# A Survey Report on Extracting Frequent Patterns using FP-Growth Algorithm and Apriori Algorithm

*Kannasani Srinivasa Rao*[1], *M Krishnamurthy*[2], *A Kannan*[3]

[1]Hindustan University, Chennai, India

[2]KCG College of Technology, Chennai, India

[3]Anna University, Chennai, India

**E-mail:** srinu532@gmail.com

***Abstract***

*The proposed research work focuses on Web Usage Mining. In this research work extraction of users interests from web log data can be done, which are based on visit time and visit density which can be obtain from an analysis of web users web log data. Also, this research work will focus on updating user's interests. Moreover, we will also discuss the technological challenges to provide personalization and examine new developments in data mining and its applications to personalization, including techniques to support clustering and searching in a very high dimensional data space with huge amount of data.*

***Keywords:*** *Web usage mining, user's interests, data*

## INTRODUCTION

It is not exaggerated to say the World Wide Web is the most excited impacts to the human society in the last 10 years. It changes the ways of doing business, providing and receiving education, managing the organization etc. Today, Web has turned to be the largest information source available. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository. Today, many recommendation systems cannot give users enough personalized help but provide the user with lots of irrelevant information. One of the main reasons is that it cannot accurately extract user's interests [1–5]. Therefore, analyzing users web log data and extracting users potential interested domains become very important and challenging research topics of Web Usage Mining. Web technology is not evolving in comfy and progressive steps; however, it is turbulent, erratic and sometimes rather uncomfortable. It is calculable that the web, arguably the foremost necessary a part of the new technological setting, has expanded by regarding 2000 attempt to that is doubling in size each six to 10 months. In recent years, the advance in pc and net technologies and, therefore, the decrease in their price have expanded the suggests that on the market to gather and store knowledge. As AN intermediate consequence, the quantity of data (Meaningful data) hold on has been increasing at a really quick pace. Ancient info analysis techniques are helpful to form informative reports from knowledge and to substantiate predefined hypothesis regarding the info. However, vast volumes being collected produce new challenges for such techniques as organizations hunt for ways in which to form use of the hold on information to realize a position over competitors. It is affordable to believe that knowledge collected over AN extended amount contains hidden data regarding the business or patterns characterizing client profile and behavior. With the zoom of the globe Wide net, the study of data discovery in net, modeling and predicting the user's access on an internet website has become important [6–10].

Jespersen et al. planned a hybrid approach for analyzing the visitant click stream sequences. A mix o fmachine-readable text probabilistic descriptive linguistics and click on reality table approach is employed to mine internet logs that might be conjointly used for general sequence mining tasks. Mobasher et al. planned the net personalization system that consists of offline tasks associated with the mining if usage knowledge and on-line method of automatic online page customization supported the data discovered. LOGSOM (LOGSOM, a system that utilizes Kohonen's self-organizing map (SOM) to prepare web content into a two-dimensional map) planned by Smith et al., utilizes a self-organizing map based mostly entirely on the users' navigation behavior, instead of the content of the net pages. Lumber Jack planned by Chi et al. builds up user profiles by combining each agglomeration of user sessions and ancient applied math traffic analysis exploitation k–means formula. Joshi et al. used relative on-line analytical process approach for making an internet log warehouse exploitation access logs and mined logs. Web mining may be divided into 3 areas, particularly website mining, internet structure mining and internet usage mining. Website mining focuses on discovery of data hold on on the net. Internet. Structure mining focuses on improvement in structural style of an internet site. internet usage mining, the most topic of this paper, focuses on data discovery from the usage of people websites [11–15].

Global Internet Usage Average Usage shows the current usage around the globe and in United States. Month of September 2003, Panel Type: Home

| | **September** | **August** | **% Change** |
|---|---|---|---|
| Number of Sessions per Month | 22 | 22 | 1.65 |
| Number of Unique Domains Visited | 55 | 54 | 0.89 |
| Page Views per Month | 901 | 899 | 0.3 |
| Page Views per Surfing Session | 41 | 41 | 0 |
| Time Spent per Month | 11:59:20 | 11:50:30 | 1.24 |
| Time Spent During Surfing Session | 0:32:29 | 0:32:37 | -0.4 |
| Duration of a Page Viewed | 0:00:48 | 0:00:47 | 0.94 |
| Active Internet Universe | 252,672,070 | 253,054,814 | -0.15 |
| Current Internet Universe Estimate | 419,054,724 | 416,339,888 | 0.65 |

United States: Average Web Usage

Month of October 2003, Panel Type: Home

| | |
|---|---|
| Sessions/Visits Per Person | 71 |
| Domains Visited Per Person | 103 |
| PC Time Per Person | 80:46:37 |
| Duration of a Web Page Viewed | 0:01:00 |
| Active Digital Media Universe | 47,003,165 |
| Current Digital Media Universe Estimate | 51,012,930 |

## WEB USAGE MINING AND PATTERN DISCOVERY

Web usage mining is that the application knowledge of information mining techniques to find usage patterns from internet data, so as to know and higher serve the requirements of Web-based applications. Internet usage mining consists of 3 phases, specifically preprocessing, pattern discovery, and pattern analysis. A high level internet usage mining method is given in Figure one. Mobasher *et al.* proposes that the online mining method is often divided into 2 main elements. The primary half includes the domain dependent processes of remodeling the online knowledge into appropriate dealing kind. This includes preprocessing, dealing identification, and knowledge integration parts. The second half includes some data processing and pattern matching techniques like association rule and ordered patterns. Within the absence of cookies or dynamically embedded session Ids within the URIs, the mix of science address are often used as a primary pass estimate of distinctive users. This estimate is often refined victimization the referrer field as delineated. Some authors have planned world architectures to handle the online usage mining method. Cooley *et al.* planned a web site data filter, named WebSIFT that establishes a framework for internet usage mining as shown in Figure two. The WebSIFT performs the mining in distinct tasks.

In this case, episodes are either all of the page views in a server sessions that the user spent a significant amount of time viewing, or all of the navigation page views leading up to each content page view. The details of how a cutoff time is determined for classifying a page view as content or navigation are also contained. The click-stream or click-flow for each user is divided into sessions based on a simple thirty-minute timeout. The notion of what makes discovered knowledge interesting has been addressed. A survey of methods that have been used to characterize the interestingness of discovered patterns. Four dimensions employed by to classify interest measures square measure pattern-form, illustration, scope, and class.  Pattern-form defines what style of patterns a live is applicable to, like association rules or classification rules. The illustration dimension defines the character of the framework, like probabilistic or logical. Scope could be a binary dimension that indicates whether or not the live applies to single pattern, or to the whole discovered set. The ultimate dimension, category is additionally a binary dimension that may be labeled as subjective or objective [16–20].

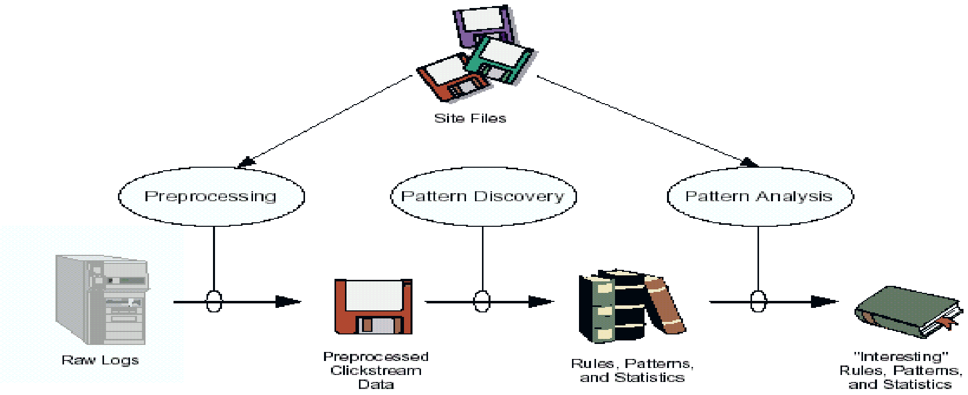
Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure 1: High Level *Web Usage Mining* Process

We Sift system divides the net Usage Mining method into 3 main elements, as show in Figure 1. For a specific computing device, the 3 server logs access, referrer, and agent (often combined into one log), the hypertext markup language files, example files, script files or databases that compose the location content, and any facultative knowledge like registration knowledge or remote agent logs give data to construct the various information abstractions. The preprocessing part uses the computer file to construct a server session file supported the tactic and heuristics mentioned. So, as to preprocess a server log, the log should initial be "cleaned", that consists of removing unsuccessful requests, parsing relevant CGI name/value pairs and rolling up file accesses into page views. Once the log is reborn into a listing of page views, users should be known. Within the absence of cookies or dynamically embedded session Ids within the URIs, the mix of IP address.

The first is preprocessing state in which user sessions are inferred from log data. The second searches for patterns in the data by making use of standard data mining techniques, such as association rules or mining for sequential patterns. In the third stage an information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns. Links between pages and the similarity between contents of pages provide evidence that pages are related. The preprocessing phase allows the option of converting the server sessions into episodes prior to performing knowledge discover.

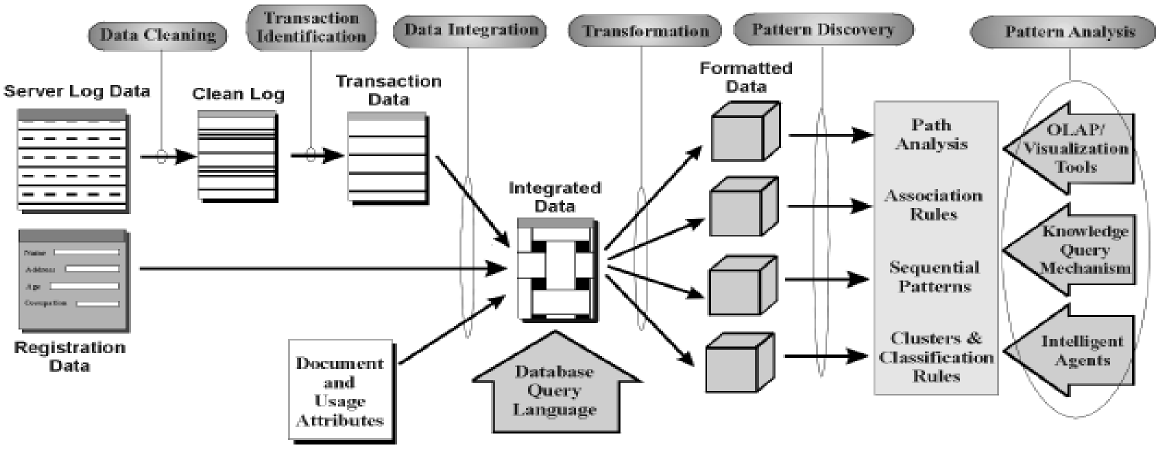

***Fig. 2:*** *General Architecture for Web Usage Mining.*

Preprocessing for the content and structure of a site involves assembling each page view for parsing and /or analysis. Page views are accessed through HTTP requests by a "site crawler" to assemble the components of the page view. This handles both static and dynamic content. In addition to being used to derive a site topology, the site files are used to classify the pages of a site. Both the site topology and page classification and then be fed into the information filter. The knowledge discovery phase uses existing data mining techniques to generate rules and patterns. Included in this phase is the generation of general usage statistics, such as number of "hits" per page, page most frequently accessed, most common starting page, and average time spent on each page [21–28].

The WebSIFT performs the mining in distinct tasks. The primary state is preprocessing during which user sessions area unit inferred from log information. The second searches for patterns within the information by creating use of normal data processing techniques, like association rules or mining for ordered patterns. Within the third stage AN info filter bases on domain data and, therefore, the internet site structures is applied to the mining patterns in look for the fascinating patterns. Links between pages and, therefore, the similarity between contents of pages offer proof that the pages area unit connected. This info is employed to spot fascinating patterns, for instance, item sets that contain pages in some way connected area unit declared fascinating. In Mobasher *et al.* the authors propose to cluster the item sets obtained by the mining stage in cluster of URL references. These clusters area unit aimed toward real time online page personalization. A hypergraph is inferred from the mined item sets wherever the nodes correspond to pages and, therefore, the hyperedges connect pages in an exceedingly itemset. The load of a hyperedge is given by the boldness of the principles concerned. The graph is later divided into clusters and an occurring user session is matched against such clusters. For every URL within the matching clusters a recommendation score is computed and, therefore, the recommendation set consists by the entire URL whose recommendation score is higher than a threshold.

In Masseglia *et al.* projected Associate in Nursing integrated tool for mining access patterns and association rules from log file. The techniques enforced pay explicit attention to the handling of your time constraints, like the minimum and most time gap between adjacent requests in an exceedingly pattern. The system provides a true time generator of dynamic links that geared toward mechanically modifying the machine-readable text organization once user navigation matches an antecedently strip-mined rule. Fundamental strategies of information cleanup and preparation are well studied by Srinivasa *et al.* The most techniques historically used for modeling usage patterns in an exceedingly computing device square measure cooperative filtering (CF), bunch pages or user sessions, association rule generation, successive pattern generation and Markoff Models. The prediction step is that the data processing of the model that considers the active user session and makes recommendations supported the discovered patterns. The time spent on a page may be a smart live of the user's interest therein page, providing Associate in Nursing implicit rating for it. If a user is interested in the content of a page, she will likely spend more time there compared to the other pages in her session. They presented a new model that uses both the sequences of visiting pages and the time spent on those pages which reflects the structural information of user session and handles two-dimensional information. Data preprocessing consists of knowledge filtering, user identification, session/transaction identification, and topology extraction. Information filtering filters out some noise, i.e., unsuccessful requests, mechanically downloaded graphics, or requests from robots, to induce additional compact coaching information. Currently, individuals use some heuristic rules to spot user, like science address, cookies, etc. Preprocessing consists of changing the usage, content, and structure data contained within the numerous on the market information sources into the information abstractions necessary for pattern discovery.

### Usage Preprocessing

Usage preprocessing consists of sites, like science addresses, page references and, therefore, the date and time of accesses. Typically, the usage information comes from Associate in Nursing Extended Common Log Format (ECLF) Server log.

### Content Preprocessing

Content preprocessing consists of changing the text, images, scripts, and multimedia system information into forms that square measure helpful for the net usage mining method. Usually, this consists of playing content mining like classification or cluster. Within the context of net usage mining, the content of websites will be accustomed filter the input to the pattern discovery algorithms.

### Structure Preprocessing

Net structure mining analyses the link structure of the net so as to spot relevant documents. The structure of a website is made by the machine-readable text links between page views. Intra-page structure data includes the arrangement of varied hypertext markup language or XML tags inside a given page. The principal quite inter-page structure data is hyper-links connecting one page to a different. The Google computer programme makes use of the net link structure within the method of determinative the connection of a page. The Google computer programme achieves sensible results as a result of whereas the keyword similarity analysis ensures high preciseness the utilization of likelihood live ensures prime quality of the pages came back.

### Server-Level Collection

An online server log records the browsing behavior of website guests. The information recorded in server logs mirror the co-occurring and interleaved access of an online website by multiple users. These log files will be keep in numerous formats like Common Log Format (CLF) or Extended Common Log Format (ECLF). ECLF contains shopper science address, User ID, time/date, request, status, bytes, referrer, and agent. Following of individual users is not a simple task as a result of the homeless affiliation model of the hypertext transfer protocol protocol. So, as to handle this downside, net servers also can store different quite usage data like cookies in separate logs, or appended to the CLF or ECLF logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Packet sniffing technology (also referred to as "network monitors") is an alternative method for collecting usage data through server logs. Packet sniffers monitor network traffic coming to a Web server and extract usage data directly from TCP/IP packets. Besides usage data, the server side log also provides access to the "site files", e.g., content data, structure information, local databases, and Web page meta-information such as the size of a file and its last modified time.

### Client Level Assortment

Shopper-side assortment will be enforced by employing a remote agent (such as Java scripts or Java applets) or by modifying the ASCII text file of associate existing browser (such as Mosaic or Mozilla) to boost its information collection capabilities. Proxy Level Collection: the net Service supplier (ISP) machine that users hook up with through a model could be a common variety of proxy server. An internet proxy acts as associate negotiant between shopper browsers and internet servers. Proxy-level caching will be wont to scale back the loading of your time of an internet page intimate with by users in addition because the network traffic load at the server and shopper sides.

### Pattern Discovery

Pattern discovery uses ways and algorithms developed from many fields like statistics, data processing, machine learning and pattern recognition. Zaiane *et al.* planned the utilization of On-Line Analytical process (OLAP) technology in internet usage mining. OLAP and also the information cube structure provide a extremely interactive and powerful information retrieval and analysis atmosphere. The data that may be discovered is pictured within the variety of rules, tables, charts, graphs, and different visual presentation forms for characterizing, comparing, predicting, or classifying information from the net access log. Mental image also can be utilized in internet usage mining, and it presents the info within the means that may be understood by users a lot of simply.

### Statistical Analysis

Applied mathematics techniques square measure the foremost common methodology to extract data regarding guests to an internet web site. By analyzing the session file, one will perform completely different types of descriptive applied mathematics analyses (frequency, mean, median, etc.) on variables like page views, viewing time and length of a direction path. As an example e-Trade developed in West Germanic language for European country and scrapped it as a result of German folks were visiting a people site instead of the German site. Several internet traffic analysis tools turn out a periodic report containing applied mathematics info like the foremost oftentimes accessed pages, average read time of a page or average length of a path through a web site. This kind of data will be probably helpful for up the system performance, enhancing the safety of the system, facilitating the location modification task, and providing support for selling selections. There square measure a lot of business tools offered for applied mathematics analysis.

## Association Rules

Association rule generation will be wont to relate pages that square measure most frequently documented along in a very single server sessions. Within the context of internet usage mining, association rules ask sets of pages that square measure accessed beside a support price exceptional some such threshold. Association rule mining has been well studied in data processing, particularly for basket dealings information analysis. Several association rule algorithms are used, like Apriori, Partition. Apart from being applicable for e-Commerce, business intelligence and selling applications, it will facilitate internet designers to structure their electronic computer. The results about the usefulness of such rules in supermarket transaction or in web application have not been reported. People also put some constraints over the mining process, and prune the extracted rules. The association rules may also serve as heuristic for pre fetching documents in order to reduce user-perceived latency when loading a page from a remote site. In electronic CRM, an existing customer can be retained by dynamically creating web offers based on associations with threshold support and/or confidence value.

## Clustering

Cluster may be a technique to cluster along a group of things having similar characteristics. Cluster will be performed on either the users or the page views. Cluster analysis in net usage mining intends to search out the cluster of user, page, or sessions from diary file, wherever every cluster represents a gaggle of objects with common fascinating or characteristic. User cluster is meant to search out user teams that have common interests supported their behaviors, and it is crucial for user community construction. Page cluster is that the method of cluster pages per the users' access over them. Such information is particularly helpful for inferring user demographics so as to perform market segmentation in e-Commerce applications or give personalised online page to the users. On the opposite hand, cluster of pages can discover teams of pages having connected content. This data is helpful for the web search engines and net help suppliers. In each applications, permanent or dynamic hypertext mark-up language pages will be created that counsel connected hyperlinks to the user per the user's question or past history of data wants. The intuition is that if the likelihood of visiting page, given page has additionally been visited, and is high, then perhaps they will be sorted into one cluster. For session cluster, all the sessions are processed to search out some fascinating session clusters. Every session cluster is also one fascinating topic at intervals the net web site. Mobasher *et al.* generated recommendations from URL clusters to make Associate in Nursing adaptative computing machine by victimization ARHP (Association Rule Hypergraph Partitioning). Abhrahum *et al.* projected Associate in Nursing ant-clustering formula to get net usage patterns and a linear genetic programming approach to research the visitant trends. They projected hybrid framework, that uses Associate in Nursing hymenopteran colony improvement formula to cluster net usage patterns. The data from the log files are cleansed and preprocessed and, therefore, the ACLUSTER formula is employed to spot the usage patterns. The developed clusters of information are fed to a linear genetic programming model to research the usage trends.

## Classification

Classification is the task of mapping a data item into one of several predefined classes. In the internet marketing, a customer can be classified as 'no customer', 'visitor once' and 'visitor regular' based on their browsing patters and discovered rules for attracting the customers by displaying special offers. In the web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Baysian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc. For example, classification on server logs may lead to the discovery of interesting rules such as: 30% of users who placed an online order in /Product/Music are in the 18-25 age groups and live on the west coast. The Classification algorithms such as C4.5, CART, BAYES, and RIPPER can be used to predict if page is of interest to the user.

### Sequential Patterns

The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. A new algorithm MiDAS (Mining Internet data for Associative Sequences) for discovering sequential patterns from web log files has been proposed that provides behavioral marketing intelligence for e-commerce scenarios. MiDAS contains three phases: 1. A priori phase is the input data preparation, which consists of data reduction and data type substitution. 2. Discovery Phase discovers the sequences of hits and generates the pattern tree. 3. A posteriori Phase filters out all sequences that do not fulfill the criteria laid in the specified navigation templates and topology network and also pruning is done in this phase. By using this approach, Web marketers can predict future visit patterns, which will be helpful in placing advertisements aimed at certain user groups. Other types of temporal analysis that can be performed on sequential patterns include trend analysis, change point detection, or similarity analysis. WUM employs an innovative technique for the discovery of navigation patterns over an aggregated materialized view of the web log. After performing the classical preparation steps (i.e., user and session identification) the user sessions are merged into Aggregated Tree. An Aggregated Tree is a tree constructed by merging trails with the same prefix. WUM provides a query language called MINT to let the users specify their query, concerning the content, structure and statistics of navigation patterns. MINT supports the specification of criteria of statistical, structural, and textual nature. The "WEBMIER" tool provides a query language on top of external mining software for association rules and for sequential patterns.

### Deviation/Outlier Detection

It contains techniques aimed at detecting unusual changes in the data relatively to the expected values. Such techniques are useful, for example, in fraud detection, where the inconsistent use of credit cards can identify situations where a card is stolen. The inconsistent use of credit card could be noted if there were transactions performed in different geographic locations within a given time window.

### Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process as described in Figure 2. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

## 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 1**: 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 1 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.

## 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.

## The Analysis of Frequent Patterns from the Web Log Data

After knowledge preprocessing, we tend to apply the subsequent conditions. The subsequent may be a formal statement of the problem: Let, L= be a collection of literals, known as things. Let, D be a collection of transactions, wherever, every dealings T may be a set of things specified T may be a set of L. Related to every dealings, we are saying that a dealings T contains X, a collection of some things in L. For instance, if varied users repeatedly access identical series of pages, a corresponding series of log entries can seem within the log file, Associate in Nursing this series are often thought of as an access pattern. We have studied the performance of the FP-growth technique compared with the essential apriori algorithms in massive databases. Then, we have ascertain frequent patterns from diary knowledge by victimisation the FP-growth technique, our performance study shows that the strategy mines each short and long patterns with efficiency in massive databases, The FP-growth algorithmic rule is one amongst the quickest approaches for frequent item set mining. The FP-growth algorithmic rule uses the FP-tree system to realize a condensed illustration of the information dealings and staff a divide-and conquer approach to decompose the mining drawback into a collection of smaller issues.

## COMPARISON BETWEEN FP-GROWTH ALGORITHM AND APRIORI ALGORITHM

Apriori formula searches massive for big item sets throughout its initial information pass and use its result because the seed for locating different large datasets throughout resulting passes. Rules having a price on top of the minimum ar known as massive or frequent itemsets and people below are known as little item sets. The formula relies on the big itemset property that states: Any set of an oversized itemset is large and any set of frequent item set should 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].

### 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.

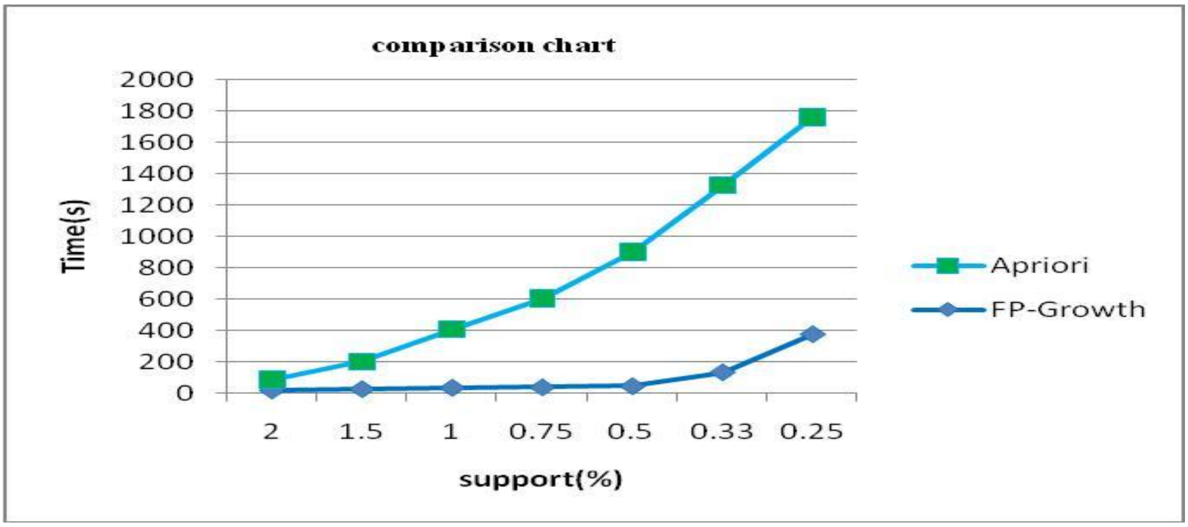

***Fig. 3:*** *The Above Graph shows the Comparison done using 80K Database.*

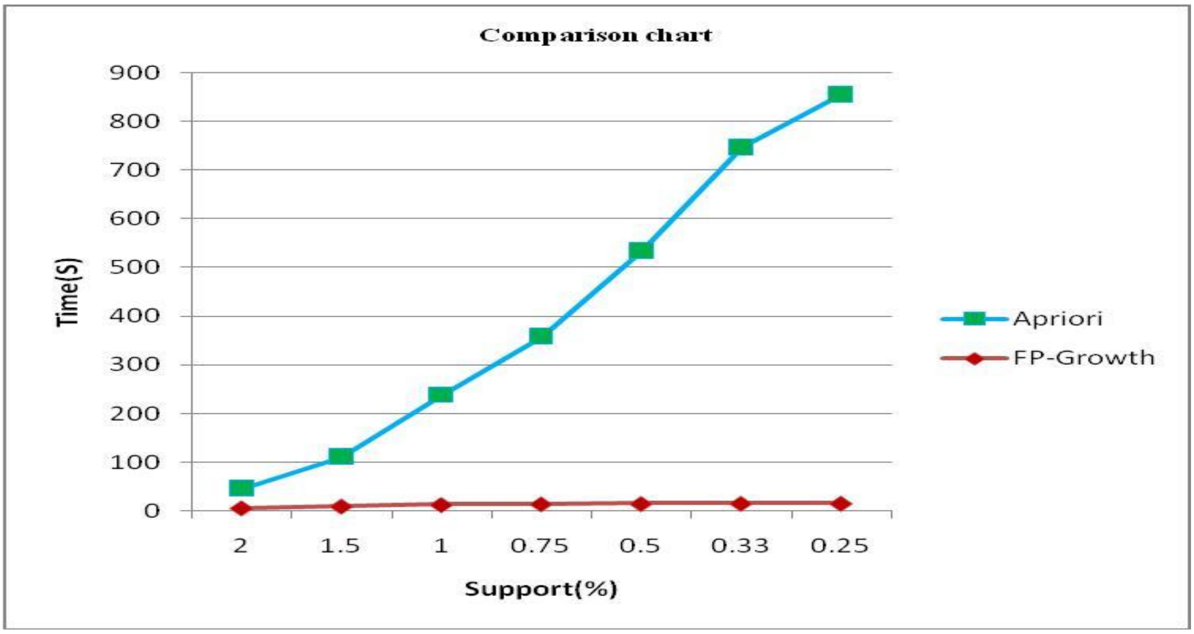

***Fig. 4:*** *The Above Graph shows the Comparison done using 50K Database.*

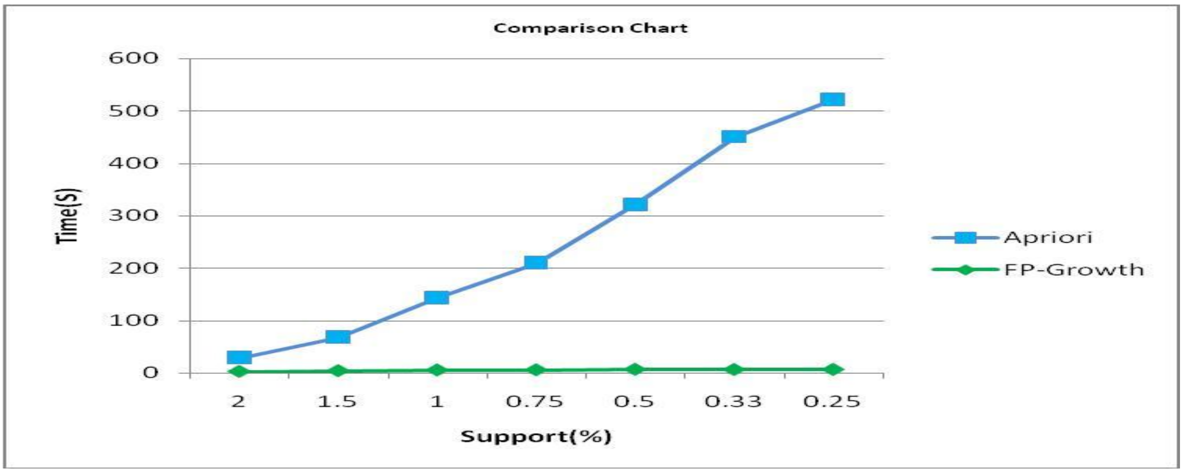

***Fig. 5:** The Above Graph shows the Comparison done using 30K Database.*

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.

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.

## EXPERIMENTAL RESULTS

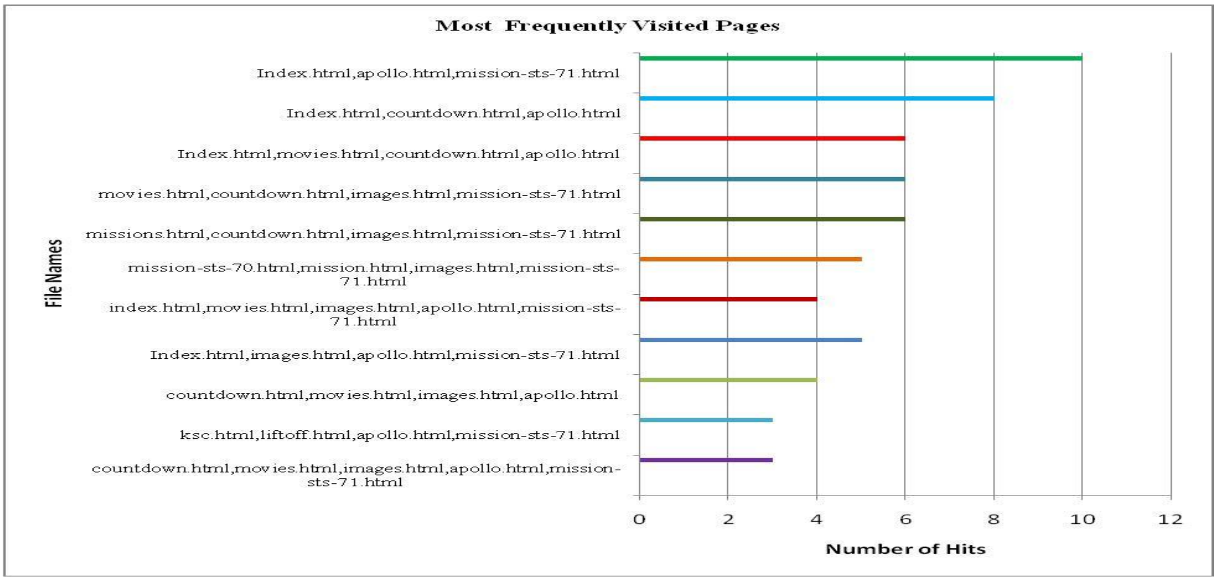

***Fig. 6:** The Above Graph shows the Result for the Most frequently Visited Pages from the Website.*

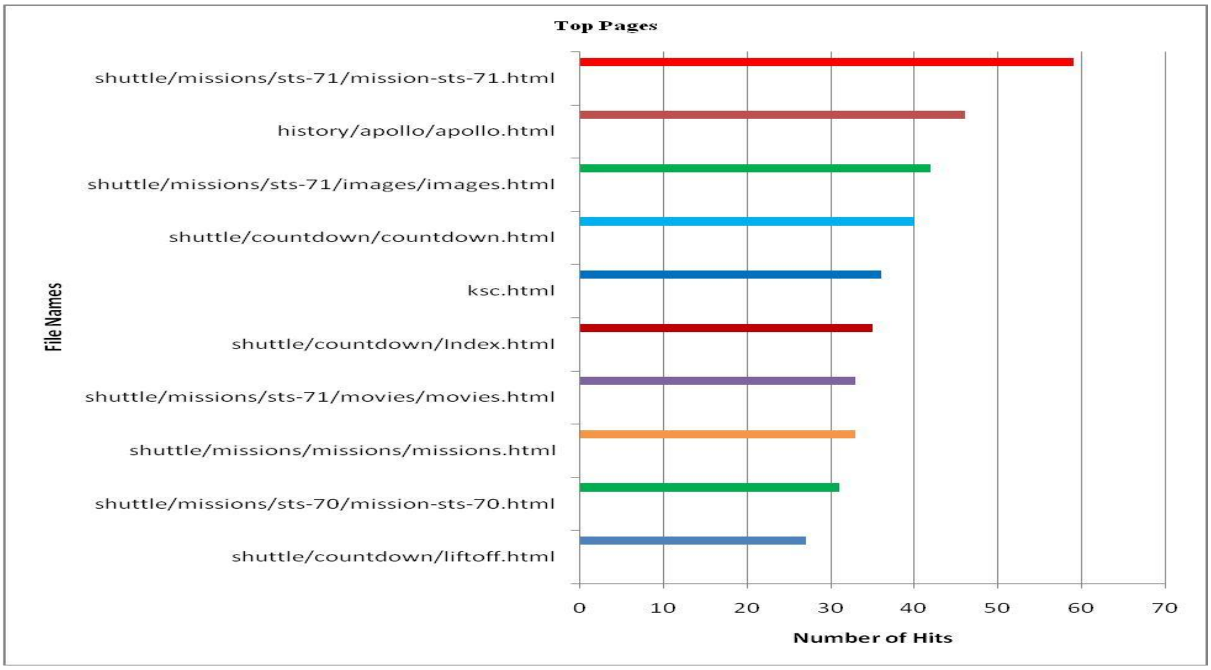

***Fig. 7:*** *The Above Graph shows the Result for the Top Downloaded Pages from the Website.*

.

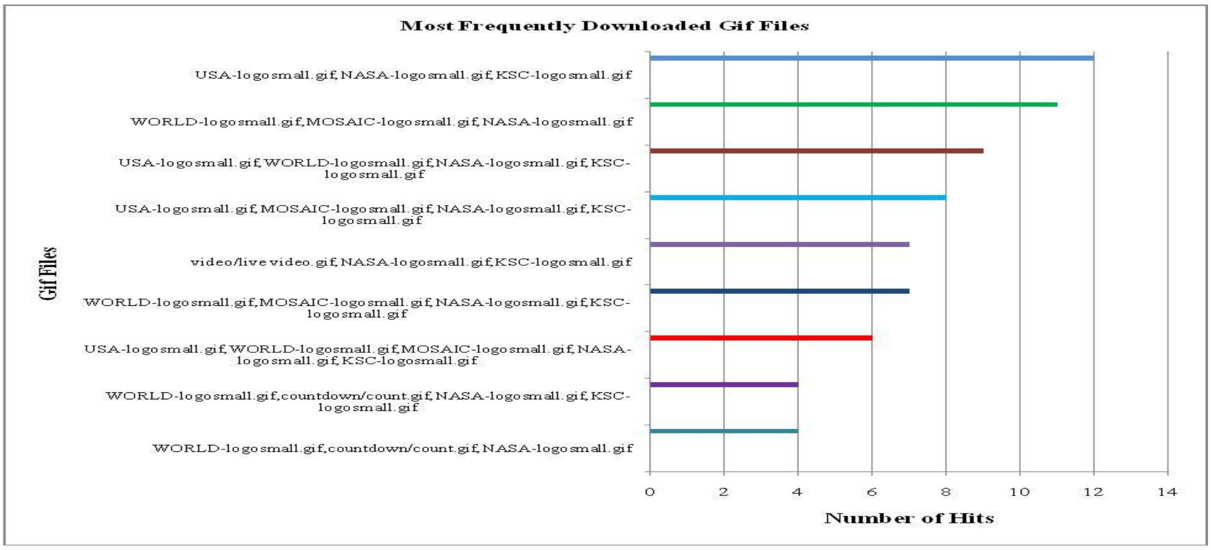

***Fig. 8:*** *The Above Graph shows the Result for the Most Frequently Downloaded Gif Files from the Website.*

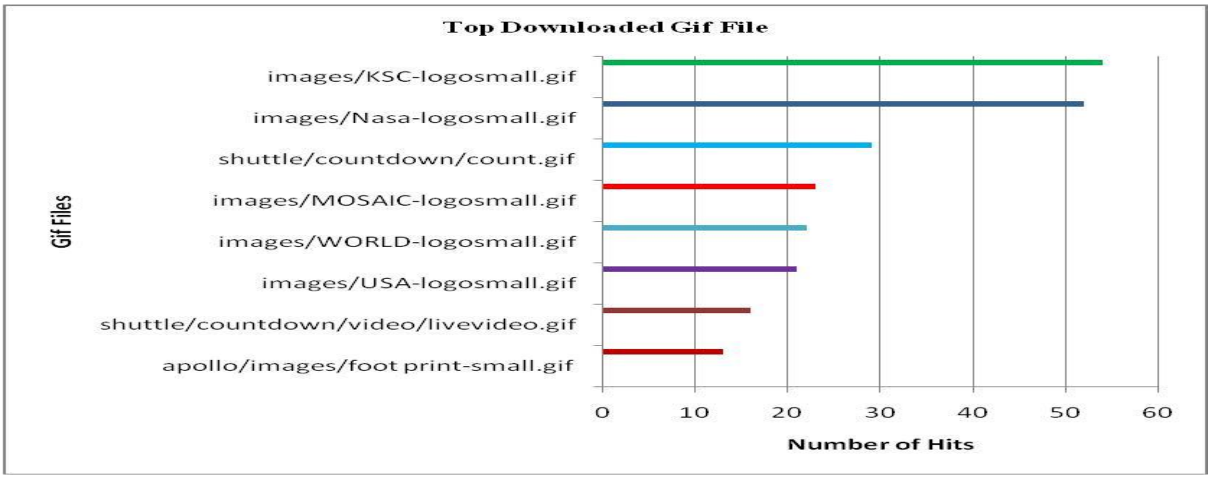

***Fig. 9:** The Above Graph shows the Result for Top Downloaded Gif File from the Website.*

## CONCLUSION

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. Ajit Abhraham, Vitorino Ramos. Web usage mining using artificial ant colony clustering and linear genetic programming, to appear in CEC´03-congress on evolutionary computation. *IEEE Press*; 2003.
2. A.G. Büchner, M. Baumgarten, S.S. Anand, et al. navigation pattern

discovery from internet data. *In WEBKDD*; 1999.

3. Jos'e Borges, Mark Levene. Data mining of user navigation patterns. *WEBKDD*; 1999.
4. A.G. Büchner, M.D. Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *ACM SIGMOD*. 1998; 27(4): 54–61p.
5. V. Cadez, D. Heckerman, C. Meek, et al. Model-based clustering and visualization of navigation patterns on a website. *Journal of Data Mining and Knowledge Discovery*. 2003; 7(4).
6. Robert Cooley, Bamshad Mobasher, Jaideep Srivastava. Web mining: information and pattern discovery on the world wide web (A Survey Paper) (1997). *In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97);* 1997.
7. Robert Cooley, Bamshad Mobasher, Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*. 1999; 1(1).
8. Chi E.H., Rosien A., Heer J. Lumber jack: Intelligent discovey and analysis of web user traffic composition. *In Proceedings of ACM-SIGKDD Workshop on Web Mining for Usage Patterns and User* Profiles; 2002.
9. Available at: http://www.crisp-dm.org.
10. Robert Cooley, Pang-Ning Tan, Jaideep Srivastava. WebSIFT: The web site information filter system. *Proceedings of the Web Usage Analysis and User Profiling Workshop*; 1999.
11. Sule Gunduz, M. Tamer Ozsu. A web page prediction model based on click-stream tree representation of user behavior. *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2003.
12. Robert J. Hilderman, Howard J. Hamilton. Knowledge discovery and interestingness measures. A survey, Technical Report, University of Regina; 1999.
13. Joshi K. P., Joshi A., Yesha Y., Krishnapuram, R. Warehousing and mining we logs, proceedings of the 2nd ACM CIKM workshop on web information and data management. 1999; 63–68p.
14. Jespersean S.E., Throhauge J., Bach T. A hybrid approach to web usage mining, data warehousing and knowledge discovery, (DaWaK’02),

LNCS 2454. *Springer Verlag Germany*. 2002; 73–82p.

15. Soren E. Jespersen, Jesper Thorhauge, Torben Bach Pederson. A hybrid approach to web usage mining, Technical Report 02-5002; 2002.
16. Margaret H. Dunham. Data mining introductory and advanced topics. Prentice Hall; 2003.
17. Levene, M., Loizou, G. Computing the entropy of user navigation in the web, Department of Computer Science, University College London; 1999.
18. Available at: http://www.nielsen-netratings.com.
19. Bamshad Mobasher, Robert Cooley, Jaideep Srivastava. Creating adaptive web sites through usage-based clustering of URLs. *In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*. 1999.
20. Masseglia, F., Poncelet, P., Cicchetti, R. Webtool: An integrated framework for data mining. *In proceedings of the Ninth International Conference on Database and Expert System Application (DEXA'99)*; 1999.
21. Shigeru Oyanagi, Kazuto Kubota, Akihiko Nakase. Mining WWW access sequence by matrix clustering. *SIGKDD Explorations*. 4(2): 125p.
22. Robert W. Cooley. Web usage mining: Discovery and application of interesting patterns from web data. A Ph. D. Thesis; 2000.
23. Jaideep Srivastava, Robert Cooley, Mukund Deshpande et al. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*. 2000; 1(2).
24. Smith K.A., Ng A. Web page clustering using a self-organizing map of user navigation patterns. *Decision Support Systems*. 2003; 35(2): 245–256p.
25. Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, Vishal Shah. Knowledge discovery from users web-page navigation. *IEEE RIDE*; 1997.
26. Myra Spiliopoulou, Lukas C. Faulstich. WUM: A web utilization miner, in International workshop on the web and databases (WebDB98), Valencia, Spain; 1998.
27. Zarkesh, J. Adibi. Pathmining: Knowledge discovery in partially ordered databases; 1997.
28. O. R. Zaiane, M. Xin, J. Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. *In Proc. Advances in Digital Libraries Conference (ADL'98);* 1998.