An Approach for Frequent Access Pattern Identification in Web Usage Mining

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Information Security

Submitted By
Murli Manohar Sharma
Roll No. 801233009

Under the supervision of:
Ms. Anju Bala
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004
July 2014

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"An Approach for Frequent Access Pattern Identification in Web Usage Mining"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Information Security* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Ms. Anju Bala* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Murli Sharma

Murli Manohar Sharma

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Anju Bala

Ms. Anju Bala
Assistant Professor,
Computer Science and
Engineering Department

Countersigned by

Deepak Garg

Dr. Deepak Garg
Head
Computer Science and Engineering Department
Thapar University
Patiala

Dr. S. K. Mohapatra
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

I express my sincere and deep gratitude to my guide Ms. Anju Bala, Assistant Professor in Computer Science & Engineering Department, Thapar University Patiala, for the invaluable guidance, support and encouragement. She provided me all resource and guidance throughout thesis work.

I am heartfelt thankful to Dr. Deepak Garg, Head of Computer Science & Engineering Department Thapar University Patiala, for providing us adequate environment, facility for carrying out thesis work.

I would like to thank to all staff members who were always there at the need of hour and provided with all the help and facilities, which I required for the completion of my thesis.

At last but not the least I would like to thank God and my parents for not letting me down at the time of crisis and showing me the silver lining in the dark clouds.

Murli Sharma

Murli Manohar Sharma
Roll No.- 801033009

Abstract

In the consent of this internet world, nobody is untouched with the internet for their usage. For such kind of scenario, data mining becomes an essential part of computer science. Data mining is a sub-field, which computationally processes the data, collected and is able to help the analysts of the research and development department, and the scientists, for proposing the ideas for some betterment of the organization. The user access is recorded in log files. The information regarding the website traversal of a user is always tracked by using these log files. These log files are generally stored at the server or at the client side. The web server logs provide important information. In the field of web mining the analysis of the web logs is done to identify the users' search patterns. In the existing approaches of finding the patterns, tree have been created which is based on the frequent access pattern identification. The creation of tree has increased the overhead of web usage. Therefore, Single Scan Pattern Algorithm has been proposed, which is based on use of database scan without creating any tree. The proposed algorithm would be able to increase the efficiency and decrease the overhead of unnecessary database scanning.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

In this chapter the concepts of data-mining and web-mining is discussed. The subtasks and the challenges are the important portion of this section.

Data mining is a field of computer science which commingles various traditional data analysis methods with the sophisticated algorithms to process the large amount of data. There is a rapid advancement in the data collection and the data storage technologies to cumulate such kind of vast data. To extract a useful information form a greater amount of data is not an easy task. The traditional data analysis tools [1] and techniques are not able to serve the purpose, because of the large scaled data.

The question may arise that why the data now a days, is becoming larger than before. One of the basic reasons behind this is to increase the customer relation quality. In business, data collection is done by various ways like bar code scanners, smart cards etc. This allows retailers to collect as much as information they can find and analyze this data to enhance the business decisions. The retailers can utilize this information, along with the other business critical data. This data may include web logs from e-commerce web site and the customer service records from call centers, to help them better understand the requirements of their customers and make the business more effective and customer satisfactory.

The data mining techniques [2] can be used to support a huge variety of business applications like consumer profiling, targeted marketing, work flow management, store layout and fraud expectations. Such kind of data analysis may also provide them the names of the customers who may be beneficial for their business.

1.1 Introduction to Data Mining

Data mining is all about the collecting data and then finding and extracting the useful information from such kind of large amount of data chunks. The data-mining techniques are used to purge large amount of data to find the useful and interesting patterns [3] about the user and the business. Some of the techniques are capable of predicting the outcome of the current schemes such as whether the new customer will be willing to spend more money in the services provided by the company or not.

Data-mining is an indispensable part of knowledge-discovery. Knowledge discovery in database, is the overall process to convert the raw data into some useful information. This process consists of some steps. In these steps the input data is stored in the form of a variety of formats and may reside in a centralized manner. The related steps are given in the Figure 1.1.

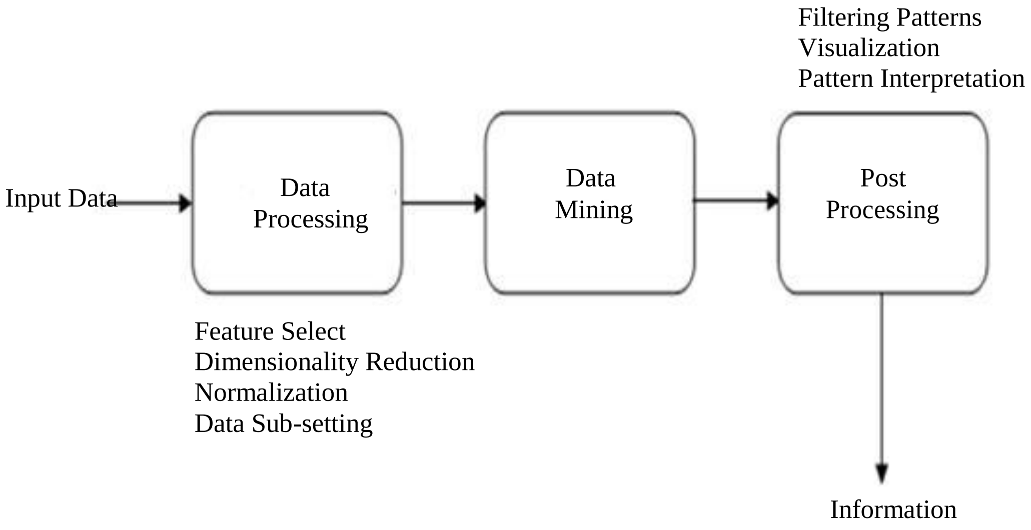


Figure 1.1 The process of knowledge discovery in databases [4]

Here the data is residing in the centralized data repository to be distributed in the multiple sites. Now the data is preprocessed. The objective of data preprocessing is to transform the raw data into an appropriate format. This helps in further analysis. There are many steps involved in the steps such as fusing data from multiple sources, cleaning data to remove noise and duplicate observations, select the records and the features that are relevant to the data mining task. Since many ways are available to collect the data and the data storage can be done in many other ways.

So to make similarities in the data collected by these ways data preprocessing is a necessary step. After this, here come the data mining steps, which is the most important objective of this report. Data mining employs some algorithms to find out the user usage patterns, algorithms for helping business decision making and others. In the business applications, the information offered by data mining results need to be integrated with the other management tools. This integration is needed for effective marketing promotions. So this is an important goal and to achieve this data postprocessing step is involved. This process ensures that only the valid and the useful results are obtained by the decision support system.

1.1.1 Challenges in data mining

The field of data analysis is posed by some challenges [5, 6]. These are the challenges which motivated the field of data mining. These challenges are as follows.

- **Scalability**

In the field of data analysis, data generation and collection has grown the size of data up to terabytes or even petabytes. Every algorithm which is handles such kind of massive data must also be capable of handling scalability. These algorithms should also involve a special data structure to access individual records in an efficient way.

- **Heterogeneous and Complex Data**

The role of data mining in business, science, medicine and others has grown. Because of involvement of such kind of variety of data, there is a need for the techniques to handle the heterogeneous data. In the recent years there have been the emergence of the complex data e.g. collection of the web pages that contain semi-structured text and the hyperlinks etc. Techniques are developed to mine such kind of complex data objects.

- **Data ownership and distribution**

Often it happens that the data to be analyzed, is stored in scattered locations and owned by various organized. Due to this feature there is a requirement of developing distributed data mining. But the distributed algorithms also face some hindrances. The amount of communication needed, depends on the hardware. To effectively consolidate the data mining results obtained from multiple sources is not an easy task.

- **Non-traditional Analysis**

In the traditional statistical approach, a hypothesis is proposed, then a test is designed and then the data is collected and analyzed according to the hypotheses. But this process needs so much effort. The current scenario involves many tasks which often require generation and evaluation of many other hypotheses. The requirements of automation of hypothesis creation and evaluations have motivated some data mining techniques.

The researchers of various disciplines began to focus on developing efficient and robust tools and methods to overcome these challenges. The work done by these researchers is totally based upon the methodology, tools and algorithms that were devised earlier by the researchers. Specifically data mining draws upon the ideas of sampling, estimation and hypothesis testing using statistics, search method and algorithms, modeling techniques, learning theories from artificial intelligence, pattern recognition and machine learning. Data mining is also quick in adopting ideas from other areas like information retrieval, optimization, information theory etc. [6].

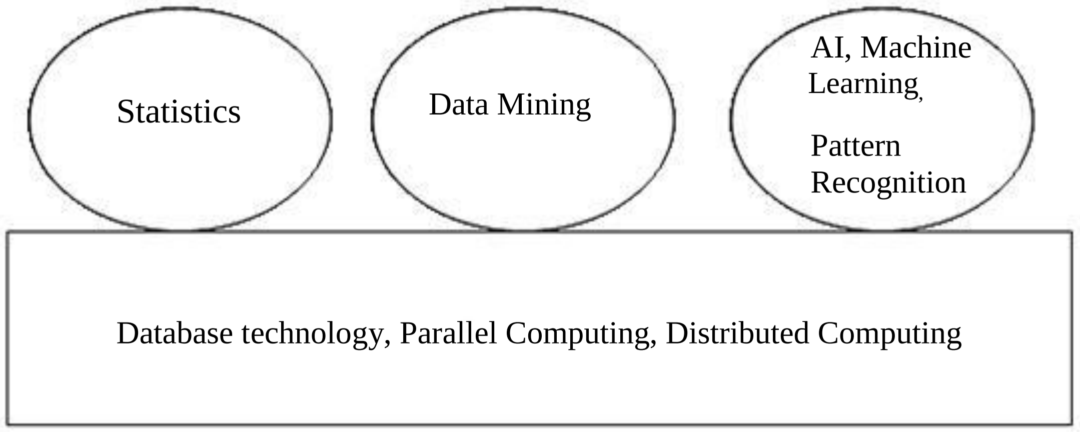

Statistics
Data Mining
AI, Machine Learning,
Pattern Recognition
Database technology, Parallel Computing, Distributed Computing

Figure 1.2 Influences of various other fields on data mining [7]

Various other areas are also affected by the field of data mining as shown in Figure 1.2. These fields support data mining for fulfilling the purpose of collecting and extracting useful information from the large scaled databases.

Data-mining gets an important place in the today's world. This has become an important research area since the amount of data available in most of the applications has become higher in size. This great amount of data should be procesed to extract the useful information and knowledge, since they are not explicit. The definition of data mining is given by Usama*et al.* [7] as "*Data Mining is the process of discoveringinteresting knowledge from large amount of data*".

1.1.2 Data mining sub-tasks

Data-mining itself involves six common but the important classes of sub-tasks [8].

These sub-tasks are as follows.

- **Anomaly detection**

The extraction of unusual data records which seem to be interesting or the data errors and requires further investigation in anomaly detection. The observations whose characteristics are significantly different from the usual collected data, are analyzed in the anomaly detection and such kind of observations are called as anomalies. The aim of this anomaly detection and study is to identify the real anomaly and avert wrong labeling of normal objects. Higher detection rate, greater fault tolerance and low false positives as well as false negatives are the properties of a good anomaly detector. These anomaly detectors are applied in network intrusion detections, fraud detection, ecosystem changes' study or many other varieties of areas.

- **Association rule learning**

This is task, whose purpose is to identify the patterns. These patterns are tightly associated with the features of the data. It Searches for relationships between variables. The patterns identified can be represented in the form of implication rules. The main motive of association rule analysis is to extract the most interesting patterns. For this purpose, different-different data structures are employed in the implementation phase, the database tables are needed to be scanned many a times. For example, a supermarket may gather data on the purchasing habits of each customer. In this way the data may become very large. It becomes difficult to fulfill the purpose of analyzing the customer habits. By using the association rule learning, the supermarket becomes able to decide that which products are frequently bought or searched together. The supermarket may use this information for marketing and sales purposes. This is sometimes termed as market basket-analysis.

- **Clustering**

In early batch processing systems the data of the similar kind was bunched together and processing was done in such a way that the processes which are of similar kind were processed together. Similarly, in the field of data mining, the data of similar kind are grouped together. This grouping is called as clustering. In the field of business and data analysis clustering is employed so that the observations that belong to the same clustering are more similar to each other than the observations that belong to the other clusters. It is the effort of discovering the groups and structures in the data that are in some way similar without using known structures in the data.

- **Classification**

Classification is the duty of data mining which assigns the items to a set of target categories. The motive behind the categorization is to precisely predict the target class, for each case in the data. For example, a classification model is used to identify the loan applicants as low, medium.

A categorization task starts with the data set in which the class assignments are known beforehand. For example, a classification model whichcan predict credit risk, can be developed based upon the observed data for many loan applicants over a period of time. In addition to the previous transactions, the data may track the employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit ratingsshould be the targat, the other attributes are the predictors and the data for each of the customer will constitute a case.

So the classification is the task of generalizing the known structures to apply to a new data. For example, an e-mail program may be attempted to classify it as legitimate or as spam.

- **Regression**

Regression is a function which moulds the data without high errors. Regression is a data-mining function that predicts a number, age, weight, distance, temperature, income or sales. For example, a regression model can be used to predict the children's height, where their age, weight, and other factors are given.

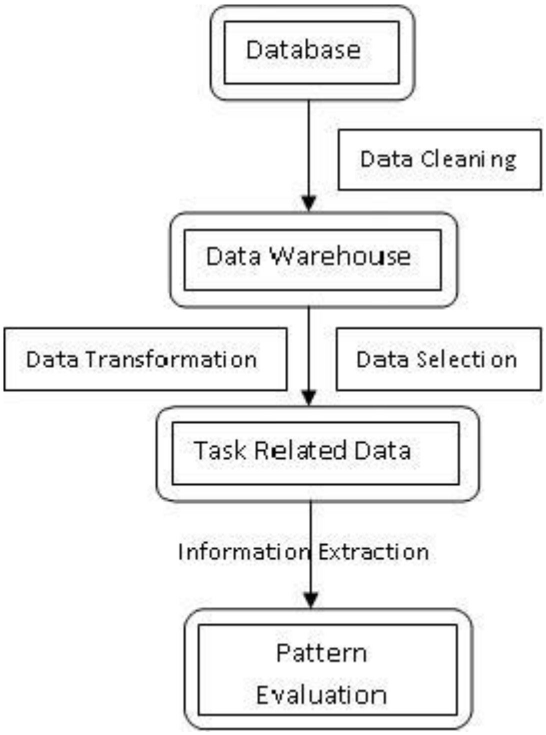
A regression task starts with a set of data, in which the target values are known. For example, a regression models which predict the children's height can be developed based on the observed data for more children over a period of time. The data might track the age, height, weight, development, heredity, and so on. Height would be the target while the other attributes would be the predictors, and the data for each child will constitute a case.

- **Summarization**

Summarization is a key concept for the data mining. This concept involves the techniques for finding a compact, precise and accurate description of the data set to be analyzed in data mining field. It provides a more precise representation of the data-set which includes the visualization and report generation. Summarization can be viewed as a compressing technique that takes a given set of transactions as an input and represents this data set into a smaller set of patterns while extracting the maximum possible information.

1.2 Data Mining Process

The process of data mining comprises of some steps starting from raw data collections to some form of new knowledge. This is an iterative process and consists of the steps shown in Figure 1.3.



1

Figure 1.3 Phases of Data-Mining [5]

First of all the data is taken from the database. After this the data is shifted to the data mining process. The further processes are done on the basis of the flow chart given above.

- **Data Cleaning**

In this phase, the missing and redundant data is identified. The real world data may be incompleted, inconsistent and corrupted. So the current phase deals with the detection and removal of the errors and redundancies from the data, to improve the quality of the data. The problemof data quality is present during the data collections, such as files and databases, e.g., may be because of misspellings during data entry, missing the information or other invalid data. In this process, missing values can be entered or removed, noise values can be smoothened, outliers are identified and each of these deficiencies are handed by different other techeniques.

- **Data selection**

In the current step, the relevant data for the analysis is decided. This is done on the basis of the requirement and the type of the data to be analyzed. After this the data is obtained from the data collected in the data warehouse.

- **Data transformation**

The process of data transformation, converts the source data into proper format to mine the data. Data transformation involves the basic data management sub-tasks,like smoothening, aggregating, generalizing, normalization and attributes construction. It is also known as data consolidation. To assure the transformations, this phase may employ data mapping, code generation etc.

- **Data integration**

The designers make create the data based on their needs, in dome different format every time. So this data needs some kind of integration. This process combines the data from different other sources. The source data can be a set of multiple databases, which may have different data definitions. In this case, data integration process inserts the data into a single-coherent data store, from these multiple sources of data.

1.2.1 Methods for pattern extraction

In the Data-mining processes, data intelligent methods are applied to extract the data patterns. These methods are as follows [10].

- **Pattern evaluation**

It is the task of discovering interesting patterns among the extracted data-set. For this purpose log files are to be studied. The log file contains the user traversal paths and other useful information.

- **Knowledge representations**

This is the last phase in which the discovered knowledge is visually represented to the user. This essential step uses some visualization techniques to help the users to understand and interpret the data-mining results. Knowledge representation includes techniques for the visualization which are used to understand the discovered knowledge to the user. Thisrepresentation is done generally in the form of tree.

World Wide Web is one of the largest and most widely known sources of data. All the data is residing on the www domain and all the data is being updated by many users every day. This total size of the whole documents can grow up to terabytes. The documents on www are distributed over millions of computers that are connected with each other by telephone lines, optical fibers, radio modems and so many other technologies. World Wide Web is growing at a very higher rate in size of the traffic, the amount of the documents and the complexity level of the web sites. Due to this trend, the demand for extracting valuable information from this huge amount of data source is increasing every day. This leads to new area called web mining which was first coined by Etzioni [11], which is the implementation of the data-mining techeniques to World Wide Web. The next section explains general overview of web-mining [12, 13].

1.3 Web Mining

Web data-mining is an important sub-field of data-mining, which deals with the extraction of interesting information from the world wide web. It was Oren Etzioni [11] who first proposed the terminologyof web-mining, in 1996 in his paper. In this paper, it was claimed that web-mining is the use of data-mining techeniques to collect and extract the information from World Wide Web documents and services automatically."Whether effective web mining is feasible in practice or not?", was the question which was coined by the author. Web-mining is the area of research, which is so much big today, due to the tremendous growth of data available on the web and the recent interest in e-commerce field. Web-mining is used to understand the customer behaviour, evaluate the effectiveness of a particular web site, and help for the success of a marketing campaign and various other decisions making for the betterment of the business.There is one simple definition of web mining [14], which is "*Web Data Mining is theapplication of data mining techniques, to extract interesting and potentially useful knowledge from web data. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process."*

1.3.1 Web Mining Challenges

- **To find relevant information**

The users involve internet, in their searches and for gaining knowledge. However today's searching methods have problems like low precision which is due to the irrelevance of many of the searching results. These problems result in the difficulty to find the relevant information. Another problem is low recall which is due to inability to index all the information available on the web [6].

- **Getting new information from the data available on the web**

  This problem is fundamentally a sub-problem of the first one. Above problem is query triggered process but this problem is a process which is triggered data. In this problem, the useful information is collected from the data stored.

- **To learn consumers' needs**

  This problem is all about that what a user needs and searches often. Inside this problem there are sub problem such as customization of the information to the intended group of consumers. This problem is more related to the web site design management and marketing [15].

1.3.2 Web Mining Subtasks

Web mining sub-tasks [16] are explained as follows.

- **Resource finding**

  This is the task of retrieving the desired web documents. The word resource finding means, the process of retriving the data which is either online or offline, from the sources available on the web, such as electronic newsletters, the text contents of HTML documents obtained by removing HTML tags and also the manual selection of web resources.

- **Information selection and pre-processing**

  Automatically selecting and pre-processing specific information from retrieved web resources, is the heart of this phase. It is like a transformation process of the original data retrieved in the IR process into a useful data-set. These transformations can be pre-processed such as stop words, stemming or a pre-processing aimed at obtaining the desired representation such as finding phrases in the training corpus, transforming the representation to relational or first order logic form.

- **Generalization**

It is the process which automatically discovers patterns of individual web sites as well as multiple sites. Machine learning or data mining techniques are used in the process of generalization.

1.4 Organization of thesis report

This document gives some introduction about the era of data mining in the introduction portion. In the section of literature survey, various other developments and the researches are discussed. This phase of the thesis tells about the importance of the data mining, web mining and the web usage mining, by giving the overview of the background of web mining, the techniques used for information retrieval. Web usage mining is the main focus of this thesis work. Here the impact of data mining in the other fields can also be seen in a very comprehend way. In third chapter, the related algorithms have been discussed. These algorithms are the base of the thesis work. In the final chapter the comparative analysis of the proposed work with the existing algorithms, is done.

Chapter 2
Literature Survey

This chapter discusses about the importance of the web mining field by describing the web mining taxonomy and it's applications. Data-mining is the analysis portion of the "Knowledge Discovery in Databases" process, or KDD, [8]. This is a modern field of computer science [17] and is the process that results in the discovery of new patterns from a large data set. The main objective of the data-mining process is to extract the information from an existing data set and transform it into a usable form for further use.

The real data-mining task is to automatically analyze the large amount of data to extract the previously unknown patterns such as groups of data items (cluster analysis), detecting unusual records (anomaly detection) and dependencies (association rule mining). These patterns can then be seen as a kind of summary of the input data and may be used for further analysis and predictive analysis.

For example, the data mining step may prompt us to identify various other multiple groups in the data which can then be used to obtain more accurate prediction results from a business decision support system.

2.1 Relation of Web Mining with Other Fields

In the recent years, the growth in number of web sites and web users to these web sites has increased with an exponential rate. Due to this much growth, a huge amount of online data is being generated. To mine the interesting information from such kind of huge pool of data, the data mining techeniques may be applied. This web data is unstructured and not arranged in a useful manner. So, we cannot apply the data-mining techeniques immediately on this data. Web-mining is implemented to get the interesting patterns, which can be implemented further in many other real world problems like improving web site structure, to understanding the visitor's requirement in a better way and the product recommendations and get the business enhancement.

In 1996, Etzioni [11] in his research paper stated that the web mining rapidly has become a new exploring field which is expanding each moment. In this paper thequestion was brought out that: “Is it practical to mine Web data?” He also suggested that web-mining can be divided into three processes. There is an increasing number of researchers are who are working in this field and they are doing some surveys around the data-mining on the web. Initially the web-mining was clearly divided into three categories: web-content mining, web-structure mining and web-usage mining. There have been researches around the content mining performance improvement and the web-structure mining based on the research of data mining and information-retrieval, information-extraction and artificial intelligence. The rapid extension of the web is creating a constant growth of the information, which leads to several other problems.

According to Kantardzic*et al.* [18], Web-content mining is all about searching the information resources automatically, which are available online. Web structure mining means mining the web documents’ structure and links, in short, mining the web structure data. Web-usage mining involves the data from the server access logs, user registration and their profiles, user sessions or transactions.

Web mining subtasks [19] are as follows.

(a) Finding the resources that can give us the raw data

(b) Information selection and pre-processing

(c) Patterns inspection

(d) Verification and interpretation

(e) Visualization

Web-mining is often related with the IR or IE. The explanation to this statement is given by Xindong*et al.*[20]. The overview of this relation is given below.

- **Web-mining and Information-Retrieval (IR)**

For the automatic retrieval of relevant and important documents, information retrieval [21] is used. According to some researchers, the resources identification and the document collection on the web are under the area of web content mining while some of them co-relate the web mining with the intelligent IR. The primary goal of IR is searching and indexing text for useful document. Research in IR includes document categorization and classification modeling, data visualization, user interfaces and filtering [6]. The task of web document classification or sub-fieldassignment can be used for indexing. This can be considered as an example of web mining.

- **Web-Mining and Information-Extraction (IE)**

The aim of IE is to transform a collection of documents. This can be achieved usually implementing the IR-system into the information that is immediately retrived and analyzed. IE [23] aims on the extraction of the relevant facts from the documents while opposite to it IR's motive is to select the relevant documents. IE's area of interest are the layout or representation of a document, while IR thinks of the text in a document is just a collection of unordered and mashed words. Thus, we can say that IE works at a finer granularity level while IR concentrates on the documents. To build IE-systems manual, is not a feasible and scalable task for the dynamic and diverse web contents. Due to this nature of the web most of the IE systems focus on specific websites to extract the information. Machine learning or the data-mining techniques can be used to learn how to extract the patterns or how to create the rules for web documents automatically semi-automatically.

2.2 Web Mining Taxonomy

Web mining is considered to have three subfields [24]

- Web content mining
- Web structure mining
- Web usage mining

These fields are discussed in Figure 2.1. These fields are having their own importance according to the need of the implementation.

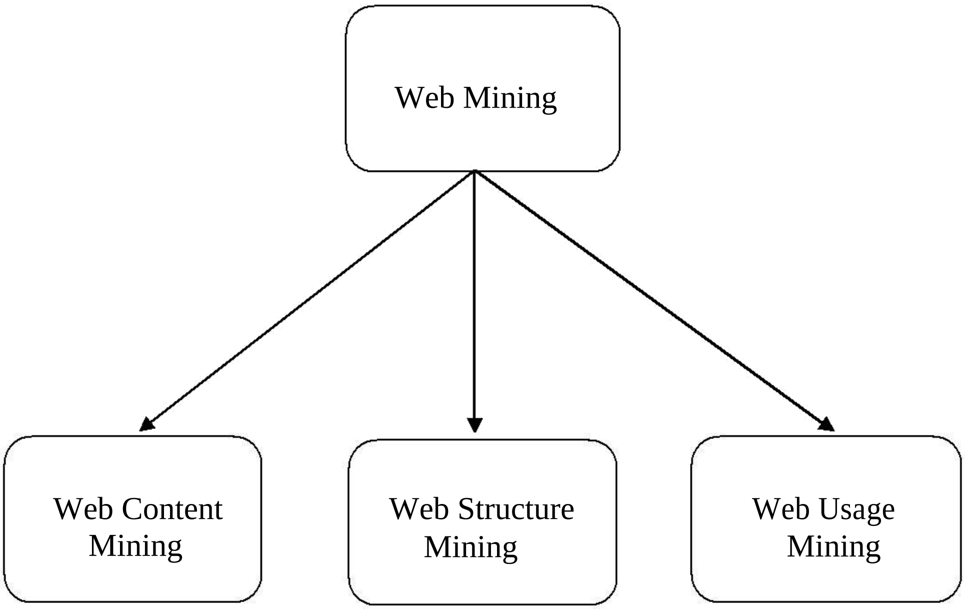
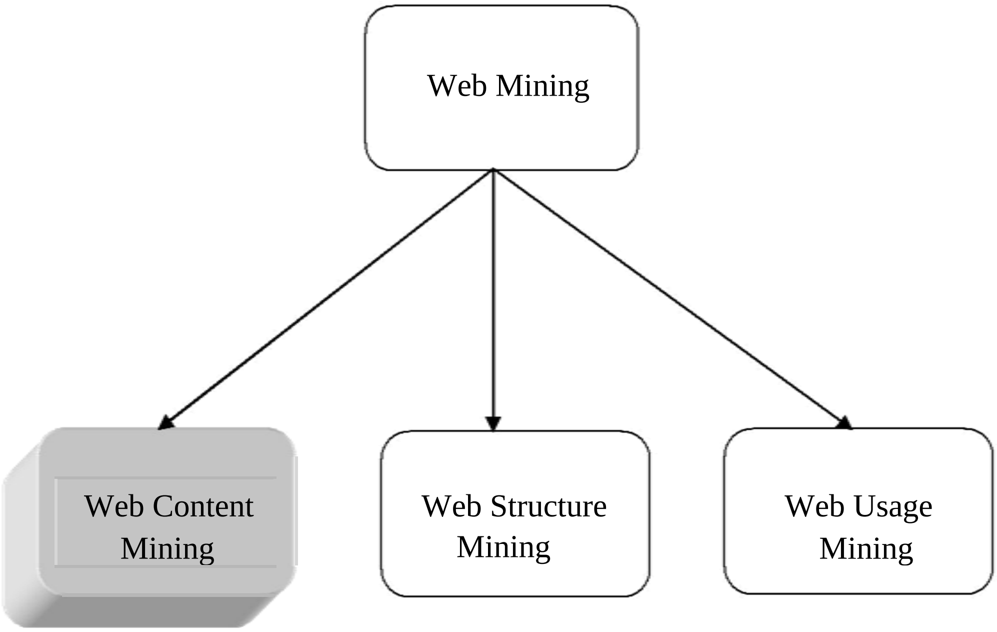


Figure 2.1 Web Mining Taxonomy [24]

These subfields of Web Mining are described as follows.

2.2.1 Web Content Mining

Web content mining describe the automatic searching of data resources which are available online [25]. This includes mining the web data contents as shown in Figure 2.2. The web documents are usually comprised of several types of data such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents some arestructured data like the data in the tables or database generated HTML pages, but most of themare unstructured text data. The unstructured phenomenon of web data, forces the web-content mining towards a more complicated approach. The web-content mining is differentiated with the two different aspects: Information Retrieval View and Database View.



- Summarization of web pages
- Summarization of web searches
- Mining multimedia web contents
- Web pages classification

Figure 2.2 Web Content Mining [25, 26]

Web content mining describes the discovery of useful information from the web contents or data or documents [15]. However, what consist of the web contents could encompass a very broad range of data. This section begins by reviewing some of the important problems that web content mining aims to solve. After that enlisting of the different approaches in this respective field are classified depending on the different types of Web content data. In each approach listing of the most used techniques is being done. It is often said that the web offers an unprecedented opportunity and challenge for data mining. This is due to the following characteristics of the web environment [27].

- The amount of data or information on the web is huge and still growing exponentially.
- The information coverage of web is very wide and diverse. One can find the information about almost everything on the web.

- The data of any type exist on the web e.g. structured tables, text and multimedia data.
- Heterogeneous information are available even for the similar topic. Multiple web pages may present the similar information using completely different formats.
- Most of the information over the web is semi-structured due to the nested structure of HTML script. According to a designer there is a need of web page to present information in a simple and regular fashion.
- The information on the web are linked. There are links among pages within a site and across different sites. These are the hyperlinks.
- A web page typically contains a mixture of many kinds of information e.g. main content, advertisements, buttons etc. For a particular application only part of the information is useful and the rest are noises.
- The web is also about services. Many web sites and pages enable people to perform operations with input parameters i.e. they provide services.
- Above all the web is a virtual society. It is not only about data, information and services but also about interactions among people, organizations and automatic systems.
- The web is dynamic. Information on the web changes constantly. Keeping up with the changes and monitoring the changes are important issues for many applications.

The web content data is collection of unstructured data such as free text, semi-structured data such as HTML documents and a more structured data such as data inthe tables or the database generated HTML pages. So following two main approaches in web content mining arise. These are as follows.

- Unstructured text mining approach
- Semi-Structured and Structured mining approach.

- **Unstructured Text Data Mining (Text Mining)**

Much of the web content data is unstructured text data [11]. The research around applying data mining techniques to unstructured text is termed knowledge discovery in texts (KDT) [28] or text data mining [29] or text mining [30]. Hence we could consider text mining as an instance of web content mining. Text mining or KDT was first proposed by Feldman and

Dagan. They suggested structuring the text documents by means of information extraction, text categorization or applying NLP techniques as pre-processing step before performing any kind of KDTs. The reason is mining on the unprepared documents does not provide effectively exploitable results [31].

- **Semi-Structured and Structured data mining**

This is perhaps the most widely studied research topic of web content mining. One of the reasons for its importance and popularity is that structured data on the web are often very important as they represent their host pages. Structured data [32] is also easier to extract compared to unstructured texts. Semi-structured data is a point of convergence for the web and database communities: the former deals with documents, the latter with data. Emergent representations for semi-structured data (such as XML) are variations on the Object Exchange Model (OEM) [33]. In OEM data is in the form of atomic or compound objects. Atomic objects may be integers or strings; compound objects refer to other objects through labeled edges. HTML is a special case of such intra-document structure.

One can differentiate the research done in web content mining for semi-structured and structured data from two different points of view: IR and DB views. The goal of web content mining from the IR view is mainly to assist or to improve the information finding or filtering the information to the users usually based on either inferred or solicited user profiles while the goal of web content mining from the DB view mainly tries to model sophisticated queries other than the keywords based search that could be performed [15].

2.2.2 Web Structure Mining

Most of the web information retrieval tools only use the textual information while ignoring the link information that could be very valuable. The goal of web structure mining is to generate structural summary about the web site and web page.

Technically, web content mining focuses on the structure of the inner-document, while web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, web structure mining will categorize the web pages and generate the information such as the similarity and relationship between different web sites and the web portals. Web structure mining can also have another direction which is discovering the structure of web document itself. This type of structure mining can be used to reveal the structure of web pages, as shown in Figure 2.3, this would be good for navigation purpose and make it possible to compare or integrate web page schemes [34].

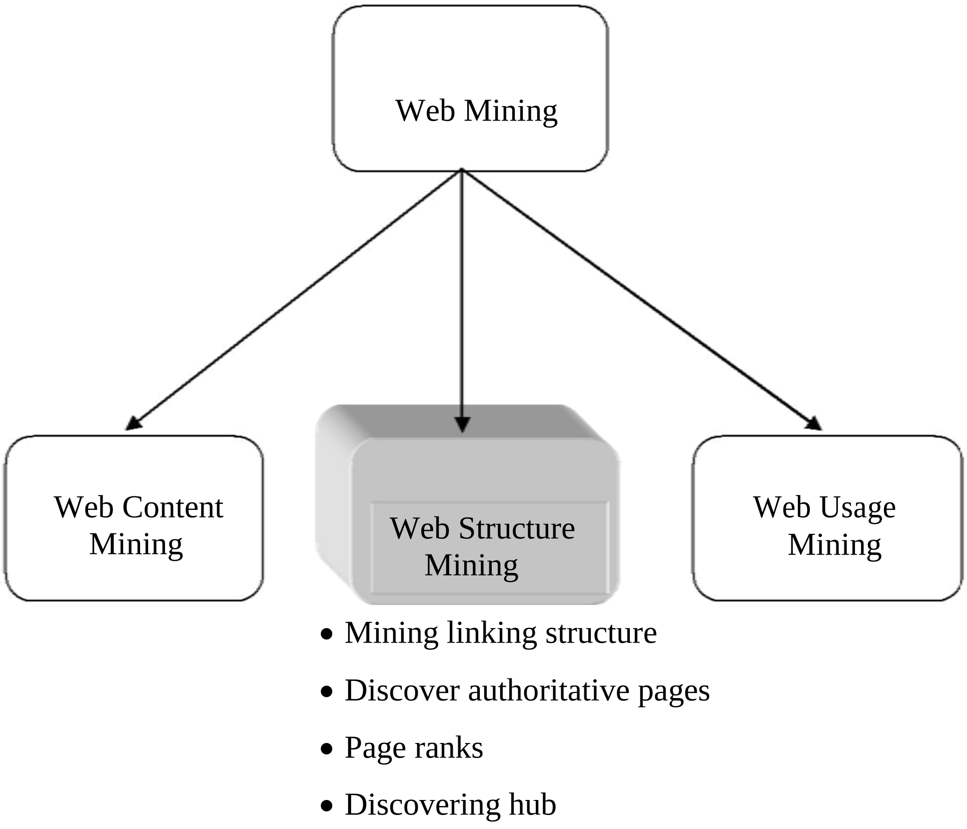


Figure 2.3 Web Structure Mining [26, 34]

Another task of web structure mining is to discover the nature of the hierarchy or the structure of the network of hyperlinks in the web sites of some particular domain. This may be helpful to generalize the flow of information in the web sites that may represent some particular domain so that the query processing may become easier and more efficient. Web structure mining [35] is indispensably related with the web content mining since it is very likely expected that the web documents contain links and they both use the real-world and raw data on the web. It is quite often to combine these two mining tasks in an application.

Web structure mining [36] is the process of discovering the structure of hyperlinks within the web. Web structure mining discovers the link structures at the inter-document level with a focus on the hyperlink structure of the web. The different objects are linked in some way. The objects in the WWW are web pages. Attributes include HTML tags, word appearances and anchor text [37]. The appropriate handling of the links can result in the potential correlations and help to improve the predictive accuracy of the learned models. The challenge for web structure mining is to dealwith the structure of the hyperlinks within the web site itself. Two algorithms that have been proposed to lead with those potential correlations are HITS [38] and Page-Rank [39]. These are as follows.

- **Hyperlink-induced topic search (HITS)**

In this concept, two kinds of pages are identified from the web hyperlink structure authorities like the pages with the good sources of content and hubs i.e. the pages with good sources of links. A good hub is a page that point outs to many other good authorities [38].

- **Page-Rank**

Page rank idea is all about counting the in-links [39]. These are equally normalized and extended by using the number of links within a web-page. The Page Rank algorithm is defined in [40] as: Assuming page A has pages T1...Tn, which point to it. The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. Also C (A) is defined as the number of links going out of page A.

The Page Rank of a page A is given as follows:

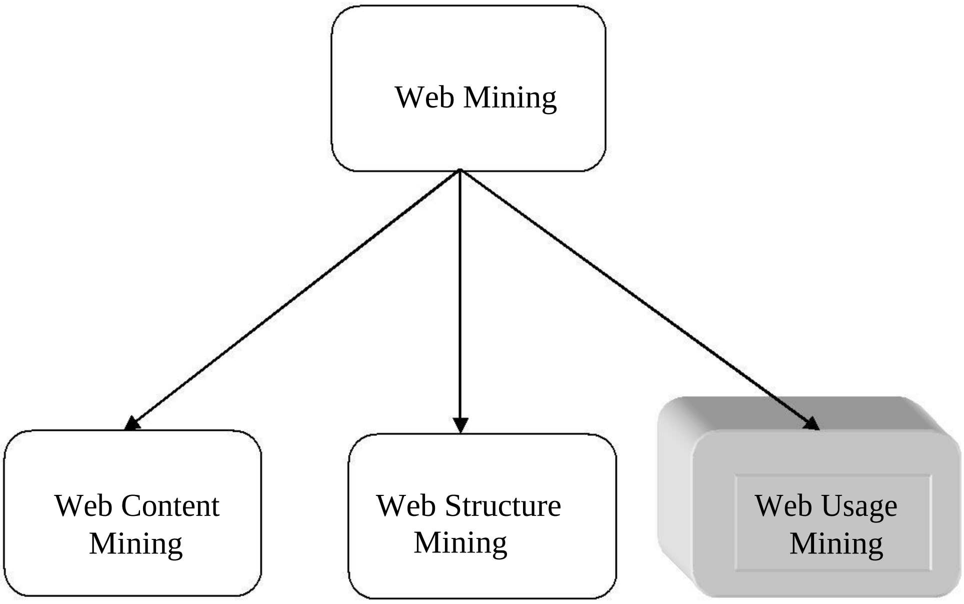
$$PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn))$$

Note that the Page Ranks form a probability distribution over web pages so the sum of all web pages Page Ranks will be one. And the damping factor d is the probability at each page.

2.2.3 Web Usage Mining

Web usage mining [41] is the application of data mining techniques, to discover the interesting usage patterns from web data, to understand and better serve the needs of web-based applications and the web users. It tries to make sense of the data generated by the web surfer’s sessions.

While the web content and structure mining utilize the primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web. The web usage data includes the data from web server logs, proxy server logs, browser logs, and user profiles. The usage data can also be split into three different kinds on the basis of the source of its collection: on the server side there is an aggregate picture of the usage of a service by all users, the client side, while on the client side there is complete picture of usage of allservices by a particular client, and the proxy side with the proxy side being somewhere in the middle. Web usage mining analyzes results of user interactions with a web server, including web logs; click streams and database transactions at a web site [14]. Some properties of web content miming is given in Figure 2.4.



- Mining web logs to discover usage pattern.
- Application-
  - ➢ Personalization of web content
  - ➢ Improving web design

Figure 2.4 Web Usage Mining [26, 41]

2.2.3.1 Web Usage Mining Process

Web usage mining process can be regarded as a three-phase process consisting [14]

- **Preprocessing or Data Preparation**

The data of web log files is preprocessed to clean the data which implies removing log entries that are not needed for the mining process, data integration and identification of the users and to improve the session making task.

- **Pattern Discovery**

Statistical methods as well as data mining methods such as path analysis, association rule, anomaly detection, sequential patterns, clustering and classification rules are applied to detect various interesting patterns or the user behavior.

- **Pattern analysis**

Patterns discovered in the previous section are analyzed here, using OLAP tools, knowledge query management mechanisms and intelligent agent to filter out the irrelevant rules or patterns.

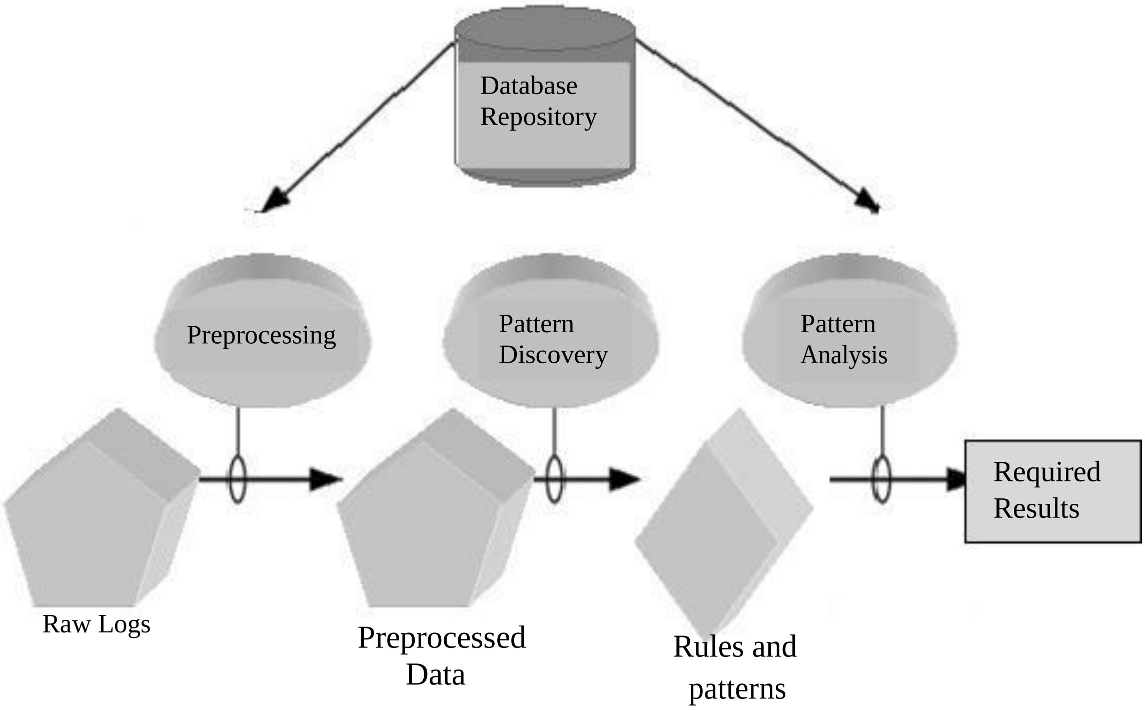


Figure 2.5 Web Usage Mining Process[14, 42]

After discovering the patterns from data usage, an analysis needs to be done. The most common ways of analyzing these patterns are: to use the query or to load the results onto a data cube and after that perform the OLAP operations. After these steps various visualization techniques are used for the interpretation to get the necessary information. The rules employed and the identified patterns needs to be used for the improvement of the system performance and in employing the upgradation to the web site.

Applying the statistical techniques and other data-mining techniques to these pre-processed website-log data to discover the useful patterns, are the two purposes of web usage mining. Data-usage mining tools are able to discover and anticipate the user behaviour. This will help the designers to upgrde the web-site architecture, to attract the visitors or to give the steady users a personalized and flexible services according to their needs. This web-usage mining process is shown in Figure 2.5.

Web-usage mining is the task of getting the information about the activities of the users during browsing session with a website and navigating through-out the web [42]. In such cases, the log files can be considered as a secondary data, while the the secondary data is the web pages' contents which are available for access.

Five most prominent steps to be followed in the web-usage mining are.

- **Data collection**– Web-log files are needed to be collected, which are able to keep track of visits of aparticular visitor.
- **Data Integration**– Multiple log-files are integrated into one single file
- **Data pre-processing**– Cleaning and re-construction of the data is done.
- **Pattern-extraction**– Interesting patternsare identified.
- **Pattern analysis and visualization**– Scrutinize the patterns extracted
- **Pattern applications**–The pattern can be applied in the real-world problems**.**

WWW Log files can be residingat the server-side, at the client-side orat the proxy-servers. It is difficult to collect all the metadata from the client-side. Most of the algorithms' work,is depended only on the server-side data. Some commonly used data-mining algorithms are the association-rule mining, sequence-mining and clustering [42]. Web-usage mining is one of the prominent research areas due to following reasons.

a) The web pages accessed before by a user, can be kept into consideration. These web-pages are then further used to scrutinize the typical behavior of the web-user and to comment about the mostly desired pages beforehand.

b) Frequent-access behaviour shown by the users, can be cast-on to identify the needed links to improve the overall performance of the future visits. Pre-fetching and caching policies can be adopted based on these results.

c) Common-access behaviours of the users can be used to improve the actual design of web-pages.

d) Usage patterns can be employed for the business intelligence. In other words, the improvement in the sales and advertisement can be done by providing product recommendations.

2.3 Web Usage Mining Applications

- **Pre-fetching and caching**

  The output delivered by the web-usage mining is useful for improving the overall efficiency of the web-servers and web based applications. Web usage mining [43] is useful for developing the proper pre-fetching and caching strategies. Therefore it is useful in reducing server response time.

- **Support to the design**

  The output delivered by web-usage mining techniques is useful for availing guidelines to improve the design of web applications. Some studies suggest that strato-grams [44] can be used to evaluate the efficiency of the web-sites from the user's perspective and some exploits the web-usage mining techniques to propose the desired changes in the web sites. Data mined from the user's behavior can be used to dynamically restructure the contents and the layout of the web-site.

- **E-commerce**

  Web based companies use data-mining techniques to mine the business intelligence which is important for the e-commerce of the company. Web usage mining techniques provides a fruitful advantage to Customer Relationship Management (CRM) [45]. The main goal is for business-specific issues such as: alluring customers, customer confinement, cross-sales and customer retrieval.

- **Personalization of web content**

  Web usage mining techniques are useful for providing personalized web user experience. This can be done by anticipating the user-behavior in real-time by matching up the current users' pattern with the patterns which were mined from the previous web logs. Recommendation systems are the most common application in this field. Recommendation systems are useful in recommending the fruitful links to the products which may be interesting to users. An example for the recommendationsystems are Personalized Site Maps are for the links proposed an dynamic technique to restructure the product catalogue in accordance to the extract user-profile [23].

2.4 Apriori Algorithm

In data mining, apriori algorithm is an algorithm to learn the association-rules. Apriori is designed to operate on the databases which containsthe transactions. Other algorithms are designed for finding association rules in data which do not haveany transaction.

It is usual forassociation-rule mining in a given set of items that the algorithm attempts to find subsets, which are the intersection to a minimum number of the item-sets. Apriori [52] uses a "bottom up" approach, where the frequent items are continued with one item at a time and a candidate’s group are tested against this data. This algorithm gets terminated when there is no further successful extensions. Apriori uses the breadth-first search and a hash tree structure to count candidate item-sets.

This algorithm creates the candidate item-sets of length “a”, from the item-sets of length “a−1”. Then it trims the candidates which have an in-frequent sub-patterns. According to the “downward closure lemma”, the candidate-set have all the frequent “a”-length item-sets. After that, it scans the transaction database to determine frequent item sets among the candidates. For determining frequent items quickly, the algorithm uses a hash tree to store candidate item sets. Apriori, while having historical significance, suffers from a number of deficiencies or the trade-offs, which have engendered other algorithms.

Apriori is a powerful algorithm for mining frequent item sets for Boolean association-rules. The name of this algorithm is based on the fact that the algorithm uses pre-knowledge of the frequent item-sets’properties. Apriori employs an iterative approach known as a level-wise search, where k-item sets are used to explore (a + 1)-item sets. 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, until no more frequent k-item sets can be found. The finding of each L_a requires one full scan of the database.

In all these the very first step is to find out the maximum forward reference. When a user traverses a website, he or she can go to the hyperlinks which are provided in that particular page. The user clicks on these hyperlinks and move forward. So generally the paths are in the forward manner. But sometimes these paths may be reversed. In other words the user can go to the backward direction and go to another node. In this situation the path which has the maximum length in the forward direction are taken into account.

After calculating the maximal forward references, now there is need to scan the database tables.

Many algorithms have been proposed by the researchers. Some of them are as follows.

- Apriori Algorithms
    - Full Scan
    - Selective scan
- FP-Growth Algorithm
- Reference Scan

Some of them are explained in the following section.7

2.4.1 Full Scan

This is a scanning method that scans the database table of the web usage logs. These web logs are collected either from the server or from the client side. The users' website traversal sample is shown in Figure 2.6.

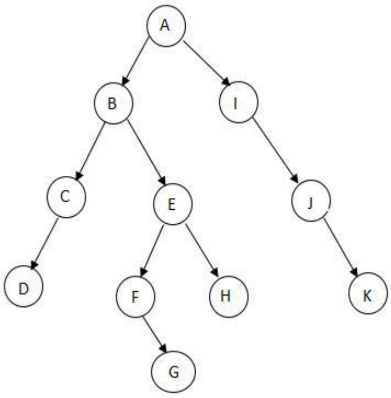


Figure 2.6 User traversal Pattern

For example the paths traversed by a user are given in the tabular form in Table 2.1.

Table 2.1 Paths Traversed by the user

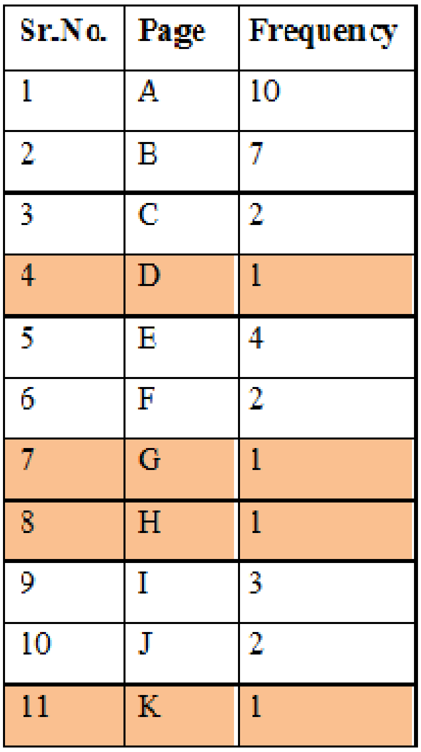
Sr. No	**Path**
1	ABCD
2	ABEFG
3	ABEH
4	AIJK
5	ABC
6	AI
7	AIJ
8	ABE
9	ABEF
10	AB

First of all find the support value. The support value is a value which shows that the pages which are below this number are less in use. In the other words this a threshold value to decide that this node can be avoided or not. This value is necessary because the data collected is very large and this data needs to be optimal in size, but to handle such kind of bigger data is very difficult. Now to find out the maximum forward reference, some levels of the refining are needed.

Level 1st

Frequency of the pages are calculated, which are referred by the user for his search. The frequency of the user to access any page is shown in the Table 2.2.

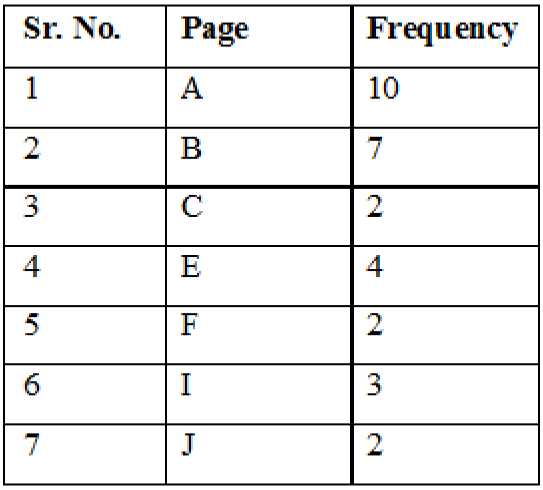
Table 2.2 Frequency of the paths



Sr.No.	Page	Frequency
1	A	10
2	B	7
3	C	2
4	D	1
5	E	4
6	F	2
7	G	1
8	H	1
9	I	3
10	J	2
11	K	1

Now, check that which node is having the frequency below the threshold support value. These item sets can be discarded for our further process. So, after applying this, the remaining data is shown in the Table 2.3.

Table 2.3 Frequencies after the minimization



Sr. No.	Page	Frequency
1	A	10
2	B	7
3	C	2
4	E	4
5	F	2
6	I	3
7	J	2

Level 2nd-

Now there is a need to make some combinations of these nodes as shown in the Table 2.4.

Table 2.4 Combinations at the second level

AB
AC
AE
AF
AI
AJ
BC
BE
BF
BI
……… ……

Further in the process of finding the patterns, these combinations are identified in the database provided in the initial phase. After this, the paths which are having the frequency below the support value are deleted out of the analysis. So by doing this comparison, the frequency table is obtained which is shown in Table 2.5.

Table 2.5 Remaining combinations after comparing with the support value

Path	Frequency
AB	7
AI	3
BC	2
BE	4

Level 3$^{\text{rd}}$-

Now there is a need to make the combinations again and then comparing them with the initial path table values. After repeating this process once again and removing the path values which are having the frequencies below the support value the obtained table has got the results as shown in the Table 2.6.

Table 2.6 Remaining combinations at the 3rd level

Path	Frequency
ABC	2
ABE	4

So finally there are two paths, which are the contenders of the frequently traversed paths. Among these two paths ABE is containing the higher frequency. So this path is traversed frequently by the user.

2.4.2 FP-Growth Algorithm

FP-Growth Algorithm is an efficient and scalable wayto mine the complete set of frequent-patterns, by using pattern fragment growth, by using an extended prefix-tree structure for storing suppressed and crucial information about the frequent-patterns named frequent-pattern tree (FP-tree). This algorithm allows the discovery of frequent item-set without any candidate itemset generation. This is a two-step approach. These steps are as follows.

Step 1- Create a compact data structure named as FP-Tree.

a) Scan the data and find the support for each item in the raw data table.

b) Calculate the minimum support value.

c) Infrequent items are discarded.

d) Sort the infrequent items in the decreasing order.

Step 2- Construct the FP-Tree.

a) Read the transaction.

b) Create a node in the tree according to their connections.

c) Increase the count of these nodes by one.

d) Take another transaction.

Step 3-Continue until al l the transactions are taken into consideration.

To calculate the minimum support value, the minimum percentage of support should be decided, say 20% and the no. of transaction are 10.

Minimum Support Value =10*(20/100)

= 2

After calculating this value, this algorithm is applied and a tree is obtained. This tree is called as FP-Tree. So at the end this FP-Tree states that which node as well as the path is most frequently visited by the user during the internet surfing. This may help in business decision making and the enhancements.

Chapter 3

Problem Statement

So in the web site industries, there is a need to identify that what are the items that a user or a group of users tried to search for. For this purpose various algorithms have been proposed. In Chapter 3 it is seen that existing algorithms are also used for web usage mining that first calculate the maximum forward reference and then creates the table. The nodes in the tables are arranged according to the frequency of traversal. To create the tree whole process is repeated again and again. Hence it will increase the complexity of the algorithm. Every time the analysts need to take a great amount of data, in the form of log files and the other data collection sources. In some cases even the information retrieval itself becomes a tedious task.

So in that situations the web usage mining techniques need some different approach, which do not need the tree creation step every time the analysis starts. Therefore there is a need of an algorithm which does not scan the whole data nodes for finding the most frequent pattern that a user generally follows. Once a web site owner gets this information they may employ some improvements in the product or the services. This information may also help them in various other decision making. Hence the requirement is to device an algorithm that is capable of achieving the goal of user access pattern identification, without the repeating scans.

Chapter 4
Implementation

In this section an attempt is made to solve our purpose in an efficient way. In this proposed work, the website structure itself is used during the analysis. The website structure means that the connectivity of the web pages that constitute the web data. In this algorithm, this architecture is considered only.

This algorithm takes an array named node_path_table as an input. The basic entity must be as follows.

```
Struct  node
{
  Char x; //for transaction id
  Char y[10]; //for storing the items
}
```

5.1 Proposed Algorithm

The related algorithm is as follows.

Algorithm 1: SSPRA (Single Scan Pattern Recognition Algorithm)

Input: Database, Website Structure, node_Path_table

```
1) Foreach transaction in database d.
      a. If Transaction t is not starting from the root
            i. Get the first node of the transaction t in s
           ii. Repeat the node_path_table until s=node_path_table[i].x
          iii. Prefix node_path_table[i].y to the transaction t.
//now we have to work on the counts
2) Foreach node in the transaction
      a. Increase the count value of node by 1
3) Get the nodes with the maximum counts and the preceding
   sequence form the node_path_table.
```

Because this scans the database only once it may be named as Single Scan Pattern Algorithm. Let us suppose that we have the website structure as shown in Table 4.1.

Here the transaction id of the user transaction and the path followed by the user is given.

Table 4.1 Input Paths (Traversal Paths)

Tid	**Database**
100	ABCD
200	ABEFG
300	ABEH
400	AIJK

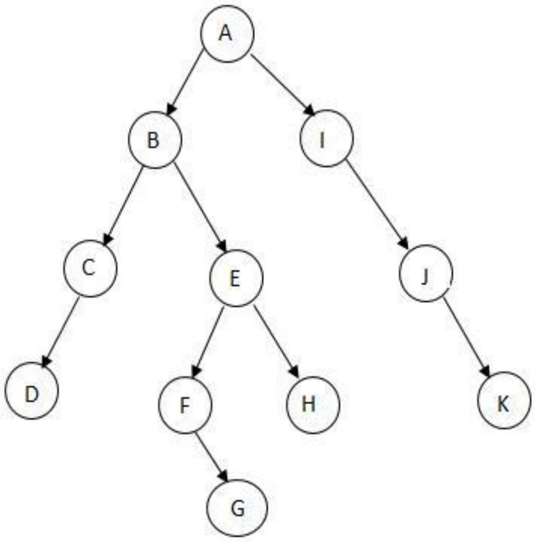


Figure 4.1- Website Structure

In Figure 4.1 the website structure is shown. This shows the way the web pages are connected to each other via the hyperlinks. Now by analyzing the transactions we can get the Table A. This table shows the transaction id and the path that the user followed. In this algorithm, the first transaction is taken and then that path is followed. Initially each node has count as 0. This count value is increased by one, whenever the node is traversed in between any path. The maximal paths are identified by using the algorithm named maximal forward reference.

Initially each node is having the count as 0, as shown in Figure 4.2. This is the very initial situation at the log file sources. Here the circles represent the web pages and the links represent the hyperlinks that connect the nodes in the data.

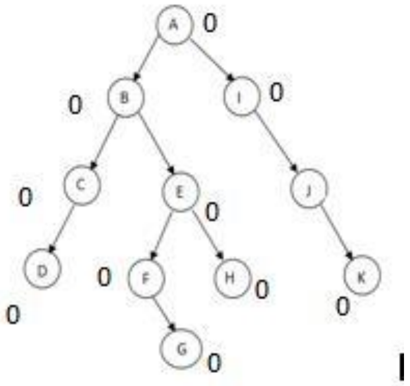


Figure 4.2 Initial Situation

After this, the database table is scanned again. Here the path ABCD is scanned first. This is the path which is traversed during the web site access. Initially, nodes A, B, C and D are having the count value as 0. These nodes are now updated to one, as shown in Figure 4.3.

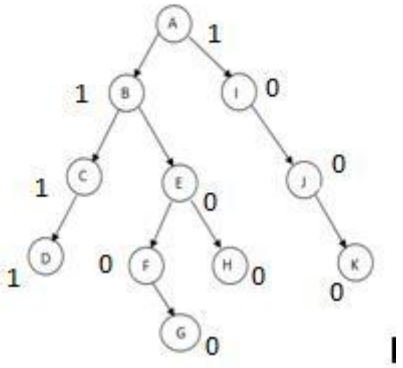


Figure 4.3 For ABCD

Now comes the path ABEFG, in the reference table. It can be noticed that the nodes A and B are repeated again, so their values are increased to 2, while the nodes E, F and G will have the count value as 1. This situation can be seen in Figure 4.4.

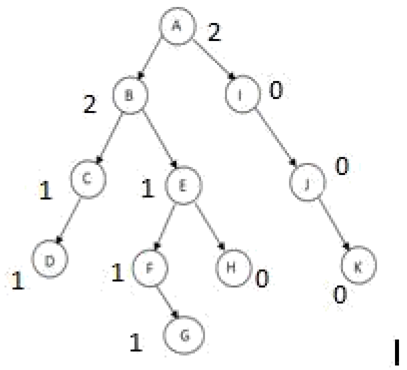


Figure 4.4 For ABEFG

The next path is ABE. Here, the nodes A and B are repeated again, so their values are increased to 3, while the node E, which is also repeated will have the count value as 2, as shown in the Figure 4.5.

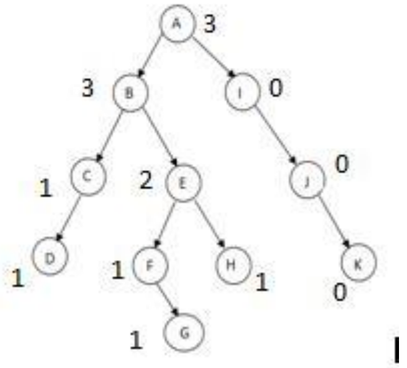


Figure 4.5 For ABE

Next is the path AIJK. Here, only the nodes A is repeated again, so it’s value is increased to 4, while the node I, J and K are not repeated at all, will have the count value as 1, as shown in the Figure 4.6.

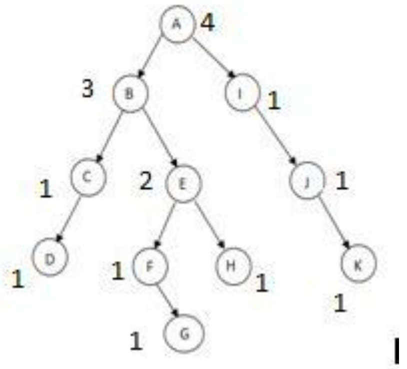


Figure 4.6 For AIJK

The next path is ABC. Here, the nodes A, B and C are repeated again, so their values are increased to 5, 4 and 2 respectively. There is no node which is repeated. The current situation can be seen in the Figure 4.7.

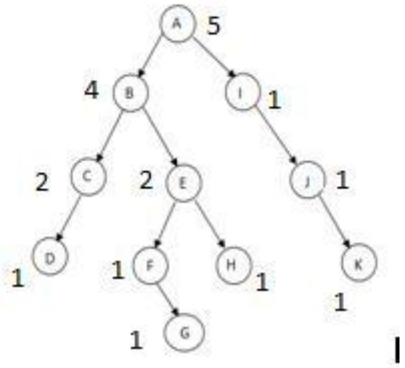


Figure 4.7 For ABC

Now there is a path AI. In this case A and I both are repeated, so these both will get there count value increased by 1, as shown in Figure 4.8.

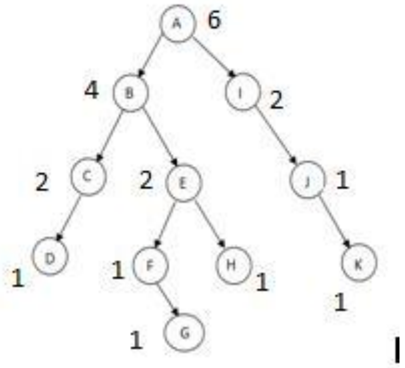


Figure 4.8 For AI

In path AIJ, all the node A, I and J are repeated , so their count value will also get increased by 1. So the count value will be A-7, I-3 and J-2, as shown Figure 4.9.

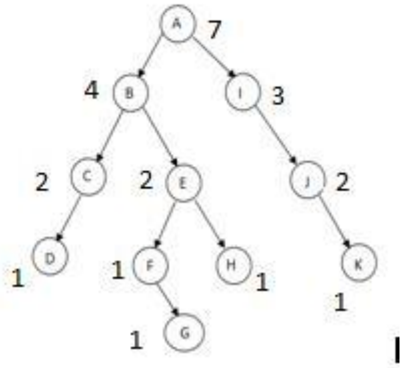


Figure 4.9 For AIJ

Now there is a path named as ABE. In this path the user goes to A, then B and then E. so the count values will be updated according to Figure 4.10.

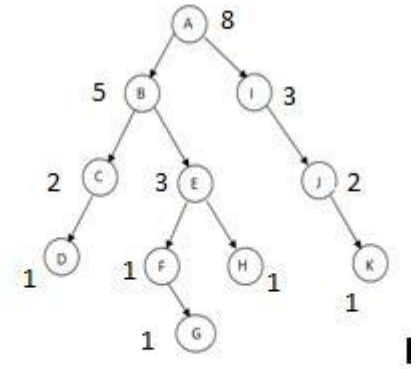


Figure 4.10 For ABE

ABEF is the path which the user traversed next. The count value of A, B, E and F is increased by one. So the situation is like Figure 4.11.

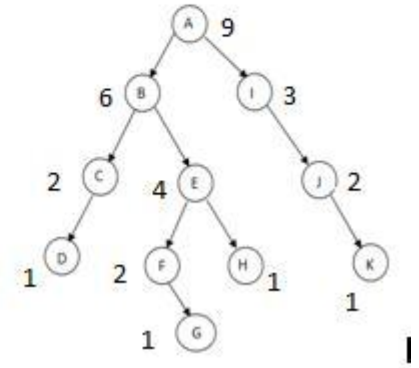


Figure 4.11 For ABEF

At the last the use traversed the path as AB. The count value of A becomes 10 and the count value of B becomes 7, as shown in Figure 4.12.

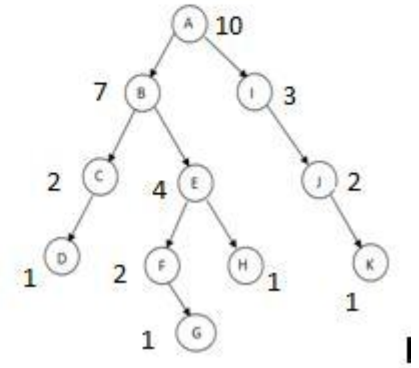


Figure 4.12 For AB

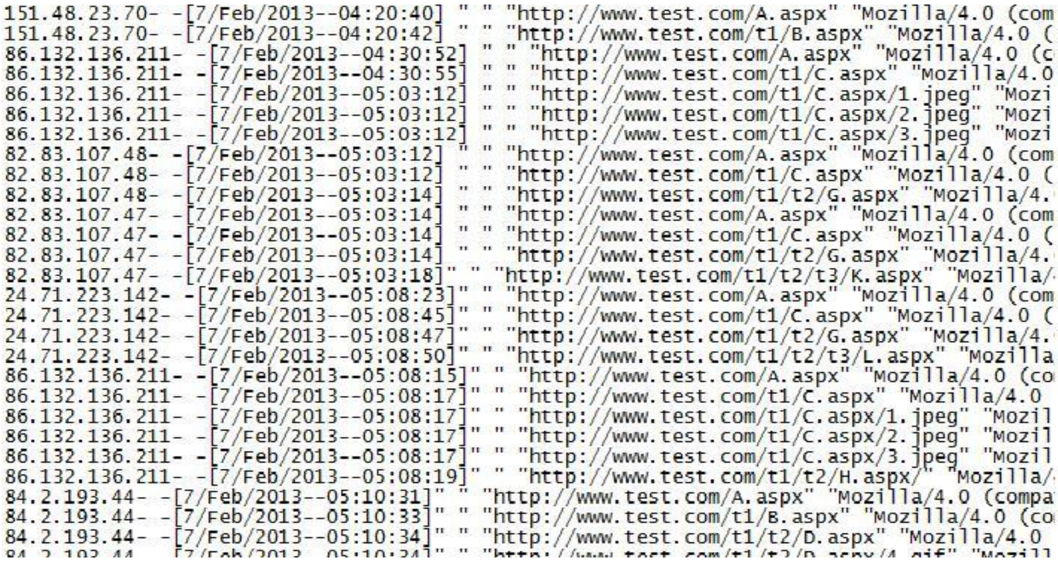
Figure 4.12 shows the situation when all the paths are identified and the reference count of each node. Here the numbers shown are the reference counts based on this, it can be identified that user traversed what kind of products and what kind of improvements are needed in the business layouts, or may recommend users for any suitable purchase.

Chapter 5
Testing and Results

In this chapter the implementation part is discussed. Along with the testing portion, the results are also compared by using different-different cases.

The log files, as shown in Figure 5.1, is the main content of any web usage mining techniques. This file may reside in the server side. This file may be at the client side but in that situation it may be difficult to extract that log file.

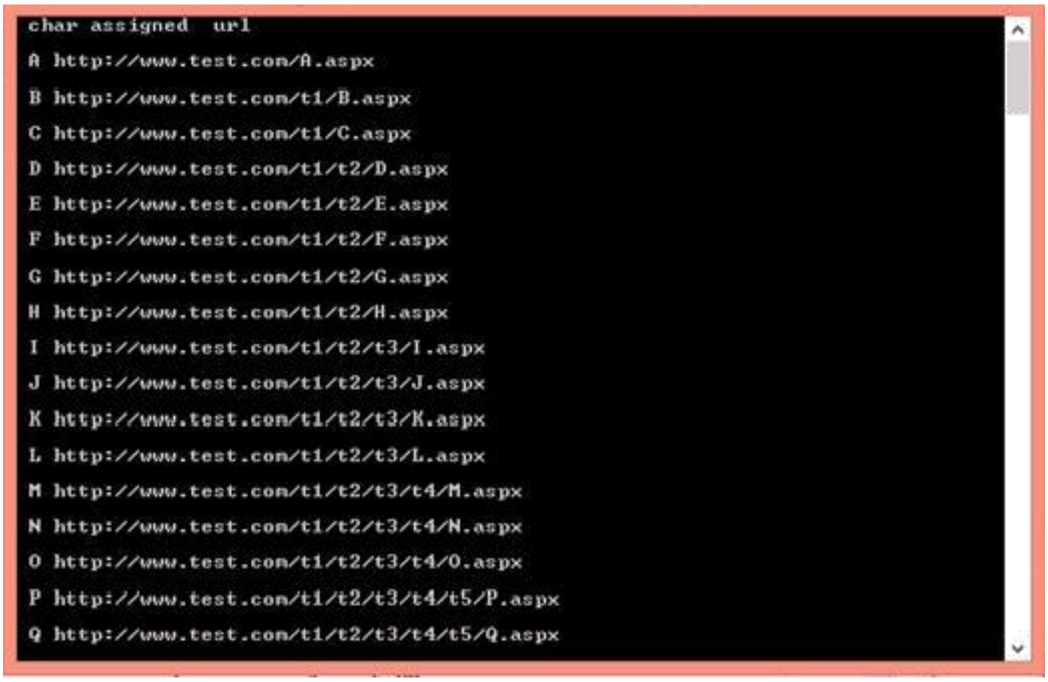
Case 1:



```
151.48.23.70- -[7/Feb/2013--04:20:40] " " "http://www.test.com/A.aspx" "Mozilla/4.0 (com
151.48.23.70- -[7/Feb/2013--04:20:42] " " "http://www.test.com/t1/B.aspx" "Mozilla/4.0 (
86.132.136.211- -[7/Feb/2013--04:30:52] " " "http://www.test.com/A.aspx" "Mozilla/4.0 (c
86.132.136.211- -[7/Feb/2013--04:30:55] " " "http://www.test.com/t1/C.aspx" "Mozilla/4.0
86.132.136.211- -[7/Feb/2013--05:03:12] " " "http://www.test.com/t1/C.aspx/1.jpeg" "Mozi
86.132.136.211- -[7/Feb/2013--05:03:12] " " "http://www.test.com/t1/C.aspx/2.jpeg" "Mozi
86.132.136.211- -[7/Feb/2013--05:03:12] " " "http://www.test.com/t1/C.aspx/3.jpeg" "Mozi
82.83.107.48- -[7/Feb/2013--05:03:12] " " "http://www.test.com/A.aspx" "Mozilla/4.0 (com
82.83.107.48- -[7/Feb/2013--05:03:12] " " "http://www.test.com/t1/C.aspx" "Mozilla/4.0 (
82.83.107.48- -[7/Feb/2013--05:03:14] " " "http://www.test.com/t1/t2/G.aspx" "Mozilla/4.
82.83.107.47- -[7/Feb/2013--05:03:14] " " "http://www.test.com/A.aspx" "Mozilla/4.0 (com
82.83.107.47- -[7/Feb/2013--05:03:14] " " "http://www.test.com/t1/C.aspx" "Mozilla/4.0 (
82.83.107.47- -[7/Feb/2013--05:03:14] " " "http://www.test.com/t1/t2/G.aspx" "Mozilla/4.
82.83.107.47- -[7/Feb/2013--05:03:18]" " "http://www.test.com/t1/t2/t3/K.aspx" "Mozilla/
24.71.223.142- -[7/Feb/2013--05:08:23]" " "http://www.test.com/A.aspx" "Mozilla/4.0 (com
24.71.223.142- -[7/Feb/2013--05:08:45]" " "http://www.test.com/t1/C.aspx" "Mozilla/4.0 (
24.71.223.142- -[7/Feb/2013--05:08:47]" " "http://www.test.com/t1/t2/G.aspx" "Mozilla/4.
24.71.223.142- -[7/Feb/2013--05:08:50]" " "http://www.test.com/t1/t2/t3/L.aspx" "Mozilla
86.132.136.211- -[7/Feb/2013--05:08:15]" " "http://www.test.com/A.aspx" "Mozilla/4.0 (co
86.132.136.211- -[7/Feb/2013--05:08:17]" " "http://www.test.com/t1/C.aspx" "Mozilla/4.0
86.132.136.211- -[7/Feb/2013--05:08:17]" " "http://www.test.com/t1/C.aspx/1.jpeg" "Mozil
86.132.136.211- -[7/Feb/2013--05:08:17]" " "http://www.test.com/t1/C.aspx/2.jpeg" "Mozil
86.132.136.211- -[7/Feb/2013--05:08:17]" " "http://www.test.com/t1/C.aspx/3.jpeg" "Mozil
86.132.136.211- -[7/Feb/2013--05:08:19]" " "http://www.test.com/t1/t2/H.aspx/" "Mozilla/
84.2.193.44- -[7/Feb/2013--05:10:31]" " "http://www.test.com/A.aspx" "Mozilla/4.0 (compa
84.2.193.44- -[7/Feb/2013--05:10:33]" " "http://www.test.com/t1/B.aspx" "Mozilla/4.0 (co
84.2.193.44- -[7/Feb/2013--05:10:34]" " "http://www.test.com/t1/t2/D.aspx" "Mozilla/4.0
84.2.193.44- -[7/Feb/2013--05:10:34]" " "http://www.test.com/t1/t2/D.aspx/4.gif" "Mozill
```

Figure 5.1 Web Log File

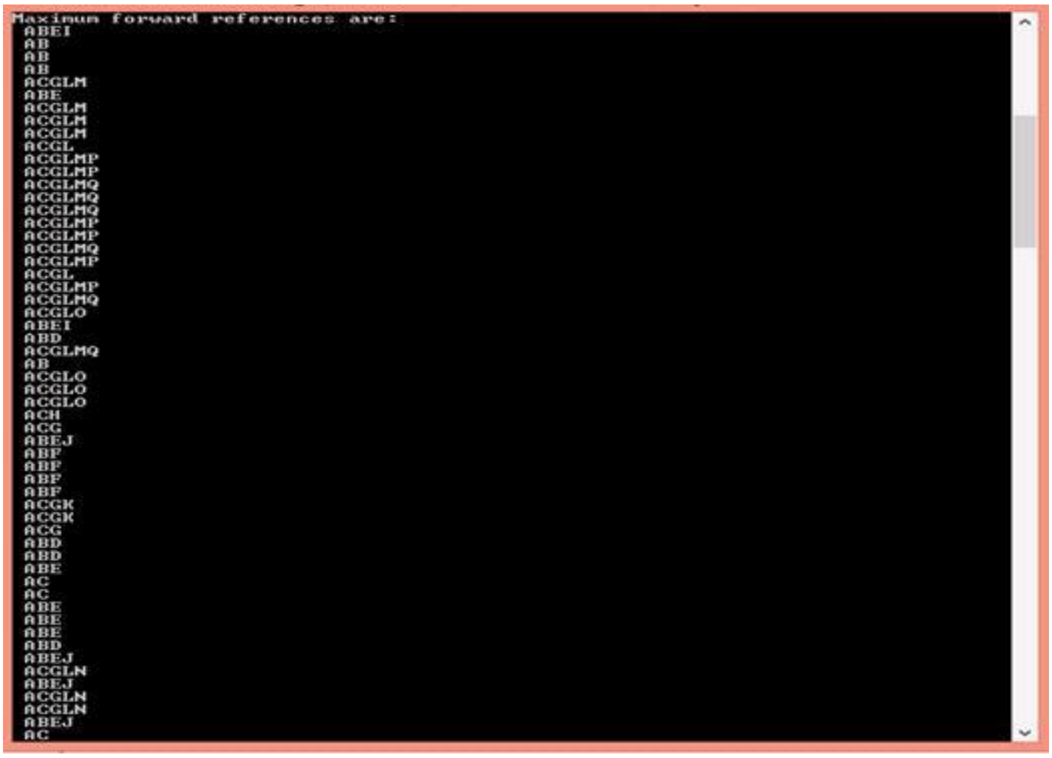
In the Figure 5.2, the various web pages are shown. Here the list of URL of the web pages are also shown. As there may be hundreds of pages in a single website, so these are the sub-pages a single web site. So these are the web pages which are open for the user to be accessed.



```
char assigned  url
A http://www.test.com/A.aspx
B http://www.test.com/t1/B.aspx
C http://www.test.com/t1/C.aspx
D http://www.test.com/t1/t2/D.aspx
E http://www.test.com/t1/t2/E.aspx
F http://www.test.com/t1/t2/F.aspx
G http://www.test.com/t1/t2/G.aspx
H http://www.test.com/t1/t2/H.aspx
I http://www.test.com/t1/t2/t3/I.aspx
J http://www.test.com/t1/t2/t3/J.aspx
K http://www.test.com/t1/t2/t3/K.aspx
L http://www.test.com/t1/t2/t3/L.aspx
M http://www.test.com/t1/t2/t3/t4/M.aspx
N http://www.test.com/t1/t2/t3/t4/N.aspx
O http://www.test.com/t1/t2/t3/t4/O.aspx
P http://www.test.com/t1/t2/t3/t4/t5/P.aspx
Q http://www.test.com/t1/t2/t3/t4/t5/Q.aspx
```

Figure 5.2 List of URL’s of the web pages

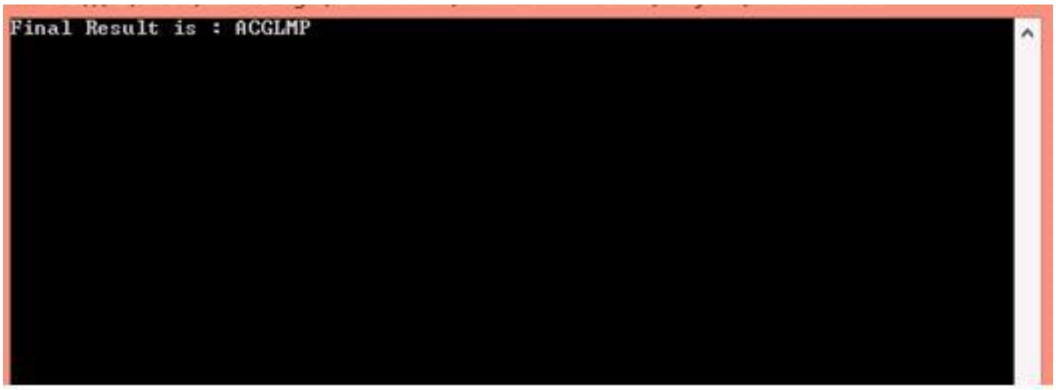
Now there is a need to calculate the maximum forward reference by using the web log files. The algorithm to calculate the MFR is same as the other algorithms uses it. These are shown in the Figure 5.3.



```
Maximum forward references are:
 ABEI
 AB
 AB
 AB
 ACGLM
 ABE
 ACGLM
 ACGLM
 ACGLM
 ACGL
 ACGLMP
 ACGLMP
 ACGLMQ
 ACGLMQ
 ACGLMQ
 ACGLMP
 ACGLMP
 ACGLMQ
 ACGLMP
 ACGL
 ACGLMP
 ACGLMQ
 ACGLO
 ABEI
 ABD
 ACGLMQ
 AB
 ACGLO
 ACGLO
 ACGLO
 ACH
 ACG
 ABEJ
 ABF
 ABF
 ABF
 ABF
 ACGK
 ACGK
 ACG
 ABD
 ABD
 ABE
 AC
 AC
 ABE
 ABE
 ABE
 ABD
 ABEJ
 ACGLN
 ABEJ
 ACGLN
 ACGLN
 ABEJ
 AC
```

Figure 5.3 Maximum Forward References

After employing the algorithm in JAVA, the results obtained are shown in the Figure 5.4.



```
Final Result is : ACGLMP
```

Figure 5.4 Most frequent path that the user traversed

This is the path which is traversed by the user most of the time, he or she traverses at the time of searching anything on the internet.

Case 2:

The results can be compared in graphical representation also, as shown in Figure 6.5.

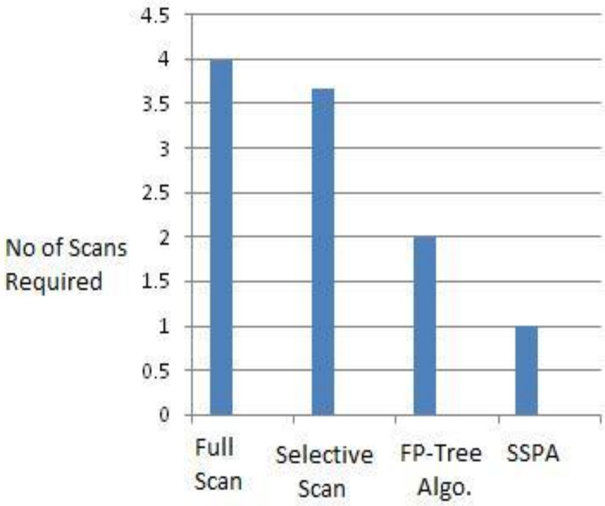


Figure 5.5 Comparison between the existing algorithm and SSPRA

As it is already said that the proposed algorithm is better in terms of number of comparisons and the number scans.

Now say number of level in a tree are N and the tree contains M nodes and the no of branches a node may have at a time is X, then we can have the equation (1).

$$N = \log_X M \tag{1}$$

Where

N: no. of levels in a tree

M: no. of nodes in the tree

X: The maximum no. of branches in a tree

This graph tells that in full scan the no of database scans needed are 4. So the no. of nodes in the tree are 81, the levels are 4 and the maximum branches a node can have are 3. In case of selective scan the levels in the tree are 3.67 (about 4), the no. of nodes are 13 (approx.) and the maximum allowable branches that a node may have are 2. For the FP tree algorithm and the SSPA algorithm the no. of scans are constant i.e. 2 and 1 respectively.

There is no need to think that FP-tree algorithm is far better than any other previous designs, but in this SSPA algorithm no tree is created. Creating an FP tree becomes an overhead sometimes, especially when there is large sized web structure. But in this proposed algorithm only the web structure is used and we can mine the data to find out the behavior of the user and help in business decision making.

Case 3:

As shown in the Table 5.1, the existing algorithms employ tree creation and the candidate generation. These two steps are very hard to execute. While in the proposed algorithm there is no need of either candidate generation or the tree creation. The Full Scan algorithm at least n number of database scans is required. Here n is the level of the tree in these algorithms. While in the proposed algorithm, there is no need to scan the database n times at all, only one scan is enough for the analysis.

Table 5.1 Comparison between other methods and the proposed algorithm

Factor	Full/Selective scan	SSPA
Consecutive Patterns	Perfect for the consecutive patterns.	Those hyperlinks are also considered in this algorithm, which are not even consecutive.
Tree creation	Needed	Not needed
Database Scan	With each pass database scan need to be done and it is equal to the level of the tree in both of these algorithms.	Single scan is needed.
Database size	It is efficient only for the small size databases.	Both the small and the large databases can be handled by this algorithm.
I/O Overhead	I/O cost is higher.	The amount of I/O overhead is lesser than other algorithm.
Candidate Generation	Candidate generation is performed.	No candidate generation is needed.

Because there is no need of any kind of tree creation, only one scan is needed. Only the website structure is used for the purpose of analysis. By using the website structure, each node gets a count variable. Each time the node is traversed; this variable gets increased by one. This is the variable "count", which plays an important role in analyzing the patterns.

The algorithms which are already in existence suffer from the I/O overhead and induce more cost to the execution. But the proposed algorithm does not suffer from the I/O overhead.

Therefore, it can be inferred from Case 1, Case 2 and Case 3 that the proposed algorithm is better than the existing algorithms such as FP-Tree algorithm, Full Scan algorithm and the Reference Scan algorithm. The numbers of scans in the proposed algorithm are reduced, so the I/O overhead is minimized.

Chapter 6
Conclusion and Future Scope

In thesis work various pattern recognition algorithms have been discussed and compared with the proposed algorithm. By the comparative apriori analysis, it can be seen that this SSPRA algorithm is working better than the existing algorithms. In case of scanning it takes only one scan for finding the most frequent data item the user is searching for. SSPRA is better than other algorithms in terms of the number of scans and would be useful for finding the frequent data item, the user is searching for.

Future Aspects

The work done in the thesis is useful in identifying the frequent patterns. In future, this work can be implemented in the big data environment. The proposed algorithm is suitable for the large data scans. As it is known that in the cloud environment the data becomes so much big. The web hosting is also supported by the clouds. So to analyze the data and get the user behavior this algorithm may get a good place.

References

[1]. C. Tsai, C. Lai and M. Chiang, "Data mining for internet of things: A survey," *IEEE Communication Surveys & Tutorials*, Vol. 16, No. 1, 2014.

[2]. D. V. Talele, V. Dipali. and C. D. Badgujar, "A Literature review of Opinoin Mining From Online Customers' Feedback and It's Applications Domains", *Asian Journal of Computer Science & Information Technology,* Vol.11, pp.301-305, 2013.

[3]. F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing" *Applicationof Information and Communication Technologies (AICT), 2011 5th International Conference on*. IEEE, pp. 1-4, 2011.

[4]. B. Roberto Espinosa, J. Zubcoff and J. Maz′on, "A Set of Experiments to Consider Data Quality Criteria in Classification Techniques for Data Mining", *International Conference on Computational Science and It's Applications" (ICCSA) 2011, Part II,* LNCS 6783, pp. 680–694, 2011.

[5]. L. Cao, "Domain-driven data mining: Challenges and prospects, "*Knowledgeand Data Engineering", IEEE Transactions, Vol.* 22, No. 6, pp. 755-769,2010.

[6]. V. Losarwar, M. Joshi, "Data Preprocessing in Web Usage Mining", *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012),* pp. 15-16, 2012.

[7]. F. Usama, P. Shapiro, and S. Padhraic, "From Data Mining to Knowledge Discovery in Databases", *American Association for Artificial Intelligence,*USA , pp.12-17, 1996.

[8]. Xu and Kaiquan, "Mining comparative opinions from customer reviews for Competitive Intelligence." *Decision support systems,* Vol. 50, No. 4, pp. 743-754, 2011.

[9]. Zhong, Ning, Li, Yuefeng, and Wu, Sheng-Tang, "Effective pattern discovery for text mining", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 24, No. 1, pp. 30-44, 2010.

[10]. O. Etzioni, “The World Wide Web: Quagmire or Gold Mine", *Communications of the ACM*, Vol. 39, No. 11, pp. 65-68, 1996.

[11]. B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification”, *International Journal of Computer Scienceand Information Security (IJCSIS),* Vol. 9, No. 4, April 2011.

[12]. K. Sharma, G.Shrivastava, and V. Kumar. "Web mining: Today and tomorrow", *Electronics Computer Technology (ICECT), 3rd InternationalConference*, Vol. 1 IEEE, 2011.

[13]. Web mining definition available: http://en.wikipedia.org/wiki/Web_mining.

[14]. R. Kosala and H. Blockeel, "Web Mining Research: A Survey", *ACMSIGKDD*, Vol. 2, No. 1, pp. 1-15, 2000.

[15]. Z. Ou, “Data Structure and Effective Retrieval in the Mining if Web Sequential Characteristic”, *International Conference on Electronic &Mechanical Engineering and Information technology*”, pp. 3551-3554, 2011.

[16]. Definition of Data Mining", Available: http://www.britannica.com/EBchecked/topic/1056150/data-mining, May 2012.

[17]. E. Han, G Karypis and V. Kumar, “Scalable Parallel Data Mining for Association Rule”, IEEE Transaction on Knowledge and Data Engineering, Vol. 12, No. 3, pp. 337-352, 2000.

[18]. Q. Zhang, and R. S. Segall, "Web Mining: A Survey of Current Research, Techniques, and Software", *International Journal of Information Technologyand Decision Making*, Vol. 07, No. 4, pp. 683-720, 2008.

[19]. W. Xindong, Z. Xingquan, G. Wu, and Wei Ding, “Design and Implementation of Patent Data Preprocessing System Based on Data Mining Theory”, *IEEE Transactions On Knowledge and Data Engineering,* Vol. 26, No. 1, 2014.

[20]. D. J. Velásquez and V. Palade, "A knowledge base for the maintenance of knowledge extracted from web data", *Knowledge-Based Systems*, Vol. 20, No. 3, pp. 238-248, 2007.

[21]. M. Efron, "Information search and retrieval in Microblogs", *Journal of theAmerican Society for Information Science and Technology,* Vol. 62, No. 6, pp.996-1008, 2011.

[22]. R.Cooley, J.Srivastava, and B.Mobasher, “Web mining: Information and pattern discovery on the World Wide Web”, *Proceedings of the 9th IEEEInternational Conference on Tools with Artificial Intelligence (ICTAI’97),* pp.558-567, 1997.

[23]. A. Abbasi, and J. Altmann, "A social network system for analyzing publication activities of researchers.",*Collective intelligence, Springer BerlinHeidelberg,* Vol. 58, pp. 49-61, 2011.

[24]. M. L. Huan, R. Setiono and Z. Zhao, “Feature Selection: An Ever Evolving Frontier in Data Mining”, *Workshop and Conference Proceedings 10,* pp. 4-13, 2010.

[25]. R. Liu and k. Chang, “Editorial: Special Issue on Web Content Mining”, *SIGKDD Explorations special issue on Web Content Mining,* vol. 6, no. 2,2004.

[26]. V. Bharanipriya and V. K. Prasad, "Web content mining tools: a comparative study" *International Journal of Information Technology and KnowledgeManagement*, Vol. 4, No. 1, pp. 211-215, 2011.

[27]. T. Jing, W. L. Zou, B. Z. Zhang, “An Efficient Web Traversal Pattern Mining algorithm Based On Suffix Array”, *Proceedings of the 3rd InternationalConference on Machine Learning and Cybernetics*, pp. 1535-1539, 2004.

[28]. D. Stiawan, I. Mohd and A. A. Hanan, "Survey on Heterogeneous Data for Recognizing Threat", *Journal of Computational Information Systems*, Vol. 7, No.12, pp. 4212-4224, 2011.

[29]. G. Siemens and R. S. Jd. Baker, "Learning analytics and educational data mining: towards communication and collaboration", *Proceedings of the 2ndinternational conference on learning analytics and knowledge*. ACM, 2012.

[30]. Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin and J. M. Hellerstein, "Distributed Graph Lab: a framework for machine learning and data mining in the cloud”, *Proceedings of the VLDB Endowment*, Vol. 5, No. 8, pp. 716-727, 2012.

[31]. C. Phua, V. Lee, K. Smith and R. Gayler, "A comprehensive survey of data mining-based fraud detection research" *arXiv preprint arXiv:1009.6119*, 2010.

[32]. K. Sharma, S. Gulshan, and K. Vikas, "Web mining: Today and tomorrow", *Electronics Computer Technology (ICECT), 3rd InternationalConference,* Vol. 1. pp. 403-399, 2011.

[33]. S. Chakrabarti, "Data mining for Hypertext: A tutorial Survey", ACM SIGKDD, Vol. 1, Issue 2, pp. 1-11, 2000.

[35]. About Web, available:
http://www.technicalsymposium.com/web_mining_not
es.html.

[36]. R. Jain and G. N. Purohit, "Page Ranking Algorithms for Web Mining", *International Journal of Computer Applications (IJCA),* Vol. 13, No.5, pp. 22-25, January 2011.

[37]. Z. Gong, "Web Structure Mining: An Introduction ", *Proceedings of the 2005IEEE International Conference on Information Acquisition*, 2005.

[38]. R. Sharma and K. Kaur, "Review of Web Structure Mining Techniques using Clustering and Ranking Algorithms", *IJRCCT, Vol.* 3, No. 6, pp. 663-668, 2014.

[39]. P. Rani, "A Review of Web Page Ranking Algorithm "*International Journalof Advanced Research in Computer Engineering & Technology (IJARCET),* Vol. 2, Issue 3, pp-1052-1054, 2013.

[40]. A.Singh and S. Sharma, "Role of Page ranking algorithm in Searching the Web: A Survey", International Journal of Engineering & Technology, Management and Applied Sciences, Vol. 1, Issue 1, June 2014.

[41]. P. Bari and P. M. Chawan. "Web Usage Mining", *Journal of EngineeringComputers & Applied Science,* Vol. 2, No. 6, pp. 34-38, 2013.

[42]. N. K. Tyagi, A. K. Solanki, and S. Tyagi, "An algorithmic approach to data preprocessing in web usage mining", *International Journal of InformationTechnology and Knowledge Management*, Vol.2, No.2, pp. 279-283, 2010.

[43]. K. Xu, S. Shaoyi Liao, J. Li and Y. Song, “Mining comparative opinions from customer reviews for Competitive Intelligence“, *Decision Support Systems*, Vol. 50, pp. 743-754, 2011.

[44]. Q. Yang, "A novel two-stage scheme built-upon clustering for sequential pattern mining", *International Journal of Innovative Computing, Informationand Control*, Vol. 7, No. 5, pp. 2809-2818, 2011.

[45]. CRM and data mining, available at: https://www.salford systems.com/ resources/whitepapers/103-customer-relationship-management-and-data-mining

[46]. J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan, “Web usage mining: discovery and applications of usage patterns from web data”, *SIGKDDExplorations*, Vol.1, No. 2, pp. 12–23, 2000.

[47]. A. K. Sharma and A. Goel. "A Framework for Prefetching Relevant Web Pages using Predictive Prefetching Engine (PPE)." *arXiv preprintarXiv:1109.6206*, 2011.

[47]. P. Britos, D. Martinelli, H. Merlino and R. García-Martínez, “Web Usage Mining Using Self Organized Maps”, *International Journal of ComputerScience and Network Security*, Vol.7, No.6, June 2007.

[48]. M. Jalali, N. Mustapha, A. Mamat. and M. N. B. Sulaiman, “Web user navigation pattern mining approach based on graph partitioning algorithm”, *Journal of Theoretical and Applied Information Technology*, Vol. 11, No. 4,2006.

[49]. N. Sujatha and K. Iyakutty, “Refinement of Web usage Data Clustering from K-means with Genetic Algorithm”, *Proceedings of European Journal ofScientific Research* , Vol. 42, No. 3, pp.464-476, 2010.

[50]. M. Khosravi and M. J. Tarokh, “Dynamic Mining of Users Interest Navigation Patterns Using Naive Bayesian Method”, *Proceedings of the IEEE sixthInternational Conference on Intelligent Computer Communication and Processing*, pp. 119-122 , 2010.

[51]. K. Pani, L. Panigrahy, V. H. Sankar, B. K. Ratha, A. K. Mandal and S. K. Padhi, *Proceedings of International Journal of Instrumentation, Control &Automation (IJICA),* Vol. 1, No. 1, 2011.

[52]. M. Suman, "A Frequent Pattern Mining Algorithm Based on Fp-Tree Structure and apriori Algorithm”, *International Journal of EngineeringResearch and Applications (IJERA)*, Vol. 2, No. 1, pp. 114-116, 2012.

List of Publications

1. M. M. Sharma and A. Bala, “An Approach for Frequent Access Pattern Identification in Web Usage Mining”, *IEEE Conference ICACCI-2014, Delhi, India: Industry Track Papers - ICACCI Industry Track.* (Status: Accepted)

2. M. M. Sharma and A. Bala, “Survey Paper on Workflow Scheduling Algorithms Used in Cloud Computing”, *International Journal of Information & Computation Technology (IJICT),* ISSN 0974-2239 Vol. 4, No. 10, pp. 997-1002, 2014.