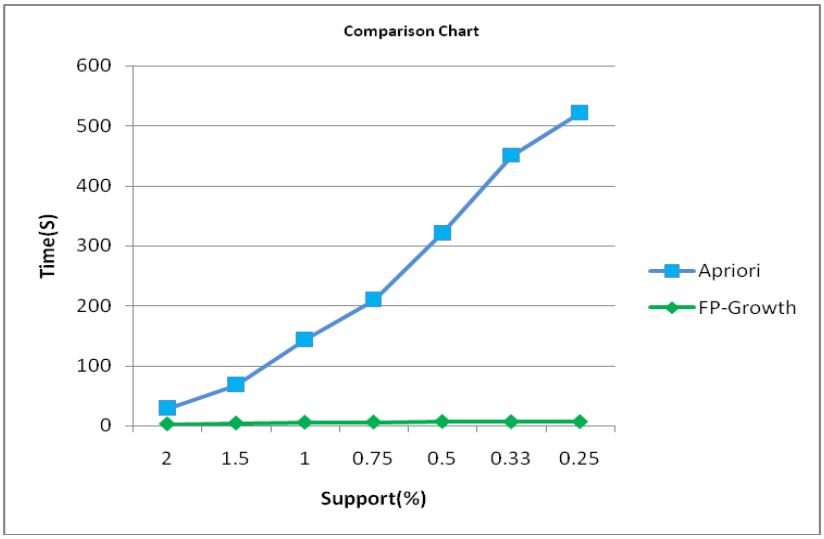

Fig.4 The above graph shows the comparison done using 30K Database

The above result shows that the FP-Growth algorithm is much more efficient than the basic apriori algorithm. Since the FP-Growth algorithm is very efficient therefore we have implemented the FP-growth algorithm for the purpose of Web usage mining.

VII. EXPERIMENTAL RESULTS

The simulation Results for the proposed algorithm is shown below, the training is performed by using first 1000 entries from the log files of NASA Kennedy Space Center WWW server and following results are drawn.

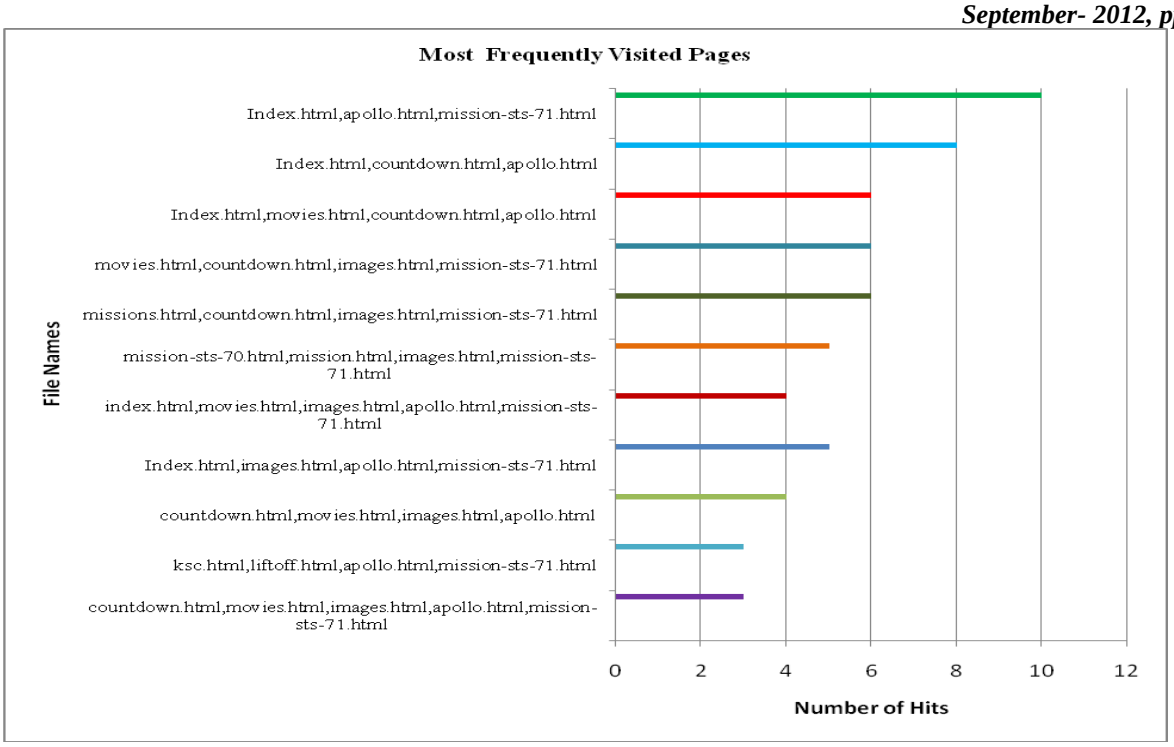

Fig.5 The above graph shows the result for the Most frequently visited pages from the website.

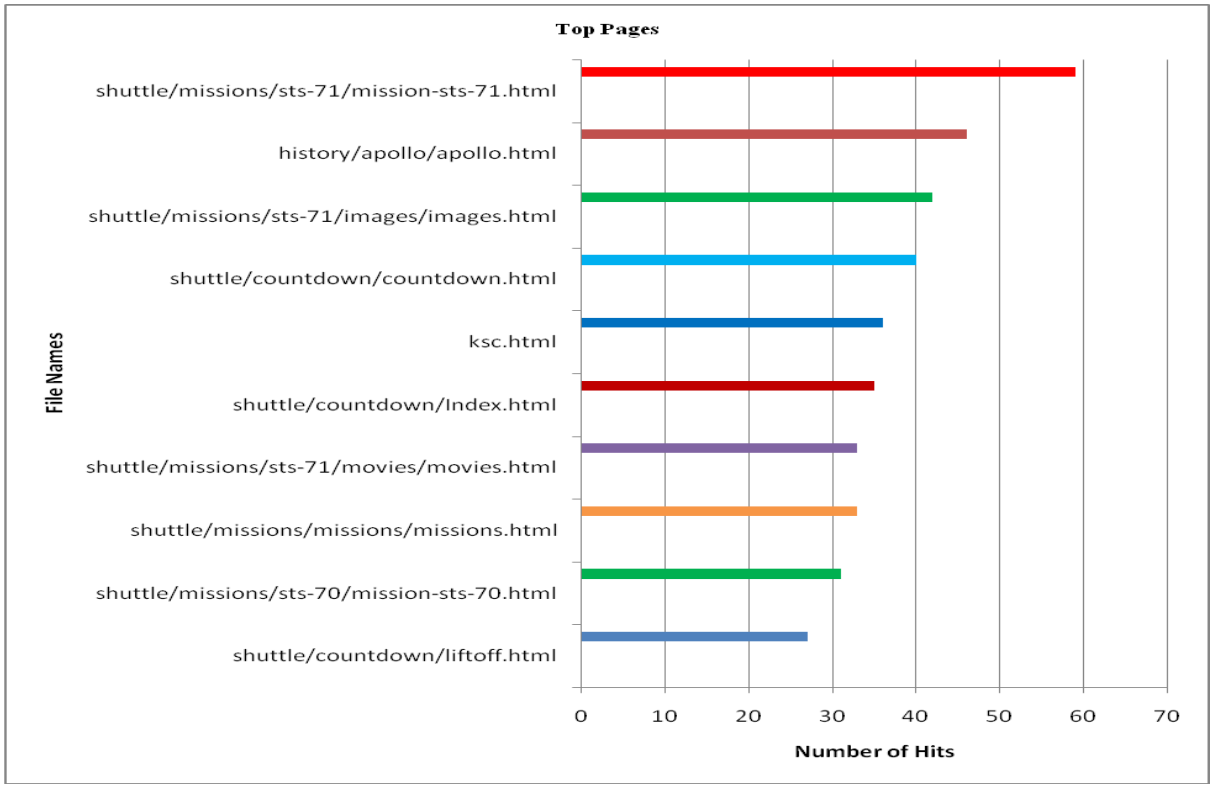

Fig.6 The above graph shows the result for the Top downloaded pages from the website.

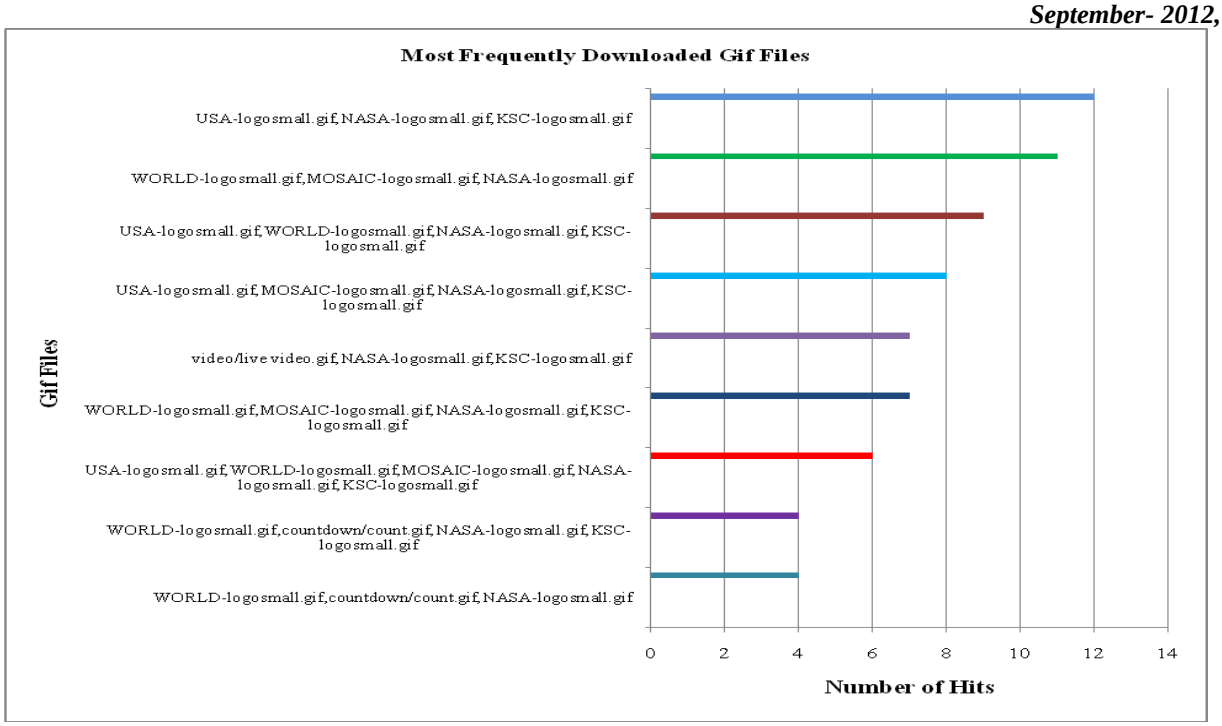

Fig.7 The above graph shows the result for the most frequently downloaded gif files from the website.

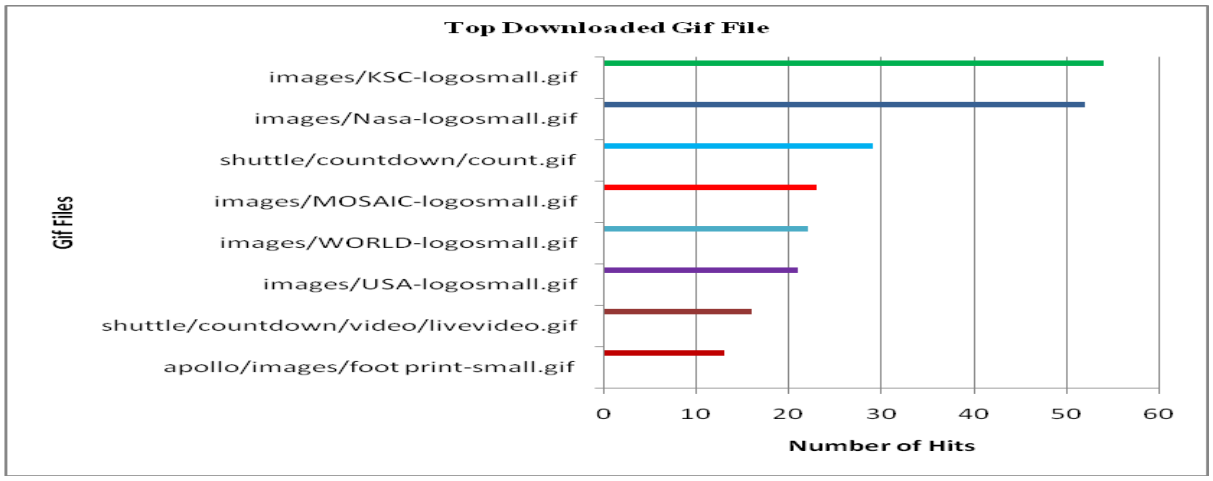

Fig.8 The above graph shows the result for Top Downloaded gif file from the website.

VIII. CONCLUSION AND SCOPE

The simulation result shows that the FP-Growth algorithm is used for finding the most frequently access pattern generated from the web log data, By using the concept of web usage mining we can easily find out the user's interest and we can modify and make our web site more valuable and more easily accessible for the users.The main goal of the proposed system is to identify usage pattern from web log files. FP Growth Algorithm is used for this purpose. Apriori is a classic algorithm for association rule mining. The main drawback of Apriori algorithm is that the candidate set generation is costly, especially if a large number of patterns and/or long patterns exist. The FP-growth algorithm is one of the fastest approaches for frequent item set mining. The FP-growth algorithm uses the FP-tree data structure to achieve a condensed representation of the database transaction and employees a divide-and conquer approach to decompose the mining problem. Our experimental result shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns. In future the algorithm can be extended to web content mining, web structure mining.

REFERENCES

[1] Huiping Peng"Discovery of Interesting Association Rules Based on Web Usage Mining" 2010 International Conference.
[2] Sanjay Kumar Malik, Nupur Prakash, S.A.M. Rizvi" Ontology and Web Usage Mining towards an Intelligent Web focusing web logs" 2010 International Conference .
[3] Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei "Web usage mining based on WAN users' behaviours" 2010 International Conference.
[4] Han J., Pei J., Yin Y. and Mao R., "Mining frequent patterns without candidate generation: A frequent-pattern tree approach" Data Mining and Knowledge Discovery, 2004.
[5] Agrawal R, Srikant R., "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99, ISBN 1-55860-153-8.
[6] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd ,Mohamad Mohsin"Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm" 2008.
[7] C.P. Sumathi, r. padmaja valli**,**"An overview of preprocessing of web log files for web usage mining" 2011.
[8] Renata Ivancsy, Istvan Vajk "Frequent Pattern Mining in Web Log Data"2006.
[9] Vaibhav Kant Singh, Vijay Shah, Yogendra Kumar Jain, "Proposing an Efficient Method for Frequent Pattern Mining" 2008
[10] K. R. Suneetha, Dr. R. Krishnamoorthi, "Identifying User Behaviour by Analyzing Web Server Access Log File"2009

