WEB LOG MINING

Project Report Submitted by

Ashik S
(4NM13CS031)

Ashish
(4NM13CS202)

Anjan Kumar L V
(4NM14CS402)

Nagesh
(4NM14CS412)

UNDER THE GUIDANCE OF
Mr.Puneeth R P
Assistant professor
Department of Computer Science and Engineering

in partial fulfillment of the requirements for the award of the Degree of

Bachelor of Engineering in
Computer Science & Engineering

from

Visvesvaraya Technological University, Belgaum

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

N.M.A.M. INSTITUTE OF TECHNOLOGY

(An Autonomous Institution under VTU, Belgaum)
(AICTE approved, NBA Accredited, ISO 9001:2008 Certified)
NITTE -574 110, Udupi District, KARNATAKA

April 2017

N.M.A.M. INSTITUTE OF TECHNOLOGY

(An Autonomous Institution under VTU, Belgaum)

(AICTE approved, NBA Accredited, ISO 9001:2008 Certified)

NITTE -574 110, Udupi District, KARNATAKA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled

"WEB LOG MINING"

is a bonafide work carried out by

Ashik S(4NM13CS031) *Ashish(4NM13CS202)*

Anjan Kumar L V(4NM14CS402) *Nagesh(4NM14CS412)*

in partial fulfillment of the requirements for the award of

Bachelor of Engineering Degree in Computer Science and Engineering

prescribed by Visvesvaraya Technological University, Belgaum

during the year 2016-2017.

It is certified that all corrections/suggestions indicated for Internal Assessment have been Incorporated in the report deposited in the departmental library.

The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the Bachelor of Engineering Degree.

Signature of Guide **Signature of HOD** **Signature of Principal**

Semester End Viva Voce Examination

Name of the Examiners	Signature with Date
1. ____________________________	_________________________
2. ____________________________	________________________

ACKNOWLEDGEMENT

We believe that our project will be complete only after we thank the people who have contributed to make this project successful.

First and foremost, our sincere thanks to our beloved principal, **Dr. Niranjan N. Chiplunkar** for giving us an opportunity to carry out our project work at our college and providing us with all needed facilities.

We express our deep sense of gratitude and indebtedness to **Dr. UdayaKumar Reddy**, Head of Department, Computer Science and Engineering, for his inspiring guidance, constant encouragement, support and suggestion for improvement during the course of our project.

We would like to thank our guide **Mr. Puneeth R P**, Assistant professor, Department of Computer Science and Engineering for his support, guidance and encouragement .Our thanks to the project co-ordinators **Mr.Raju K, Mrs.Asmita Poojary** and **Mr.Ranjan Kumar H S** for their constant support.

We also thank all those who have supported us throughout our project.

Finally, thanks to staff members of the Department of Computer Science and Engineering and all our friends for their honest opinions and suggestions throughout the course of our project.

Ashik S

Ashish

Anjan Kumar L V

Nagesh

ABSTRACT

Web Mining refers to extraction of knowledge from the web log data by application of data mining techniques. WUM generally consists of Web Log Pre-processing, Web Log Knowledge Discovery and Web Log Pattern Analysis. Web Log Pre-processing is a major and complex task of WUM. Elimination of noise and irrelevant data, thereby reducing the burden on the system leads to efficient discovery of patterns by further stages of WUM. In this paper, Web Log Pre-processing Methods to efficiently identify users and user sessions have been implemented and results have been analyzed.

Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analysing these log files gives a neat idea about the user. This paper gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn give way to an effective mining. It also provides the idea of creating an extended log file and learning the user behaviour.

TABLE OF CONTENTS

CONTENTS **PAGE NO**

4. System Design 13-16

5. System Implementation 17-19

6. System Testing 20-25

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In the current era, we are witnessing a surge of Web Usage around the globe. A large volume of data is constantly being accessed and shared among a varied type of users; both humans and intelligent machines. Thus, taking up a structured approach to control this information exchange, has what made Web Mining as a hot topic in the field of Information Technology.

The proposed research, mainly concentrates on Web Usage Mining(WUM) as a means to track the behavioural patterns of users surfing either a web site or a page. Web Usage mining consists of three main steps: Pre-processing, Knowledge Discovery and Pattern Analysis.

Web Log Pre-processing is a major and complex task of WUM. Elimination of noise and irrelevant data, thereby reducing the burden on the system leads to efficient discovery of patterns by further stages of WUM. Here, Web Log Pre-processing Methods to efficiently identify users and user sessions have been implemented and results have been analyzed.

Mining for knowledge from Web log data has the prospective of revealing useful information. In general context, Web Usage Mining is employing Data Mining algorithms over the Web Usage Data. However, this process is not just adapting existing algorithms to new data. The WUM algorithms take as input the raw web log which is a rich repository of user activities on the web. A user session represents the sequences of page accesses that a user does in his visit to a Web site.

1.2 PROBLEM STATEMENT

The Web Usage Mining is a recent research field of Data Mining. The fast and dynamic development of WUM left little time to the Data Mining analysts to understand and solve the problems arising in this field. Research dealing with the Web usage data must overcome a number of issues. We list below the current main WUM open problems:

- The quantity of data is continuously increasing;
- The pre-processing step does not receive enough analysis efforts;
- The Web sites have no or little semantic definitions for their Web pages;
- The sequential pattern mining techniques for WUM are not appropriate for dealing with the specifics of Web usage data, mainly with its huge quantity.

As previously mentioned, the main problem for the WUM analysis is represented by the huge quantity of data currently collected. More and larger Web sites and more visitors are the main causes for this. Moreover, Web sites' owners are struggling to keep and increase the number of their visitors as this is directly related to the profits the Web sites generate. Thus, understanding the needs of their users is vital for the Web sites' owners.

Rushing to analyze usage data without a proper pattern discovery method will lead to poor results or even to failure. Thus here we use APFT algorithm to efficiently analyze the web server logs and discover frequently accessed pages along with path chosen to reach them.

1.3 STUDY AREA

Our area of study includes PHP, HTML, CSS and MYSQL SERVER.

1.4 OBJECTIVE

1. Shortening Paths of High visit Pages: The pages which are frequently accessed by the users can be seen as to follow a particular path. These pages can be included in an easily accessible part of the Website thus resulting in the decrease in the navigation path length.

2. Eliminating or Combining Low Visit Pages: The pages which are not frequently accessed by users can be either removed or their content can be merged with pages with frequent access.

3. Redesigning Pages to help User Navigation: To help the user to navigate through the website in the best possible manner, the information obtained can be used to redesign the structure of the Website.

4. Redesigning Pages for Search Engine Optimization: The content as well as other information in the website can be improved from analyzing user patterns and this information can be used to redesign pages for Search Engine Optimization so that the search engines index the website at a proper rank.

5. Help Evaluating Effectiveness of Advertising Campaigns: Important and business critical advertisements can be put up on pages that are frequently accessed.

1.5 METHODOLOGY

- Web Usage Mining in particular deals with the nature of the data. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of pre-processing and parsing before the actual extraction of the required information.
- The pre-processing task within the WUM process involves Data Cleaning and Data Structuration. Pattern-Discovery task involves Association Rule Mining, Statistical Analysis, Clustering, Classification and Sequential Patterns. The pattern analysis is the final step of the WUM process.
- A new algorithm named APFT combines the Apriori algorithm and FP-tree structure which is proposed in FP-growth algorithm.

1.6 ORGANIZATION OF THE REPORT

This report is divided up into chapters, each dealing with different aspects of the project. Each chapter has a short introduction, explaining the subject of each chapter, and then the detail of each module is explained seperately.

The following is a short overview of each of the chapters:

Chapter 2: The review of literature of summarization with web log mining tasks by referring international papers are done.

Chapter 3: Discusses the System requirements in which functional and non-functional needs are defined and also the feasibility study is done.

Chapter 4: Illustration of the design of proposed system is covered. The chapter is divided into four sections: Pre-processing functions, Design model of Summarization, Data Flow Diagram, and Database Design.

Chapter 5: Focuses on the implementation aspect of the proposed systems design and Algorithms of main modules are discussed. Also method description and signature of important functions are discussed.

Chapter 6: Focuses on the testing of the proposed summarization module.

Chapter 7: Focuses on the results where snapshot of the output are discussed along with performance analysis.

Chapter8: Conclusions and feature work is discussed. Also, the references and appendices are placed at the end of the report.

CHAPTER 2

LITERATURE REVIEW

2.1 OVERVIEW

Literature survey is mainly carried out in order to analyze the background of the current project which helps to find out flaws in the existing system & guides on which unsolved problems we can work out. So, the following topics not only illustrate the background of the project but also uncover the problems and flaws which motivated to propose solutions and work on this proposed and to be implemented project. It describes the work which has already been done and innovative things or new scope of the proposed system.

2.2 EXISTING SYSTEM

ShaaH.etal has proposed EPLogCleaner Method. EPLogCleaner that can filter out plenty of irrelevant items based on the common prefix of their URLs. Experimental results show that EPLogCleaner can improve data quality of enterprise proxy logs by further filtering out more than 30% URL requests comparing with traditional data cleaning methods.

TyagiN.etal provides an algorithmic approach to data pre-processing in web usage mining. They take requests for graphical page content, or the other file which can be induced into a webpage, or navigation sessions performed by robots and web spiders into consideration.

ZhengL.eal has proposed Optimized User Identification, Optimized Session Identification. The optimized data pre-processing technology is used to improvement of the technology betters the quality of data pre-processing results. The strategy based on the referred webpage is adopted at the stage of user identification. Experiment shave proved that advanced data pre-processing technology can enhance the quality of data pre-processing results.

Nithya P and Sumathi P have proposed novel pre-processing technique. This method describes removing local and global noise and web robots. This paper continues the line of research on Web access log analysis is to analyze the patterns of website usage and the features of user's behavior. It is vital to preprocess the log data for efficient webusage mining process.

SujataV.etal has planned Prediction of User navigation patterns using Clustering and Classification (PUCC) from weblog data. In the initial stage PUCC focuses on separating the potential users in weblog data, and in the second stage clustering process is used to cluster the potential users with similar interest and within the third stage the results of classification and clustering is used to predict the user future requests.

Theint has proposed data mining techniques to discover user access patterns from web log. This paper mainly focuses on data pre-processing stage of the first phase of web usage mining with activities like field extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from the single line of the log file.

LosarwarV.etal has discussed the importance of data pre-processing methods and various steps involved in getting the required content effectively .A complete pre-processing technique is being proposed to pre-process the weblog for extraction of user patterns. Data cleaning algorithm removes their relevant entries from weblog.

Munk M.etal has tried to assess the impact of reconstruction of the activities of a web visitor on the quantity and quality of the extracted rules which represent the web user behaviour patterns. Experiment, find out to which criteria are necessary to realize this time-consuming data preparation and specifying the inevitable steps that are required for obtaining valid data from the log file.

2.3 PROPOSED SYSTEM

Data ware house is the only viable solution that can bring that dream into reality. The enhancement of future endeavours to make decisions depends on the availability of correct information that is based on the quality of data underlying. The quality data can only be produced by cleaning data prior to loading into the data ware house since the data collected from different sources will be dirty. Once the data have been cleaned it will produce accurate results when the data mining query is applied to it. So correctness of the data is essential for well-formed and reliable decision making.

Data pre-processing is an important step to filter and organize only appropriate information before applying any web mining procedure. Pre-processing reduce log file size and also increase quality of available data. The purpose of data pre-processing is to improve data quality and increase data mining accuracy. Pre-processing consists of: data cleansing, user identification, session identification. Here the main task is to clean the raw web log files and insert the processed data into a relational database. In this step we remove noisy as well as unnecessary data. Remove log entry nodes contain file extension like jpg, gif means remove request such as multimedia files, image, page style file. The proposed system, mainly concentrates on Web usage mining as a means to track the behavioural patterns of users surfing either a web site or a page.

CHAPTER 3

SOFTWARE REQUIREMENT SPECIFICATION

3.1 INTRODUCTION

A software requirements specification is a description of a software system to be developed. Software requirement consists of the functional and non functional requirements for the application being developed. It briefs about the software requirements for the system.

The software requirements specification document enlists enough and necessary requirements that are required for the project development. To derive the requirements we need to have clear and thorough understanding of the products to be developed or being developed. This is achieved and refined with detailed and continuous communications with the project team and customer till the completion of the software.

3.2 FUNCTIONAL REQUIREMENTS

Functional requirements consist of two types of requirements, Hardware requirements and Software requirements.

3.2.1 Pre-Processing

The pre-processing task within the WUM process involves cleaning and structuring data to prepare it for the pattern discovery task. Web usage data is subject to a lot of noise and missing data. The Web site's structure, its content and the Web server technologies behind the Web site heavily influence on the nature and size of the usage data collected.

- **Data Cleaning-** The first pre-processing task is the data cleaning. Here, all irrelevant and noisy data are eliminated from the log files. Usually, requests for images and multimedia files are filtered out. When requesting a Web page containing additional Web resources like images or script files, several implicit requests will be generated by the Web browser. If these requests are still present when the data mining step is performed, uninteresting patterns like “Page, Image1, Image3, Image6" may be found, making the pattern analysis step more complex.

- **Data Structuration**- During this step, the requests from the raw log file is grouped by user, user session, page view, visit and episode. Grouping requests by user, also called *user identification*, depends on the Web site policies. For Web sites requesting registration, the user identification task is straightforward and guaranteed to be correct. For all other methods (e.g. using the IP address or cookies), the accuracy of the user identification is not guaranteed.

  The *user session* contains all the requests of a user, made from the same computer and within a considered amount of time. *Visits and episodes identification* depend on the purpose of the analysis. For Some analysis, grouping the requests in user sessions and keeping one request per page view is enough and there is no need to further split the user session. But, depending on the period analyzed, one user session may span over several months and, therefore, the requests will not be related. In this case, we need to split the user session after a certain time lapse. Furthermore, in Web portals like Yahoo, one user may view during a short period, several Web pages that have completely different purposes (e.g. e-mail, online shopping, stock values). These pages will all be grouped under a single visit and analyzed together although they are not related. Therefore, it is better to group these pages in episodes according to their content and analyze them separately.

3.2.2 Pattern – Discovery

Association Rule Mining- The algorithms for mining association rules (ARs) were first developed for the market basket analysis. Apriori is the first and still the most used algorithm for this task. By using this algorithm, we can extract interesting correlations from the data, like a list of items that are frequently bought together.

Applying ARs to Web usage data means extracting items (i.e. Web pages) that frequently occur together. In this case, the notion of frequent pattern depends on a minimum support expressing the minimum number of transactions (i.e. user sessions, visits or episodes) that needs to be contained in the items forming the AR. Thus, association rules express frequent co-occurrences of Web pages together.

The general form of an AR is: X => Y, where X and Y are sets of Web pages. A support of s% for this AR means that X and Y are contained together in s% of the transactions. Association represents a valuable result as they allow making useful recommendations to Web sites' users depending on the pages they visited.

Statistical Analysis:

Statistical techniques are the most common method to ex-tract knowledge about visitor to a website. By analyzing the session file one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path. Many Web traffic analysis to produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This report may include limited low-level error analysis such as detecting unauthorized entry point. Despite lacking in the depth of its analysis. This type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for marketing decisions.

Clustering:

Clustering is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to b e discovered: usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for Internet search engines and

Web assistance providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs.

Classification:

Classification is the task of mapping a data item into 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 b e done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neigh- b or 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.

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. By using this approach, Web marketers can predict future visit patterns which will b e helpful in placing advertisements aimed at certain user groups. Other types of temp oral analysis that can b e performed on sequential patterns include trend analysis, change point detection, or similarity analysis.

3.3 NON-FUNCTIONAL REQUIREMENTS:

Non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviours.

- Forward compatibility - Forward compatibility aims at the ability of a design to gracefully accept input intended for later versions of itself.
- Response time - In technology, response time is the time a system or functional unit takes to react to a given input.
- Portability - Portability in high-level computer programming is the usability of the same software in different environments.

- Usability - Usability is the ease of use and learn ability of a human-made object.

3.4 HARDWARE REQUIREMENTS

Processor	: Intel Core* family
Processor Speed	: 2.0 GHz and above
Hard Disk space	: 30GB
Ram Memory	: 8 GB

3.5 SOFTWARE REQUIREMENTS

Operating System:	Windows
Database Server:	MS SQL
IDE Used:	Dream Viewer for HTML, PHP
Tools/Library files:	Apache Sever.

CHAPTER 4

SYSTEM DESIGN

4.1 System design

This stage of Software development is required to understand and implement the lowest level functionality of the software. The description of the most basic modules which have a single input and at least a single output are provided. These modules are combines thereafter to provide the complex functionalities of the high-level design.

4.2 Design of user interface

The input design, user-oriented inputs are converted into a computer based system format. It also includes determining the record media, method of input, speed of capture and entry on to the screen. Online data entry accepts commands and data through a keyboard. The major approach to input design is the menu and the prompt design. In each alternative, the user's options are predefined. The data flow diagram indicates logical data flow, data stores, source and destination. Input data are collected and organized into a group of similar data. Once identified input media are selected for processing.

Also the important input format is designed in such a way that accidental errors are avoided. The user has to input only just the minimum data required, which also helps in avoiding the errors that the users may make. Accurate designing of the input format is very important in developing efficient software. The goal or input design is to make entry as easy, logical and free from errors.

4.3 Design of system interface

The output design, the emphasis is on producing a hard copy of the information requested or displaying the output on the CRT screen in a predetermined format. Two of the most output media today are printers and the screen. Most users now access their reports from a hard copy or screen display. Computer's output is the most important and direct source of information to the user, efficient, logical, output design should improve the systems relations with the user and help in decision-making. As the outputs are the most important source of information to the user, better design should improve the system's relation and also should help in decision-making. The output device's capability, print capability, print capability, response time requirements etc. should also be considered.

4.4 SYSTEM ARCHITECTURE

Architectural design values make up an important part of what influences architects and designer when they make their design decisions. However, architects and designers are not always influenced by the same values and intensions differ between different architectural movements.

The modular program structure explains how system can be modularized and explain the relation between the modules to achieve the complete functionality of the system. This is a high level overview of how software design document responsibility of the system were partitioned and then assigned to the subsystems. It defines each high level subsystem and the roles or responsibilities assigned to it. This product has been developed using the modularized approach. The product developed is implemented first, tested for functionality and then the features of the module are used as feedback for the development of the next.

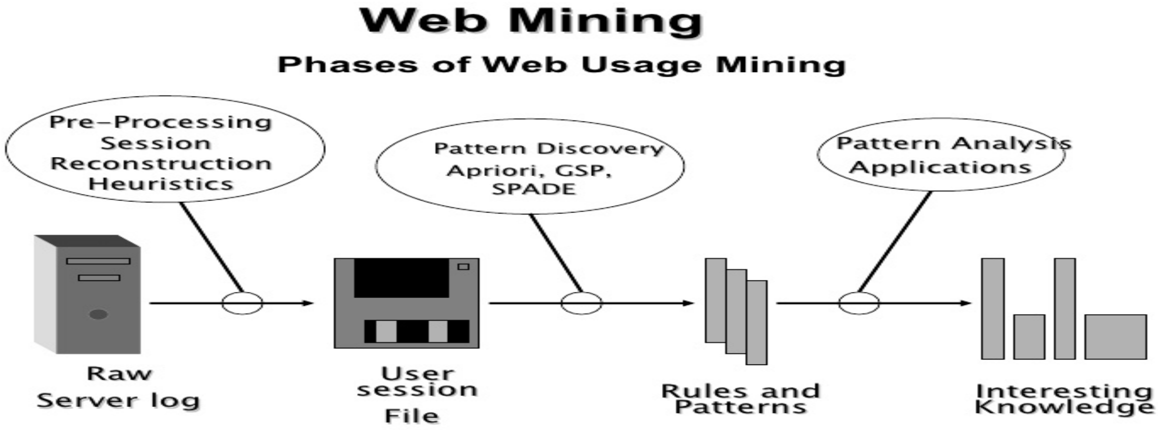

Web Mining
Phases of Web Usage Mining
Pre-Processing Session Reconstruction Heuristics
Pattern Discovery Apriori, GSP, SPADE
Pattern Analysis Applications
Raw Server log
User session File
Rules and Patterns
Interesting Knowledge

Figure 4.1: Architecture of Web Usage Mining System

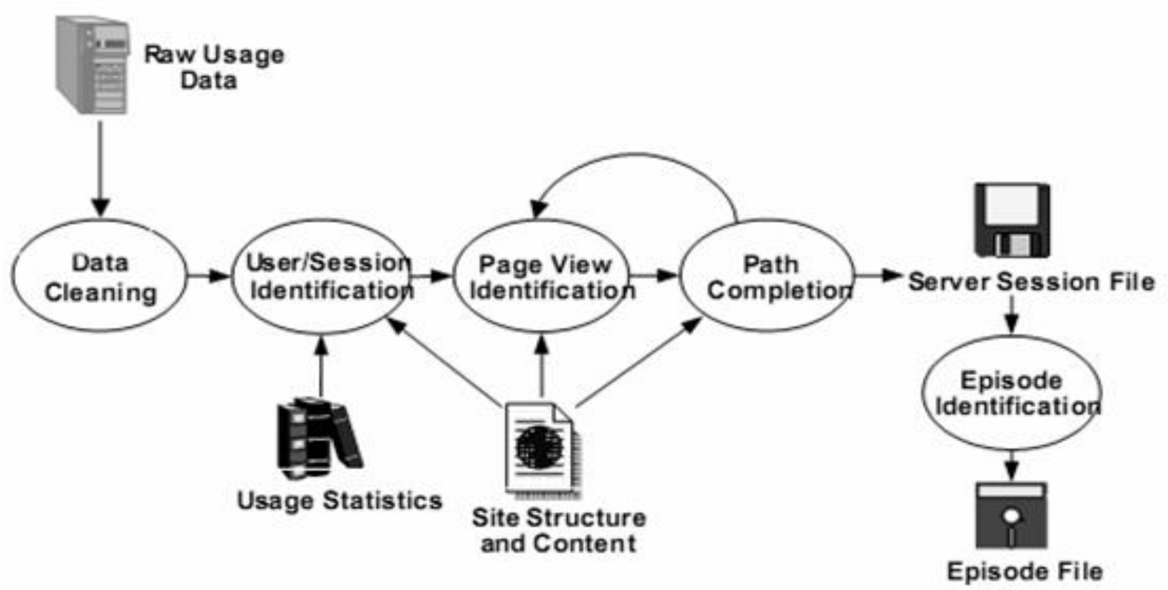


Figure 4.2: Decomposition Description

4.5 Data Flow Diagram

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.

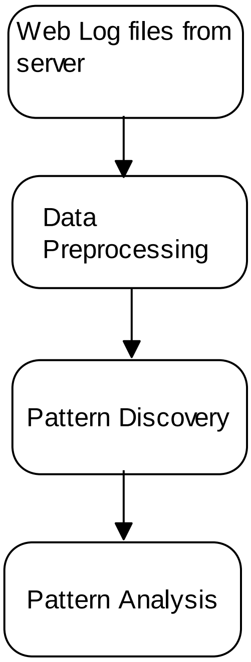


Figure 4.3: Data flow diagram for web log mining

4.6 Database Design

Database is designed to manage large facts data of information. The organizations of data engage both the definitions of structures for the storage and information. The term database design can be normally used to describe many different parts of the design of a complete database system. The database design includes an entity-relationship diagram and the tables that are used in the data

Table name : Itemset

Column Name	Data Type	Description
Item1	Varchar(8000)	Description of first field from log file.
Item2	Varchar(8000)	Description of second field from log file.
Item3	Varchar(8000)	Description of third field from log file.
Item4	Varchar(8000)	Description of fourth field from log file.

Table 4.1:Itemset.

CHAPTER 5

SYSTEM IMPLEMENTATION

5.1 MODULE DESCRIPTION:

Implementation details are given in this chapter. It is described in two parts, namely traditional algorithm and new algorithm.

5.2 Module Design

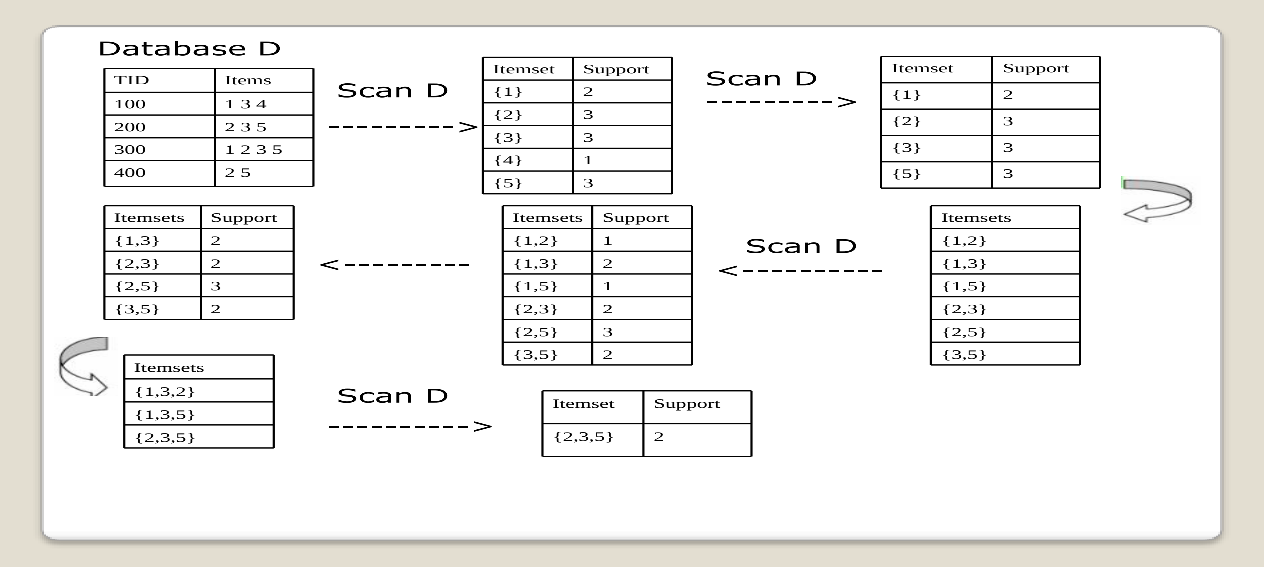
The system has the following modules:

5.3 Traditional Algorithm Module.

Apriori Algorithm

Uses a Level-wise search, where k-itemsets (An itemset that contains k items is a k-itemset) are used to explore (k+1)-itemsets, to mine frequent itemsets from transactional database for Boolean association rules.

First, the set of frequent 1-itemsets is found. This set is denoted L1. L1 is used to find L2, the set of frequent 2 itemsets, which is used to find L3, and so on.



TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Itemset	Support
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

Itemsets
{1,2}
{1,3}
{1,5}
{2,3}
{2,5}
{3,5}

Itemsets	Support
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

Itemsets	Support
{1,3}	2
{2,3}	2
{2,5}	3
{3,5}	2

Itemsets
{1,3,2}
{1,3,5}
{2,3,5}

Itemset	Support
{2,3,5}	2

Figure 5.1: Apriori algorithm.

5.4 FP-Tree Module.

Pseudo code

Coding done in 3 phases:

- Removing infrequent 1-itemsets from each sessions
- Constructing an Fp-Tree
- Finding Frequent itemsets and their path by applying Apriori Algorithm on each branches of the Fp-Tree

Removing infrequent 1-itemsets from each sessions

- Find the count (number of occurrences of each page).
- If count < Support(pre-defined value)

```
                    Eliminate the page
The pages eliminated here are infrequent
```

- Constructing an Fp-Tree
- Make the root node value as NULL
- Take the first itemset and insert it as the first
  branch(root.fp[0]) of Fp-Tree
- Take the next itemset

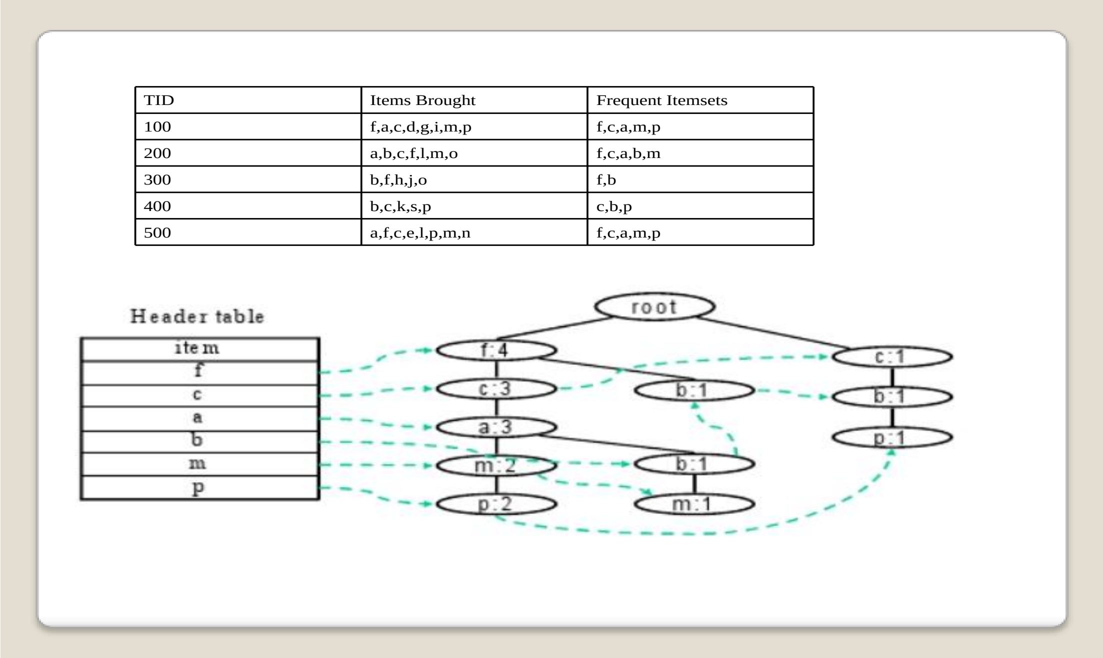
```
 Compare the first item to the first node of the previous I
  branches
  If (first item== any of the first node of previous branches)
              Go to the child of that node and take the next
  item from the itemset
        Repeat step4
        Else
              Insert the item at position i
     i->no of previous branches.
```

Finding frequent itemsets and their path by applying Apriori Algorithm on each branches of the Fp-Tree

- Consider each branch of Fp-Tree.
- Form subsets (1-itemset, 2-itemset, and so on).

- Compare the subsets with each of the pre-processed input and find their counts (occurrences of each of the subset).
- If count < Support (pre-defined value).

  Eliminate the subset.



TID	Items Brought	Frequent Itemsets
100	f,a,c,d,g,i,m,p	f,c,a,m,p
200	a,b,c,f,l,m,o	f,c,a,b,m
300	b,f,h,j,o	f,b
400	b,c,k,s,p	c,b,p
500	a,f,c,e,l,p,m,n	f,c,a,m,p

Figure 5.2: Reverse Apriori algorithm.

CHAPTER 6

SYSTEM TESTING

6.1 INTRODUCTION

Testing is a process of executing the program with the intent of finding an error, it is a set of activities that can be planned in advance and conducted systematically. Inadequate testing or non-testing may lead to errors. Nothing is completed without the testing, as it is vital kinds of data. While testing, errors are noted and corrections are made. So in general testing demonstrates that the system is working according to the specifications, and that it meets the performance requirement.

System testing is done to check whether the system works accurately and efficiently before live operation commences. A small error can conceivably explode into much larger problem. Effective testing early in the process translates directly into long term cost savings from a reduced number of errors. Testing is the systematic search for details in all project deliverable. It is the process of examining an output of a process under consideration, comparing the results against a set of predetermined expectations and dealing with variances.

Testing is vital for the success of the system. System testing makes a logical assumption that if all parts of the system are correct, the goal will be successively achieved. Testing involves operation of systems or application under controlled conditions and evaluating the results. The controlled conditions include both normal and abnormal conditions. At first, the different units are individually tested and the system as a whole is tested.

Tests are frequently grouped by where they are added in the software development process, or by the level of specificity of test. The main level during the development process are unit testing, integration testing, and system testing that are distinguished by the test target without implying a specific, process model.

6.2 Testing Levels

Testing is part of Verification and Validation. Testing plays a very critical role for quality assurance and for ensuring the reliability of the software. The objective of testing can be stated in the following ways.

- A successful test is one that uncovers as-yet-undiscovered bugs.
- A better test case has high probability of finding un-noticed bugs.
- A pessimistic approach of running the software with the intent of finding errors.

Testing can be performed in various levels like unit test, integration test and system test.

6.3 Unit Testing

Unit testing tests the individual components to ensure that they operate correctly. Each component is tested independently, without other system component. This system was tested with the set of proper test data for each module and the results were checked with the expected output. Unit testing focuses on verification effort on the smallest unit of the software design module.

Ideally, each test cases are independent than others. Substitutes such as method stubs mock objects, fakes, and test harnesses can be used to assist testing a module and isolation. In our project, Login module, Apriori algorithm module and Reverse Apriori algorithm module are tested separately. For example, while user authentication is done we check for the proper username and password.

6.4 Integration Testing

Integration testing is another aspect of testing that is generally done in order to uncover errors associated with the flow of data across interfaces. The unit-tested modules are grouped together and tested in small segment, which makes it easier to isolate and correct errors. This approach is continued until we have integrated all modules to form the system as a whole.

The purpose of integration testing is to verify functional, performance, and reliability requirements placed on major design items. These “design items”, i.e. assemblages (or group of units), are exercised through their interfaces using black box testing, success and error being simulated via appropriate parameter and data inputs. The overall idea is a “building block” approach, in which verified assemblages are added to a verified base which is then used to support the integration testing of further assemblages. In our project, various modules are integrated and then tested with different functionalities.

6.5 System Testing

System testing tests a completely integrated system to verify that it meets its requirements. System testing of software or hardware is testing conducted on a completed integrated system to evaluate the system’s compliance with its specified requirements. System testing falls within the scope of black box testing, and as such, should require no knowledge of the inner design of the code or the logic.

As a rule, system testing takes, as its input, all of the “integrated” software components that have passed integration testing and also the software system itself integrated with any applicable hardware system(s). The purpose of integration testing is to detect any inconsistencies between the software units that are integrated together (called assemblages) or between any of the assemblages and the hardware. System testing is a more limited type of testing; it seeks to detect defects both within the “inter-assemblages” and also within the system as a whole.

6.6 Acceptance testing

The test conducted by client to evaluate the system compliance as per the business requirements.

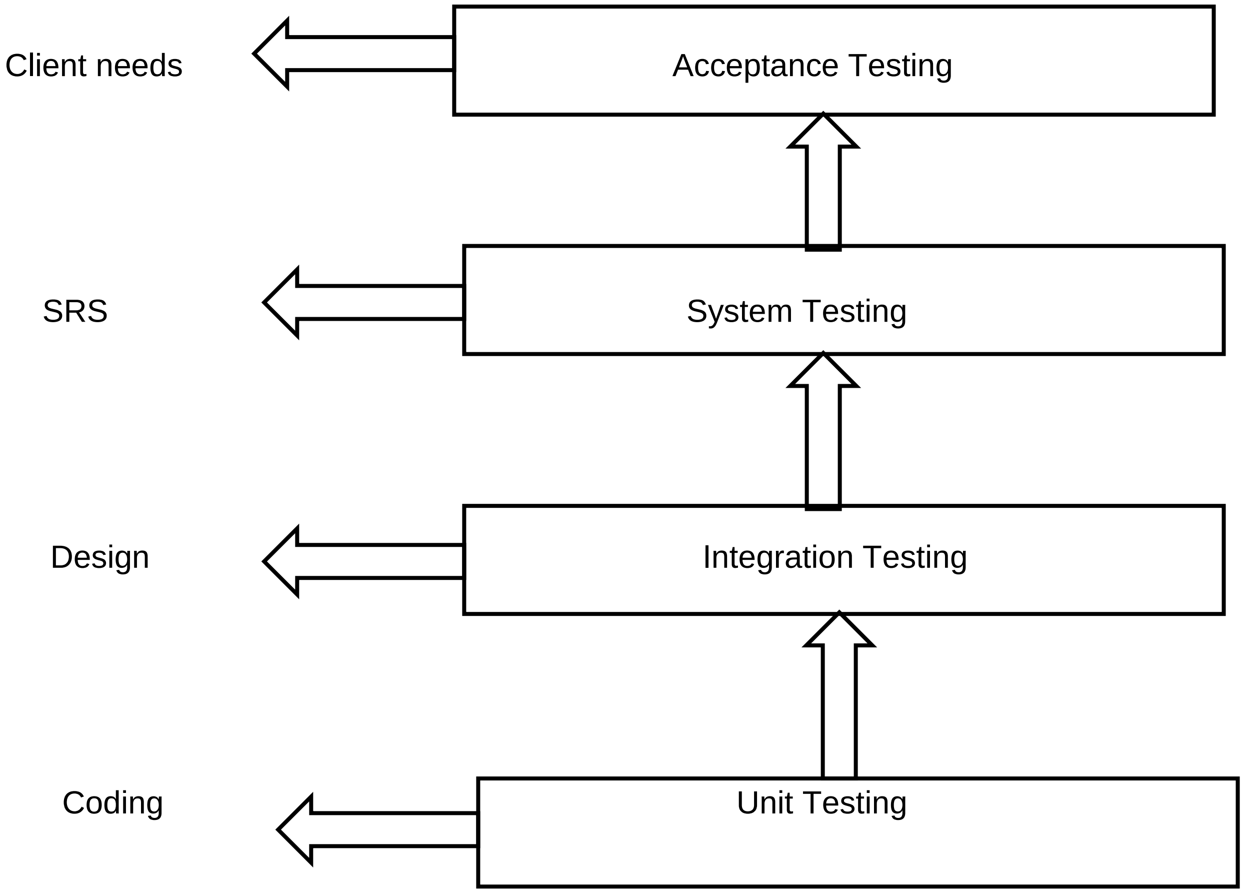


Figure 6.1: Acceptance Testing.

6.7 Test Cases

A test case is a software testing document, which consists of events, action, input, output, expected result and actual result. Technically a test case includes test description, procedure, expected result and remarks. Test cases should be based primarily on the software requirements and developed to verify correct functionality and to establish conditions that reveal potential errors.

6.7.1 Test Cases for Login

Test Cases	Description	Input	Expected Output	Actual output	Result
1	Login	Correct User name and password	Logged in successfully	Logged in successfully	Success
2	Login	Wrong User name and password	Login Failed	Login Failed	Success
3	Login	Correct User name and wrong password	Login Failed	Login Failed	Success
4	Login	Wrong User name and correct password	Login Failed	Login Failed	Success

Table 6.1: Test Cases for Login.

Table 6.7.2 Test Cases for Algorithm.

Test No.	Test case.	File Size.	Apriori Algorithm Time Efficiency in seconds.	Reverse Apriori Algorithm Time Efficiency in seconds.
1.	NASA File	11 K B.	84.407190799713	19.998001813889
2.	NASA File	10 M B.	300.13490345	200.406775394
3.	NASA File	20 M B.	500.43245432	350.3423422425
4.	NASA File	30 M B.	650.42325352	480.6554634566

Table 6.2: Test Cases for Algorithm

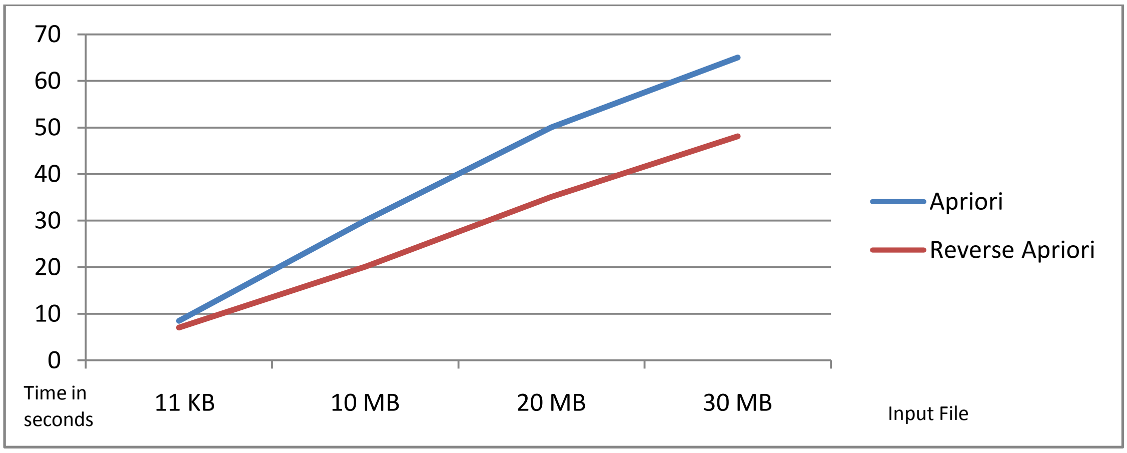


Figure 6.2 : Chart of tested files.

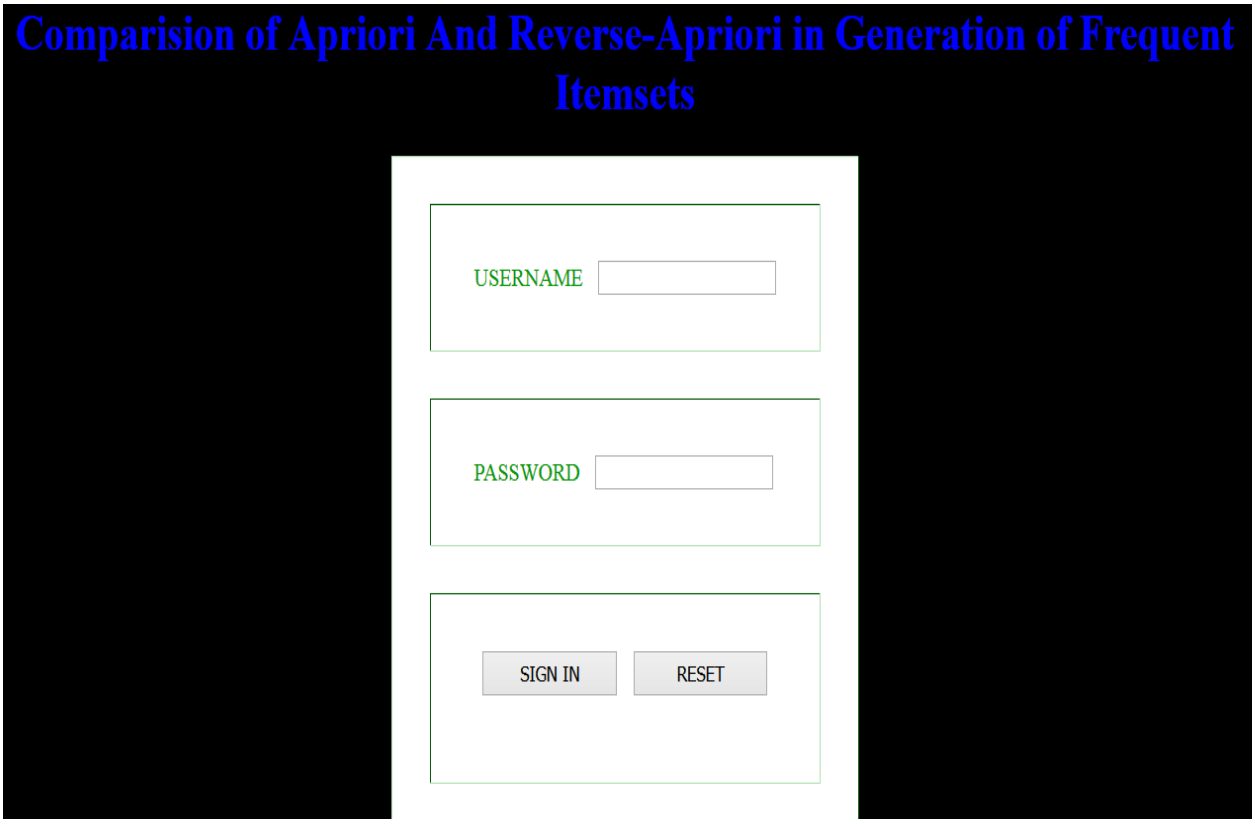
CHAPTER 7

RESULTS AND ANALYSIS

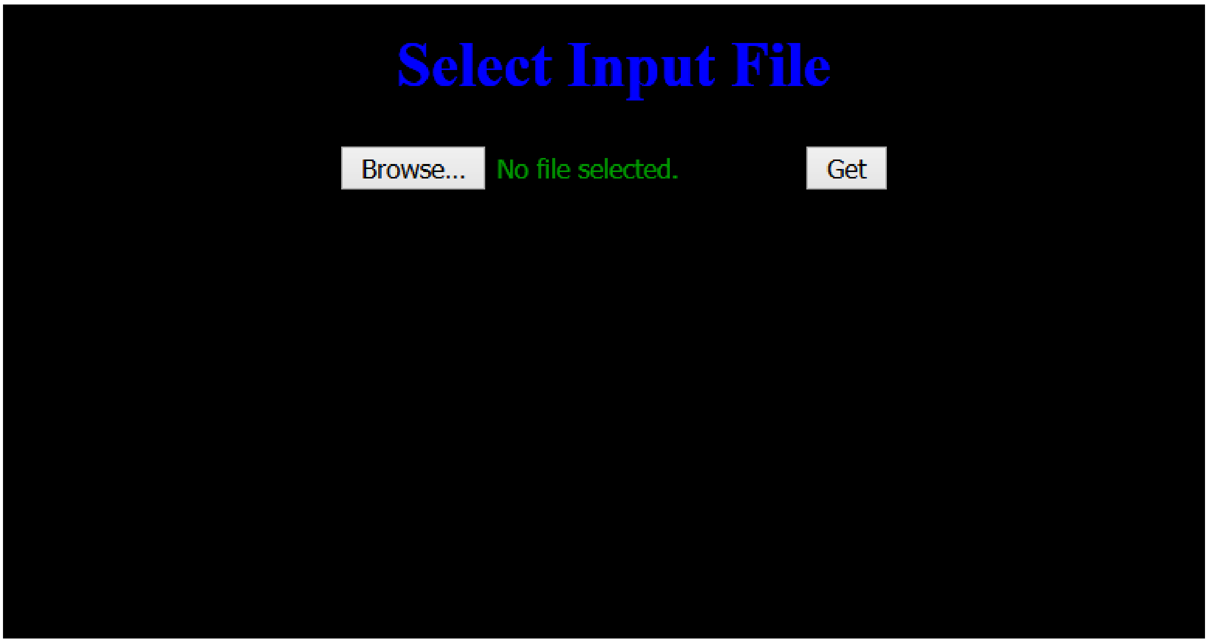
7.1 Results

The results which were obtained after the pattern discovery showed the most frequently accessed page & their path. The results which were obtained after the analysis were satisfactory and contained valuable information about the Log Files.

7.2 Login page for algorithm.

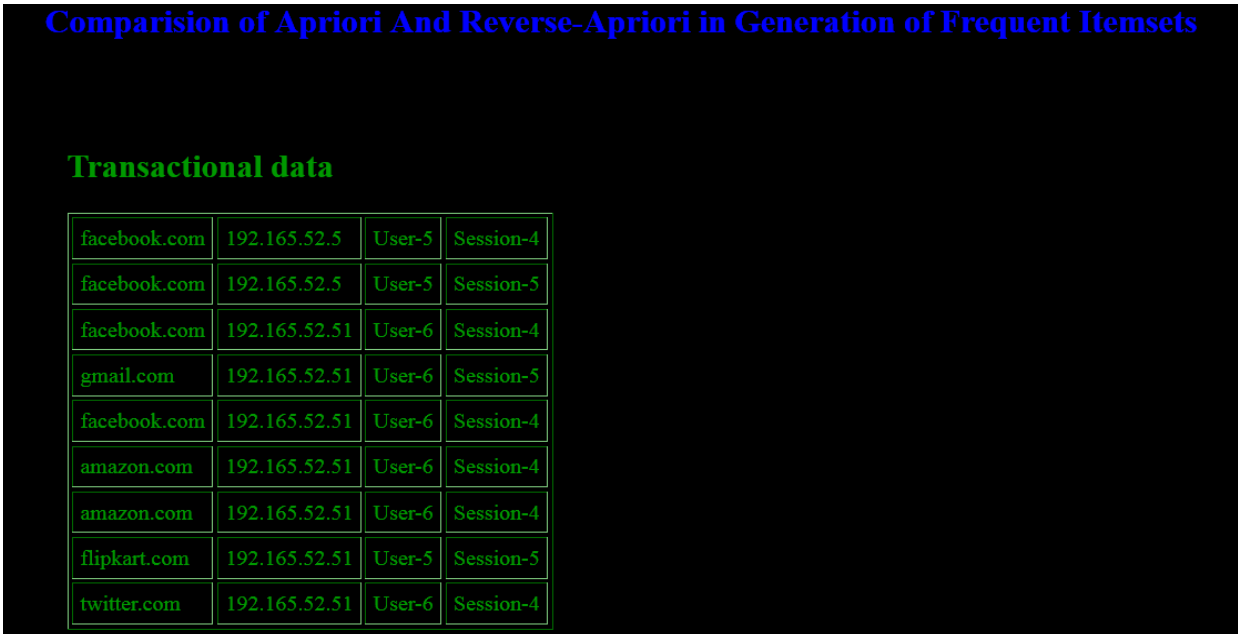

Comparision of Apriori And Reverse-Apriori in Generation of Frequent Itemsets

7.3 **Selecting Input File for algorithm.**



Select Input File

7.4 Displaying input file content.

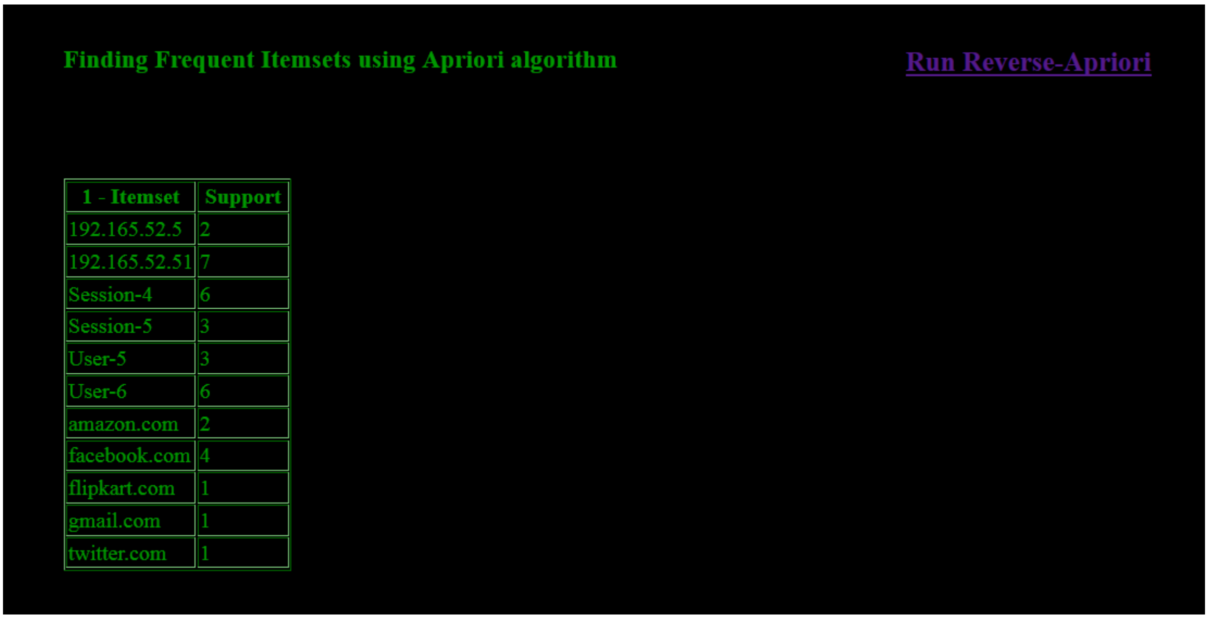


Comparision of Apriori And Reverse-Apriori in Generation of Frequent Itemsets

Transactional data

facebook.com	192.165.52.5	User-5	Session-4
facebook.com	192.165.52.5	User-5	Session-5
facebook.com	192.165.52.51	User-6	Session-4
gmail.com	192.165.52.51	User-6	Session-5
facebook.com	192.165.52.51	User-6	Session-4
amazon.com	192.165.52.51	User-6	Session-4
amazon.com	192.165.52.51	User-6	Session-4
flipkart.com	192.165.52.51	User-5	Session-5
twitter.com	192.165.52.51	User-6	Session-4

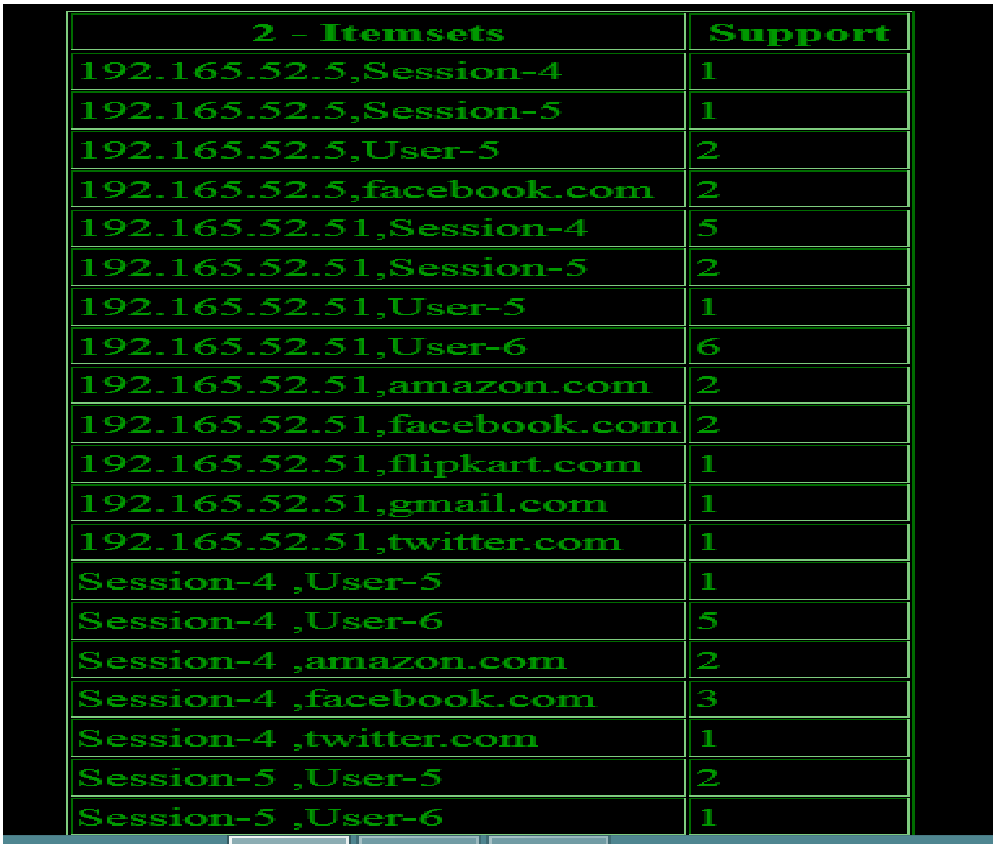
7.5 Listing Total Number of items.



Finding Frequent Itemsets using Apriori algorithm

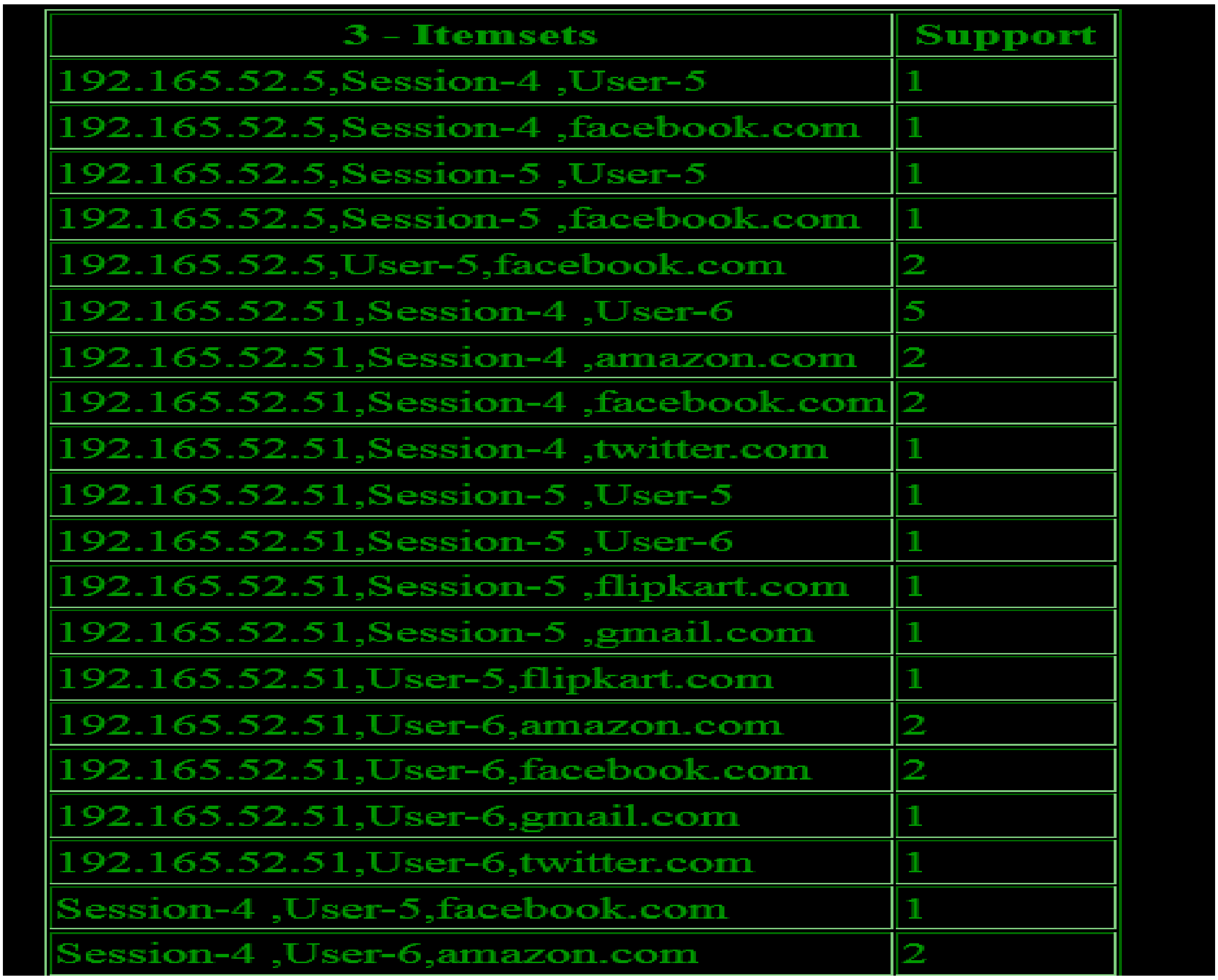
1 - Itemset	Support
192.165.52.5	2
192.165.52.51	7
Session-4	6
Session-5	3
User-5	3
User-6	6
amazon.com	2
facebook.com	4
flipkart.com	1
gmail.com	1
twitter.com	1

7.6 Comparing Two data items.



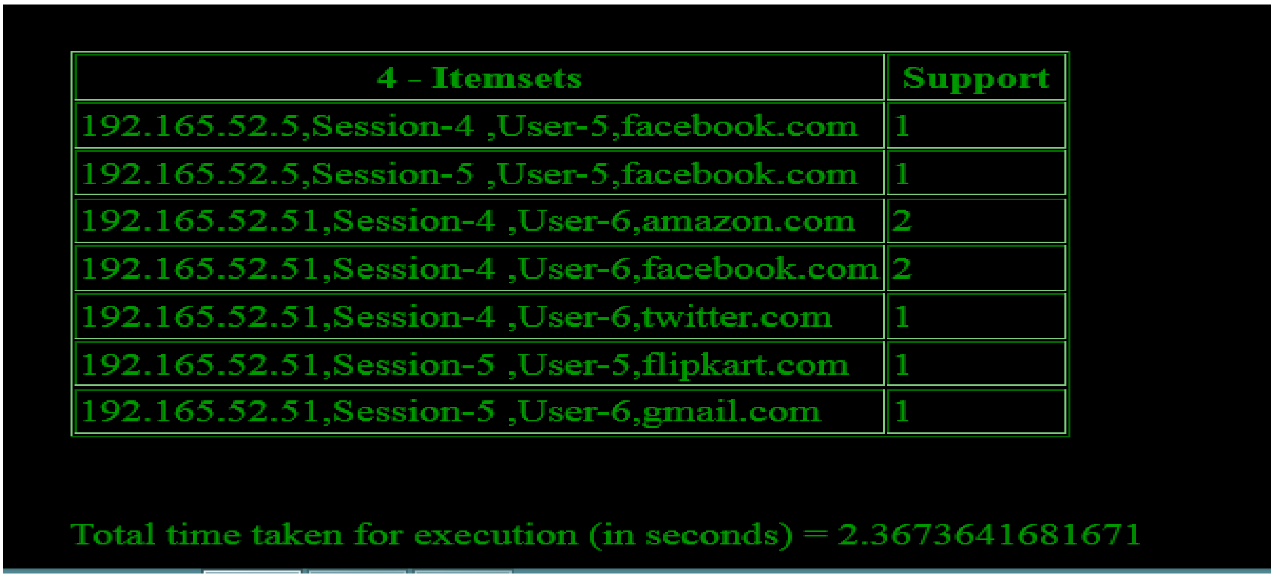
2 - Itemsets	Support
192.165.52.5,Session-4	1
192.165.52.5,Session-5	1
192.165.52.5,User-5	2
192.165.52.5,facebook.com	2
192.165.52.51,Session-4	5
192.165.52.51,Session-5	2
192.165.52.51,User-5	1
192.165.52.51,User-6	6
192.165.52.51,amazon.com	2
192.165.52.51,facebook.com	2
192.165.52.51,flipkart.com	1
192.165.52.51,gmail.com	1
192.165.52.51,twitter.com	1
Session-4 ,User-5	1
Session-4 ,User-6	5
Session-4 ,amazon.com	2
Session-4 ,facebook.com	3
Session-4 ,twitter.com	1
Session-5 ,User-5	2
Session-5 ,User-6	1

7.7 Comparing three data items.



3 - Itemsets	Support
192.165.52.5,Session-4 ,User-5	1
192.165.52.5,Session-4 ,facebook.com	1
192.165.52.5,Session-5 ,User-5	1
192.165.52.5,Session-5 ,facebook.com	1
192.165.52.5,User-5,facebook.com	2
192.165.52.51,Session-4 ,User-6	5
192.165.52.51,Session-4 ,amazon.com	2
192.165.52.51,Session-4 ,facebook.com	2
192.165.52.51,Session-4 ,twitter.com	1
192.165.52.51,Session-5 ,User-5	1
192.165.52.51,Session-5 ,User-6	1
192.165.52.51,Session-5 ,flipkart.com	1
192.165.52.51,Session-5 ,gmail.com	1
192.165.52.51,User-5,flipkart.com	1
192.165.52.51,User-6,amazon.com	2
192.165.52.51,User-6,facebook.com	2
192.165.52.51,User-6,gmail.com	1
192.165.52.51,User-6,twitter.com	1
Session-4 ,User-5,facebook.com	1
Session-4 ,User-6,amazon.com	2

7.8 Comparing four data items.

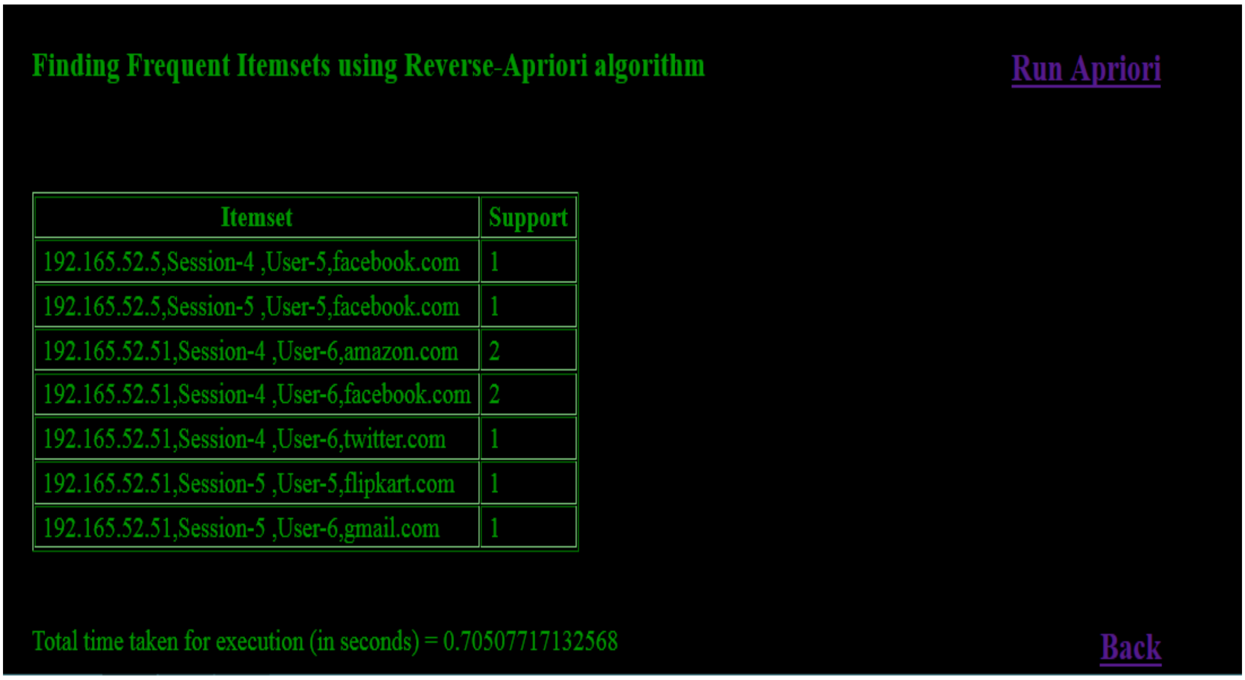


4 - Itemsets	Support
192.165.52.5,Session-4 ,User-5,facebook.com	1
192.165.52.5,Session-5 ,User-5,facebook.com	1
192.165.52.51,Session-4 ,User-6,amazon.com	2
192.165.52.51,Session-4 ,User-6,facebook.com	2
192.165.52.51,Session-4 ,User-6,twitter.com	1
192.165.52.51,Session-5 ,User-5,flipkart.com	1
192.165.52.51,Session-5 ,User-6,gmail.com	1

Total time taken for execution (in seconds) = 2.3673641681671

.

7.9 Output of reverse apriori algorithm.



Finding Frequent Itemsets using Reverse-Apriori algorithm

Itemset	Support
192.165.52.5,Session-4 ,User-5,facebook.com	1
192.165.52.5,Session-5 ,User-5,facebook.com	1
192.165.52.51,Session-4 ,User-6,amazon.com	2
192.165.52.51,Session-4 ,User-6,facebook.com	2
192.165.52.51,Session-4 ,User-6,twitter.com	1
192.165.52.51,Session-5 ,User-5,flipkart.com	1
192.165.52.51,Session-5 ,User-6,gmail.com	1

Total time taken for execution (in seconds) = 0.70507717132568

CHAPTER 8

CONCLUSION AND FUTURE WORK

The results which were obtained after the pattern discovery showed the most frequently accessed page & their path. The results which were obtained after the analysis were satisfactory and contained valuable information about the Log Files.

The implementation of web usage mining which we have proposed is on a single node (single processor). This can be implemented in a distributed manner also using applications like Hadoop, MPI etc. Further, anyone interested in this field can take a similar approach and modify these methods to expand them to a general scenario or to shift their ideas to different areas.

REFERENCES

[1] Web Usage Mining: Web log Pre-processing and Online Visitor’s frequent Pattern Discovery By Aruna Kumari G K, Sudheer Shetty in International Journal of Innovative Research in Computer and Communication Engineering April 2016

[2] Web Usage Mining: An Implementation, National Institute of Technology, Rourkela Rourkela-769008, Orissa, India.

[3] Knowledge Discovery from Web Usage Data: An Efficient Implementation of Web Log Preprocessing Techniques by Shivaprasad G., N.V. Subba Reddy and U. Dinesh Acharya in International Journal of Computer Applications February 2015.

[4] ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING By L.K. Joshila, Grace V.Maheswari and Dhinaharan Nagamalai in International Journal of Network Security & Its Applications (IJNSA) January 2011.

[5] NASA Kennedy space Centre http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html