# Reference Scan Algorithm for Path Traversal Patterns

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

**Master of Engineering**
in
**Software Engineering**

*Submitted By*
**Chintandeep Kaur**
**801031005**

Under the supervision of:
**Dr. Rinkle Rani**
**Assistant Professor**
**CSED**

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

**June 2012**

## CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"Reference Scan Algorithm for Path Traversal Patterns"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software Engineering* 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 Dr. Rinkle Rani 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.

chintandeep K.

(Chintandeep Kaur)

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

Rinkle 07/06/12

(Dr. Rinkle Rani)
Assistant Professor,
CSED

**Countersigned by**

**(Dr. Maninder Singh)**
Head,
Computer Science and Engineering Department,
Thapar University,
Patiala.

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

# Acknowledgement

It is a great pleasure for me to acknowledge the guidance, assistance and help I have received from **Dr. Rinkle Rani.** I am thankful for her continual support, encouragement, and invaluable suggestions. She not only provided me help whenever needed, but also the resources required to complete this thesis report on time.

I am also thankful to **Dr. Maninder Singh**, Head, Computer Science and Engineering Department for his kind help and cooperation.

I would also like to thank all the staff members of Computer Science and Engineering Department for providing me all the facilities required for completion of my thesis.

I would like to say thanks to my classmates for their continuous support. Also, I would thank my parents for their inspirational and ever encouraging moral support, which enabled me to pursue my studies.

# Abstract

The World Wide Web is an immense source of data that can come either from the web contents represented by the billions of pages publicly available or from the web usage represented by the log information daily collected by all the servers around the world. Web usage mining is that area of web mining which deals with the extraction of interesting knowledge from logging information produced by web servers.

Frequent pattern mining is a heavily researched area in the field of web usage mining with wide range of applications. The aim of discovering frequent patterns in web log data is to obtain information about the navigational behaviour of the users. This can be used for advertising purposes, for creating dynamic user profiles, increasing server performance and enhancing the website usage.

Many algorithms have been proposed in the same context in last decade like apriori based algorithm which are based upon candidate and test generation which suffer from repeated database scan. The aim of this study is to make comparison of the apriori based algorithm and thus propose new approach to obtain the frequent patterns that are accessed by the user while traversing a particular website.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1 Introduction

As of classical data mining, the aim of web mining is to discover and retrieve useful and interesting patterns from a large dataset. There has been huge interest towards web mining. In web mining, this dataset is the huge web data. Web data contains different kinds of information including web documents data, web structure data, web log data, and user profiles data. All of web data can be mined mainly in three different dimensions which are; web content mining, web structure mining, and web usage mining [5].

Web mining is the integration of information gathered by traditional data-mining methodologies and techniques with information gathered over the World Wide Web [60]. Just as data mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of web mining aims at finding and extracting relevant information that is hidden in the web pages which are accessed by the user in variations. Web-mining is a multi-disciplinary effort that draws techniques from fields like in-formation retrieval, statistics, machine learning and natural language processing. Web mining has some new characteristics compared to traditional data mining [52]

1. The objects of web mining are a large number of web documents which are heterogeneously distributed and each data source is heterogeneous

2. The web document itself is semi-structured or un-structured and lack the semantics which a machine can understand.

## 1.1 Data Mining

Data mining has an important place in today's world. It becomes an important research area since the amount of data available in most of the applications is huge. This huge amount of data must be processed in order to extract useful information and knowledge since they are not explicit [5].

The definition of data mining is given in as "*Data Mining is the process of discovering interesting knowledge from large amount of data*".

Data mining involves six common classes of tasks [58]

- *Anomaly detection* - The identification of unusual data records that might be interesting or data errors and require further investigation.

- *Association rule learning* – It Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

- *Clustering* – It is the task of discovering groups and structures in the data that are in some way or another similar without using known structures in the data.

- *Classification* – It is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as legitimate or as spam.

- *Regression* – It attempts to find a function which models the data with the least error.

- *Summarization* – It provides a more compact representation of the data set including visualization and report generation.

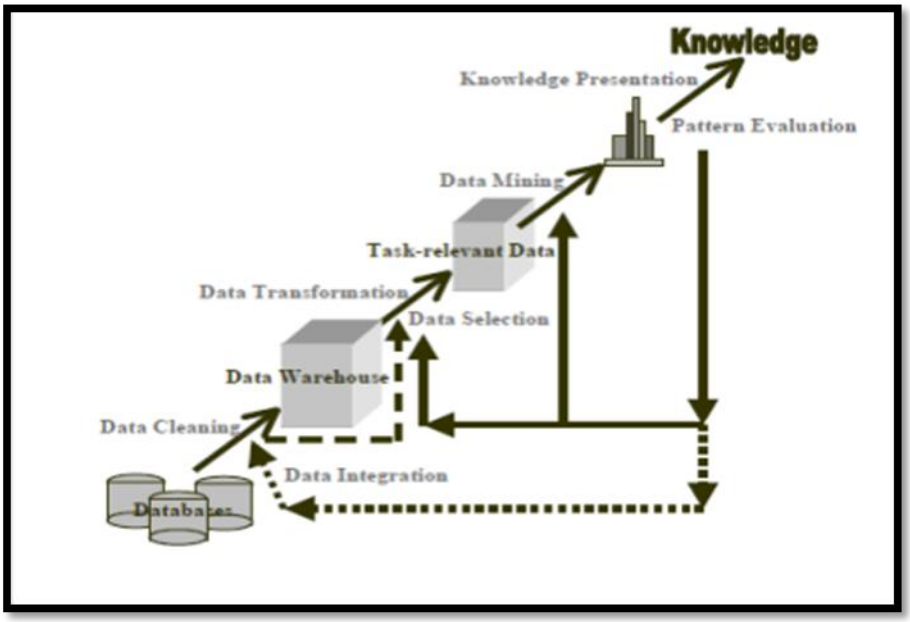


Figure 1: Phases of Data Mining [5]

### 1.1.1 Data Mining Process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps as shown in figure 1[5]

· **Data cleaning**

This is also known as data cleansing. It is a phase that handles missing and redundant data in the source file. The real world data can be incomplete, inconsistent and corrupted. In this process, missing values can be filled or removed, noise values are smoothed, outliers are identified and each of these deficiencies are handled by different techniques.

· **Data integration**

This process combines data from various sources. The source data can be multiple distinct databases having different data definitions. In this case, data integration process inserts data into a single coherent data store from these multiple data sources.

· **Data selection**

In this step, the data relevant to the analysis is decided on and retrieved from the data collection.

· **Data transformation**

It is also known as data consolidation. Data transformation process converts source data into proper format for data mining. Data transformation includes basic data management tasks such as smoothing, aggregation, generalization, normalization and attributes construction.

· **Data mining**

In Data mining process, intelligent methods are applied in order to extract data patterns.

· **Pattern evaluation**

It is the task of discovering interesting patterns among extracted pattern set. Knowledge representation includes visualization techniques which are used to interpret discovered knowledge to the user.

- **Knowledge representation**

  It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

World Wide Web is one of the largest and most widely known data source. Today, www contains billions of documents edited by millions of people. The total size of the whole documents can be interpreted in many terabytes. All documents on www are distributed over millions of computers that are connected by telephone lines, optical fibres and radio modems. World Wide Web is growing at a very large rate in size of the traffic, the amount of the documents and the complexity of 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 [21] which is the application of data mining techniques to World Wide Web. The next section explains general overview of web mining.

## 1.2 Web Mining

Web data mining is an important area of data mining which deals with the extraction of interesting knowledge from the World Wide Web. As many believe, it is Oren Etzioni [21] who first proposed the term of web mining in 1996 in his paper. In the paper he claimed that web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. In the same paper, Etzioni came up with the question: Whether effective web mining is feasible in practice? Today, with the tremendous growth of the data sources available on the web and the dramatic popularity of e-commerce in the business community, web mining has become the focus of quite a few research projects and papers. This area of research is so huge today due to the tremendous growth of information sources available on the web and the recent interest in e-commerce. Web mining is used to understand customer behaviour, evaluate the effectiveness of a particular web site, and help quantify the success of a marketing campaign.

There is no agreed definition of web data mining but we present one simple definition [60]:
"*Web Data Mining is the application of data mining techniques to find 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.*"

Web data can be [57]

- Content of actual web pages.
- Intra page structure which includes the HTML or XML node for the page.
- Inter page structure is the actual linkage structure between web pages.
- Usage data that describe how web pages are accessed by visitors.
- User profiles include demographic and registration information obtained about users.

### 1.2.1 Web Mining Challenges:

1. *Finding Relevant Information*

   People either browse or use the search service when they want to find specific information on the web. However today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.

2. *Creating new knowledge out of the information available on the web*

   This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that we already have collection of web data and we want to extract potentially useful knowledge out of it.

3. *Personalization of information*

   When people interact with the web they differ in the contents and presentations they prefer.

4. *Learning about Consumers or individual users*

   This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to web site design and management and marketing. [37]

1.2.2 Web Mining Subtasks:

1. *Resource finding*

   The task of retrieving intended web documents. By resource finding we mean the process of retrieving the data that is either online or offline from the text sources available on the web such as electronic newsletters, electronic newswire, the text contents of HTML documents obtained by removing HTML tags and also the manual selection of web resources.

2. *Information selection and pre-processing*

   Automatically selecting and pre-processing specific information from retrieved web resources. It is a kind of transformation processes of the original data retrieved in the IR process. These transformations could be either a kind of pre-processing 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.

3. *Generalization*

   It is the process which automatically discovers general patterns at individual web sites as well as across multiple sites. Machine learning or data mining techniques are typically used in the process of generalization. Humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium.

4. *Analysis*

   It involves validating and/or interpretation of the mined patterns. [7]

Kosala and Blockeel who performed research in the area of web mining suggested [37]

1. Mining for information
2. Mining the web link structure
3. Mining for user navigation patterns

**Table1: Relationship between Content, Structure and Usage Mining [45]**

| Type | Structure | Form | Object | Collection |
|---|---|---|---|---|
| **Usage** | Accessing | Click | Behavior | Logs |
| **Content** | Pages | Text | Index | Pages |
| **Structure** | Map | Hyperlinks | Map | Hyperlinks |

Mining for information focuses on the development of techniques for assisting a user in finding documents that meet a certain criterion that is web content mining. Web content mining refers to the discovery of useful information from web contents including text, images, audio and video. Mining the link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining. Web structure mining tries to discover the model underlying the link structures of the web. Model is based on the topology of hyperlinks with or without description of links. Finally, mining for user navigation patterns focuses on techniques which study the user behavior when navigating the web that is web usages mining. Web usage mining refers discovery of user access patterns from Web servers. Web usages data include data from web server access logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls or any other data as result of interaction.

## 1.3 Web Mining Taxonomy

Web mining is considered to be of three types [15]

1. Web content mining
2. Web structure mining
3. Web usage mining

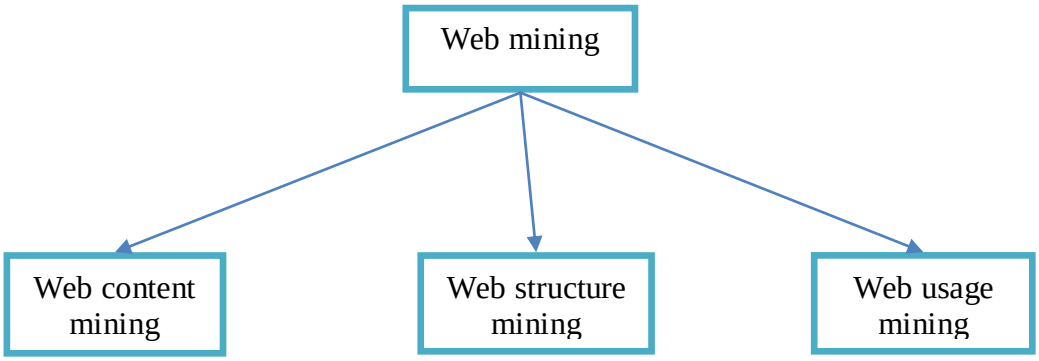


**Figure 2: Web Mining Taxonomy [15]**

### 1.3.1 Web Content Mining

Web content mining describes the automatic search of information resources available online [38], and involves mining web data contents as in figure 3. In the web mining domain, Web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in Web documents. The web document usually contains several types of data such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of web data forces the web content mining towards a more complicated approach. The web content mining is differentiated from two different points of view: Information Retrieval View and Database View.

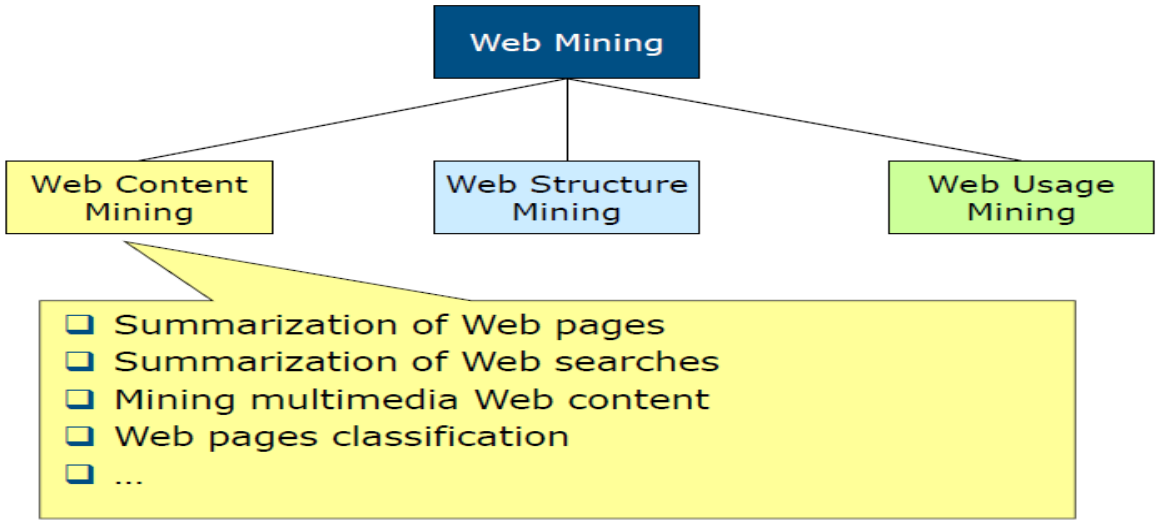


**Figure 3: Web Content Mining [28]**

### 1.3.2 Web Structure Mining

Most of the web information retrieval tools only use the textual information while ignore 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 mainly focuses on the structure of 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. 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 in figure 4; this would be good for navigation purpose and make it possible to compare or integrate web page schemes [1].

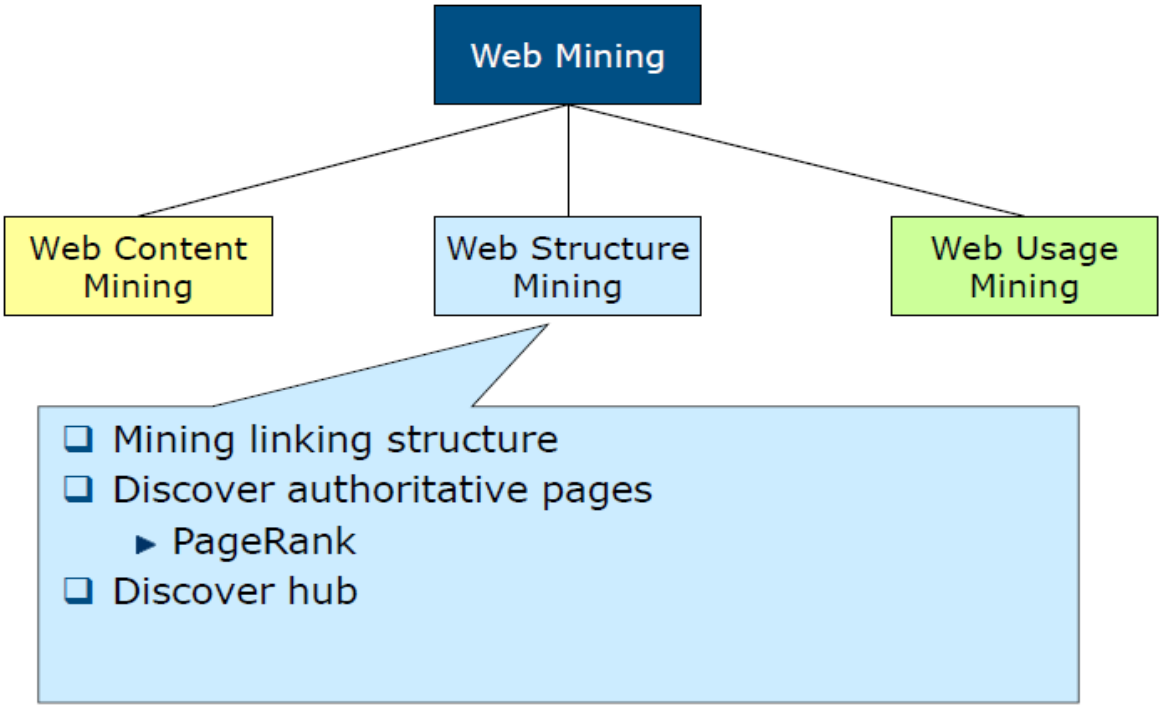


**Figure 4: Web Structure Mining [28]**

Another task of web structure mining is to discover the nature of the hierarchy or network of hyperlinks in the web sites of a particular domain. This may help to generalize the flow of information in web sites that may represent some particular domain therefore; the query processing will be easier and more efficient. Web structure mining has a natural relation with the web content mining since it is very likely that the web documents contain links and they both use the real or primary data on the web. It is quite often to combine these two mining tasks in an application.

### 1.3.3 Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web data in order to understand and better serve the needs of web-based applications. It tries to make sense of the data generated by the web surfer's sessions or behaviours. 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 3 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 all services 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 [60].

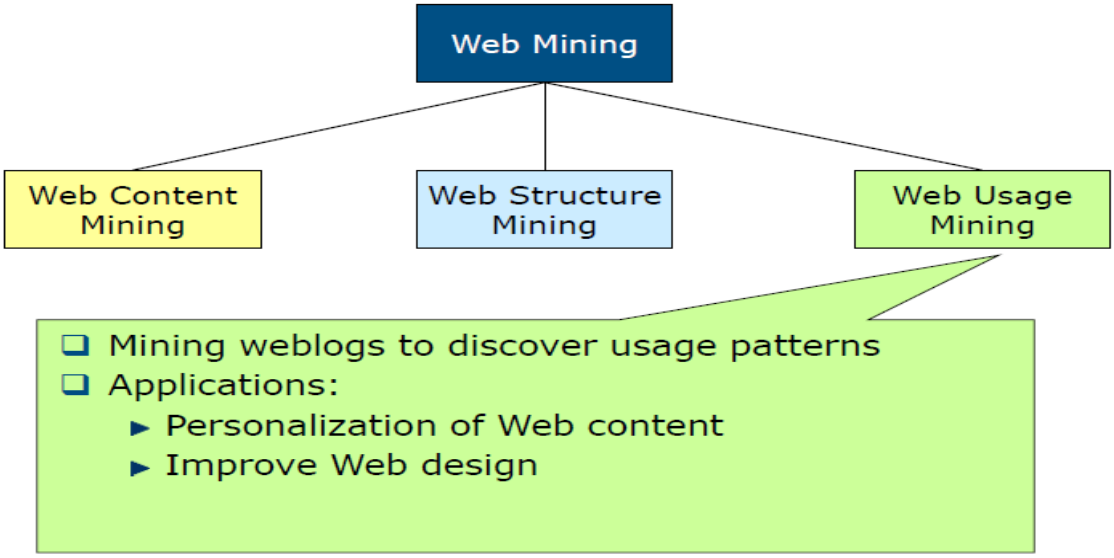


**Figure 5: Web Usage Mining [28]**

### 1.3.4 Web Usage Mining Process

Web usage mining process as shown in figure 6 can be regarded as a three-phase process consisting [60]:

i. **Preprocessing or Data Preparation**

Web log data are preprocessed in order to clean the data which implies removing log entries that are not needed for the mining process, data integration, identify users and sessions.

ii. **Pattern Discovery**

Statistical methods as well as data mining methods (Path analysis, Association rule, Sequential patterns, Clustering and Classification rules) are applied in order to detect interesting patterns.

**iii. Pattern analysis phase**

Discovered patterns are analyzed here using OLAP tools, knowledge query management mechanism and intelligent agent to filter out the uninteresting rules or patterns.

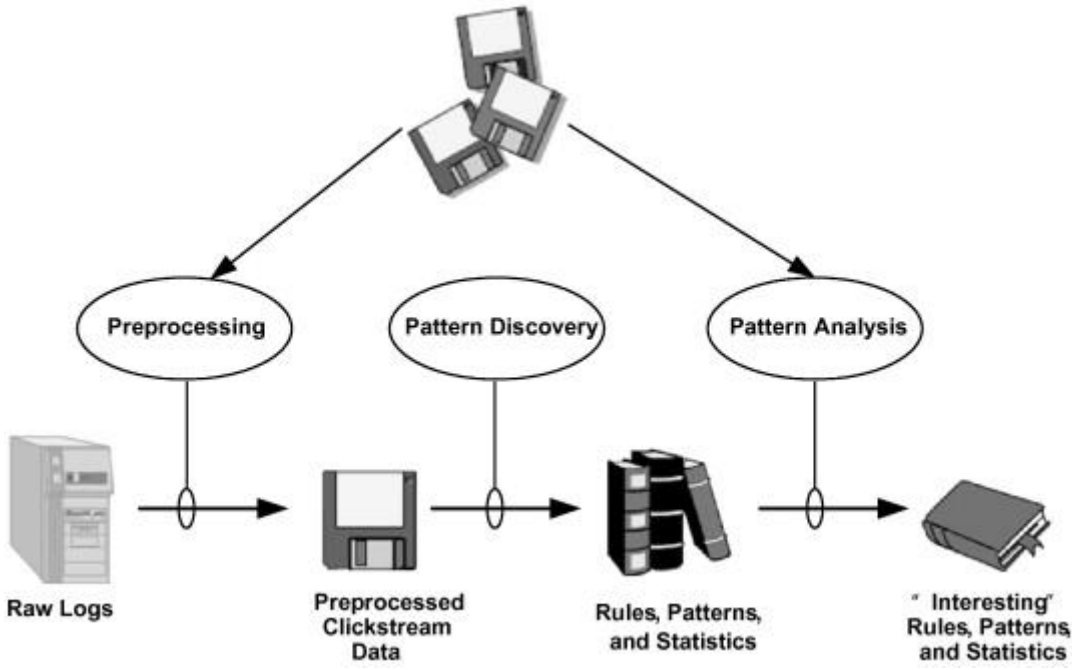


**Figure 6: Web Usage Mining Process [60]**

After discovering patterns from usage data, a further analysis has to be conducted. The most common ways of analyzing such patterns are either by using query or by loading the results into a data cube and then performing OLAP operations. Then, visualization techniques are used for results interpretation. The discovered rules and patterns can then be used for improving the system performance or for making modifications to the web site. The purpose of web usage mining is to apply statistical and data mining techniques to the pre-processed web log data in order to discover useful patterns. Usage mining tools discover and predict user behaviour in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service.

The applications are

i. Extract statistical information and discover interesting user patterns
ii. Cluster the user into groups according to their navigational behaviour
iii. Discover potential correlations between web pages and user groups
iv. Identification of potential customers for ecommerce

v. Enhance the quality and delivery of Internet information services to the end user
vi. Improve web server system performance and site design
vii. Facilitate personalization

## 1.4 Data Sources

The usage data collected at the different sources will represent the navigation patterns of different segments of the overall Web traffic ranging from single-user and single-site browsing behaviour to multi-user and multi-site access patterns.

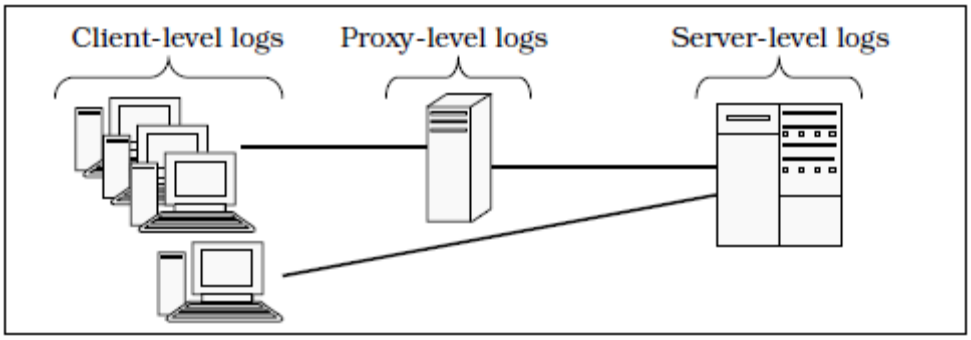


**Figure 7:Web Usage Data Sources [16]**

### 1.4.1 Server Level Collection

A web server log is an important source for performing web usage mining because it explicitly records the browsing behaviour of site visitors. The data recorded in server logs reflects the access of a web site by multiple users. These log files can be stored in various formats such as Common log or extended log formats. However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log. In addition, any important information passed through the POST method will not be available in a server log. The web server can also store other kinds of usage information such as cookies and query data in separate logs. Cookies are tokens generated by the web server for individual client browsers in order to automatically track the site visitors. Query data is also typically generated by online visitors while searching for pages relevant to their information needs. Besides usage data, the server side also provides content data, structure information and web page meta-information.

### 1.4.2 Client Level Collection

Uniform Resource Identifier (URI) is a more general definition that includes the commonly referred to Uniform Resource Locator (URL). Client-side data collection can be implemented

by using a remote agent (such as JavaScript's or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities. The implementation of client-side data collection methods requires user cooperation either in enabling the functionality of the JavaScript's and Java applets or to voluntarily use the modified browser. Client-side collection has an advantage over server-side collection because it ameliorates both the caching and session identification problems. However, Java applets perform no better than server logs in terms of determining the actual view time of a page. In fact it may incur some additional overhead especially when the Java applet is loaded for the first time. JavaScript's on the other hand consume little interpretation time but cannot capture all user clicks (such as reload or back buttons). These methods will collect only single-user and single-site browsing behaviour. A modified browser is much more versatile and will allow data collection about a single user over multiple Web sites.

### 1.4.3 Proxy Level Collection

A web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a web page experienced by users as well as the network traffic load at the server and client sides [16]. The performance of proxy caches depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple web servers. This may serve as a data source for characterizing the browsing behaviour of a group of anonymous users sharing a common proxy server.

## 2.1 Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) [58] a relatively young and interdisciplinary field of computer science [13] [33] is the process that results in the discovery of new patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics and database systems [33]. The overall goal of the data mining process is to extract knowledge from an existing data set and transform it into a human-understandable structure for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), 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 in further analysis or for example in machine learning and predictive analytics. For example the data mining step might identify multiple groups in the data which can then be used to obtain more accurate prediction results by a decision support system. The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Baye's theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage and manipulation ability. As data sets have grown in size and complexity direct "hands-on" data analysis has increasingly been augmented with indirect automated data processing aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees (1960s) and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets [33].

## 2.2 Web Mining

In the last fifteen years, the growth in number of web sites and visitors to those web sites has increased exponentially. The number of users by December 31, 2011 was 2,267,233,742 [3] which is 32.7% of the world's population. The number of active web site is 140,365,845[4] as on 13-Dec-2011. Due to this growth a huge quantity of web data has been generated. To mine the interesting data from this huge pool, data mining techniques can be applied. But the

web data is unstructured or semi structured. So we can not apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is used to discover interesting patterns which can be applied to many real world problems like improving web sites, better understanding the visitor's behaviour or product recommendation.

Sparked in 1996 by etzioni [21] in his research paper web mining rapidly became a new exploring field which was expanding each moment. He brought out a question: is it practical to mine Web data? He also suggested dividing the Web mining to three processes. The paper opened up a new active research field. There are increasing number of researcher's working on this field and they do some surveys around the data mining on the web. The web mining was clearly categorized as web content mining, web structure mining and web usage mining in till 1999. The research works have been well classified since then. There have been some works around content mining and structure mining based on the research of data mining and Information Retrieval, Information Extraction and Artificial Intelligence. The rapid expansion of the web is causing the constant growth of information leading to several problems such as an increased difficulty of extracting potentially useful knowledge.

Web content mining means automatic search of information resources available online [65] which means mining the data on the web. Web structure mining means mining the web document's structure and links, in short, mining the web structure data. Web usage mining includes the data from server access logs, user registration or profiles, user sessions or transactions, in short, mining the web log data.

Web mining subtasks are [37]

(a) Resource finding and retrieving
(b) Information selection and pre-processing
(c) Patterns analysis and recognition
(d) Validation and interpretation
(e) Visualization

Web mining is often associated with IR or IE. However, web mining is not the same as IR or IE.

### 2.2.1 Web Mining and Information Retrieval (IR)

IR is the automatic retrieval of all relevant documents while at the same time retrieving as few of the non relevant as possible [50]. Some have claimed that resource or document discovery on the web is an instance of web content mining and the others associate web mining with intelligent IR. Actually IR has the primary goal of indexing text and searching for useful documents in a collection. Research in IR includes modelling, document classification and categorization, user interfaces, data visualization and filtering [6]. The task that can be considered to be an instance of web mining is web document classification or categorization which could be used for indexing. Viewed in this respect web mining is part of the information retrieval process.

### 2.2.2 Web Mining and Information Extraction (IE)

IE has the goal of transforming a collection of documents usually with the help of an IR system into information that is more readily digested and analyzed [17]. IE aims to extract relevant facts from the documents [46] while IR aims to select relevant documents. While IE is interested in the structure or representation of a document, IR views the text in a document just as a bag of unordered words [61]. Thus, in general IE works at a finer granularity level than IR does on the documents. Building IE systems manually is not feasible and scalable for such a dynamic and diverse medium such as web contents [42]. Due to this nature of the web most IE systems focus on specific web sites to extract. Others use machine learning or data mining techniques to learn the extraction patterns or rules for web documents semi-automatically or automatically. Within this view web mining is used to improve web IE.

## 2.3 Web Content Mining

Web content mining describes the discovery of useful information from the web contents or data or documents [37]. 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 [39]:

1. The amount of data or information on the web is huge and still growing rapidly. Web data is also easily accessible.

2. The coverage of web information is wide and diverse. One can find information about almost everything on the web.

3. Data of all types exist on the web e.g. structured tables, text and multimedia data.

4. Information on the web is heterogeneous. Multiple web pages may present the similar information using completely different formats or syntaxes which makes integration of information a challenging task.

5. Much of the web information is semi-structured due to the nested structure of HTML code and the need of web page designer is to present information in a simple and regular fashion to facilitate human viewing and browsing.

6. Much of the web information is linked. There are links among pages within a site and across different sites. These links serve as information to organization tool and also as indication of trust or authority in the linked pages and sites.

7. Much of the Web information is redundant. The same piece of information or its variations may appear in many pages or sites. This property has been explored in many Web data mining tasks

8. The web is noisy. A web page typically contains a mixture of many kinds of information e.g. main content, advertisements, navigation panels and copyright notices. For a particular application only part of the information is useful and the rest are noises.

9. The web consists of surface web and deep web. Surface web is composed of pages that can be browsed using a normal web browser. Surface web is also searchable through popular search engines. Deep web is mainly composed of databases that can only be accessed through parameterized queries using query forms.

10. The web is also about services. Many web sites and pages enable people to perform operations with input parameters i.e. they provide services.

11. 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.

12. 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 in the tables or database generated HTML pages. So two following main approaches in web content mining arise

(1) Unstructured text mining approach
(2) Semi-Structured and Structured mining approach.

### 2.3.1 Unstructured Text Data Mining (Text Mining)

Much of the web content data is unstructured text data [21]. The research around applying data mining techniques to unstructured text is termed knowledge discovery in texts (KDT) [22] or text data mining [32] or text mining [55]. 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 [48].

### 2.3.2 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 is also easier to extract compared to unstructured texts. Semi-structured data is a point of convergence [18] 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) [25]. 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 labelled 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 [14]. 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 [37].

## 2.4 Web Structure Mining

With the growing interest in web mining the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [24] which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming and graph mining. There is a potentially wide range of application areas for this new area of research including Internet. The web contains a variety of objects with almost no unifying structure with differences in the authoring style and content much greater than in traditional collections of text documents which resulted in a new form known as web structure mining.

Web structure mining [11, 27, 30, 35] 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 and links are in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor text [26]. The appropriate handling of the links could lead to potential correlations and then improve the predictive accuracy of the learned models [26]. The challenge for web structure mining is to deal with the structure of the hyperlinks within the web itself. Two algorithms that have been proposed to lead with those potential correlations are HITS [36] and PageRank [43].

### 2.4.1 Hyperlink-induced topic search (HITS)

In HITS concept [36], Kleinberg identified two kinds of pages from the web hyperlink structure authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to Kleinberg, Hubs and authorities exhibit what could be called a mutually reinforcing relationship. A good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

### 2.4.2 PageRank

L. Page and S. Brin [9, 43] proposed the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. In their approach Brin and Page extended the idea of simply counting in-links equally normalizing by the number of links on a page. The Page Rank algorithm is defined as [9]: 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:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

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.

## 2.5 Web Usage Mining

Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the web. The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services, to personalize the Web portals [19] or to improve the web structure and web server performance [47]. In this case, the mined data are the log files which can be seen as the secondary data on the web where the documents accessible through the web are understood as primary data.

Five major steps followed in web usage mining are

1. Data collection – Web log files, which keeps track of visits of all the visitors
2. Data Integration – Integrate multiple log files into a single file
3. Data pre-processing – Cleaning and structuring data to prepare for pattern extraction

4. Pattern extraction – Extracting interesting patterns
5. Pattern analysis and visualization – Analyze the extracted pattern
6. Pattern applications – Apply the pattern in real world problems

There are three types of log files that can be used for web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. It is problematic to collect all the information from the client side. Thus most of the algorithms work based only on the server side data. Some commonly used data mining algorithms for web usage mining are association rule mining, sequence mining and clustering [14].

Web usage mining is one of the prominent research areas due to these following reasons:

a) One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behaviour of the user and to make prediction about desired pages. Thus personalization for a user can be achieved through web usage mining.

b) Frequent access behaviour for the users can be used to identify needed links to improve the overall performance of future accesses. Pre-fetching and caching policies can be made on the basis of frequently accessed pages to improve latency time.

c) Common access behaviours of the users can be used to improve the actual design of web pages and for making other modifications to a web site.

d) Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

### 2.5.1 Log Formats

W3C maintains a standard format for web server log files but other proprietary formats exist. For example IIS provides six different log file formats which are used to track and analyze information about IIS-based sites and services such as

1. W3C Extended Log File Format
2. W3C Centralized Logging
3. NCSA Common Log File Format
4. IIS Log File Format
5. ODBC Logging

6. Centralized Binary Logging

In addition to the six available formats custom log file format can also be configured. A log file in the W3C extended format contains a sequence of lines containing ASCII characters. Each line may contain either a directive or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by white space. If a field is unused in a particular entry dash "-" marks the omitted field. Directives record information about the logging process itself. Lines beginning with the # character contain directives. The following directives are defined in the W3C Extended format

**1. Microsoft Internet**

Information Server (IIS):

#Software: Microsoft Internet Information Server 4.0

#Version: 1.0

#Date: 1998-11-19 22:48:39

#Fields: date time c-ip cs-username s-ip cs-method cs-uri-stem cs-uri-query scstatus sc-bytes cs-bytes time-taken cs version cs(User-Agent) cs(Cookie) cs(Referrer)

c Client

s Server

r Remote

cs Client to Server.

sc Server to Client.

sr Server to Remote Server, this prefix is used by proxies.

rs Remote Server to Server, this prefix is used by proxies.

x Application specific identifier.

**2. Common Log Format**

LogFormat "%h %l %u %t \"%r\" %s %b"

**3. Combined Log Format:**

Another commonly used format string is called the Combined Log Format. It can be used as follows.

"%h %l %u %t \"%r\" %s %b\"%{Referer}i\" \"%{User-agent}i\""

This format is exactly the same as the Common Log Format, with the addition of two more fields. Each of the additional fields uses the percent-directive %{header}i, where header can be any HTTP request header.

### 2.5.2 Web Usage Mining and Related Work

Mobasher [15] proposed that the web usage mining process can be divided into two main parts. The first part includes the domain dependent processes of transforming the web data into suitable transaction form. This includes pre-processing, transaction identification and data integration components. The second part includes some data mining and pattern matching techniques such as association rule and sequential patterns. In the absence of cookies or dynamically embedded session Ids in the URIs the combination of IP address can be used as a first pass estimate of unique users.

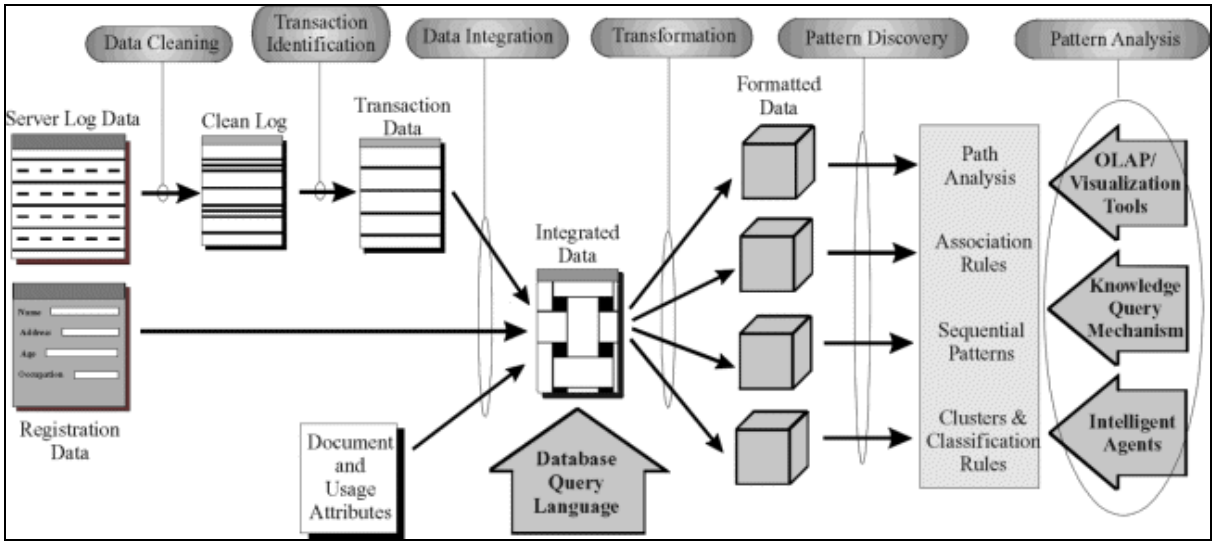


**Figure 8: Web Usage Mining Architecture [40]**

#### 2.5.2.1 Pre-processing

Data pre-processing has a fundamental role in web usage mining applications. [21] Notices that even if pre-processing techniques are widely used in web usage mining the literature on this topic is still quite limited and that the most complete reference on pre-processing dates back to 1999. Fundamental methods of data cleaning and preparation have been well studied by Srinivasa [51]. The pre-processing of web logs is usually complex and time demanding.

It comprises of four different tasks

(ii) The data cleaning

(iii)The identification and the reconstruction of users sessions

(iv) The retrieving of information about page content and structure

(v) The data formatting

#### 2.5.2.2 Web Usage Mining and Pattern Discovery

Analysis of user behavior has two aspects one concerning the interests of the users and the accessed information, the other concerning the way of accessing the information. The first aspect is solved by techniques for the construct of user profiles and is not specific to the web usage while the second one is addressed by analyzing web server logs which falls in the field of the web usage mining. In his paper, Spiliopoulou [53] proposed the exploitation of mining technology to discover access patterns with interesting statistical properties and presented Web Utilization Miner (WUM) a tool designed for this purpose. The mining model of WUM is in two aspects. First, it predicts that the importance indicators in user behavior go far beyond than frequent access to some pages such that the pattern discovery can be done in the statistical domain but also supports the subjective specification. Second, by processing aggregated sequences and applying optimization steps during the mining process the high performance can be achieved.

Buchner [10] proposed a new approach in the form of process is proposed to find marketing intelligence from Internet data. An n-dimensional web log data cube is created to store the collected data. Domain knowledge is incorporated into the data cube in order to reduce the pattern search space. They proposed an algorithm to extract navigation patterns from the data cube. The patterns conform to pre-specified navigation templates whose use enables the analyst to express his knowledge about the field and to guide the mining process. This model does not store the log data in compact form and that can be major drawback when handling very large daily log files.

Masseglia [40] proposed an integrated tool for mining access patterns and association rules from log file. The techniques implemented pay particular attention to the handling of time constraints such as the minimum and maximum time gap between adjacent requests in a pattern. The system provides a real time generator of dynamic links which aimed at automatically modifying the hypertext organization when user navigation matches a previously mined rule.

Mobasher [40] gave a new web mining technique WEBMINER which offered transaction models to extract useful information from the server logs. Spiliopoulou and Faulstich [53] collected the individual routing paths into an aggregated tree and pruned the uninteresting

patterns considering only those patterns which had desired characteristics. H. Yao [63] proposed a foundational approach to mining item set utilities from databases. This approach allows user preferences of item set as subjective values. The objective value of an item is defined according to the information stored in a transaction i.e. the quantity of the item sold in the transaction. From very large logs, Z. Chen [12] proposed two effective algorithms for finding maximal forward references longest sequences of Web pages visited by a user without revisiting some previously visited page in the sequence and their performance is relatively analyzed. An efficient web traversal pattern mining algorithm based on suffix array is given by T. Jing [32]. Yen [64] presented the modified incremental data mining algorithm for discovering web traversal patterns when the user sequences are inserted into and deleted from original database. This algorithm uses lattice structure to keep the previous mining results such that just new candidate sequences need to be computed. Hence, the web traversal pattern can be obtained rapidly when the traversal sequence database is updated. But it is unsuccessful when web site structure is changed. Zhou [67] proposed high utility path traversal pattern mining, which introduces the concept of utility into path traversal pattern mining model. A utility-based algorithm for web path traversal was improved by C. F. Ahmed [2]. They used a pattern growth sequential mining to prune a huge number of candidates. It effectively divides the search space by small projected databases recursively using the divide and conquers technique. Therefore, it saves several scanning of the whole database which is required by the exiting algorithm.

Chen [12] gave two algorithms for determining web traversal patterns: FS (full-scan) and SS (selective scan)**.** To describe algorithm FS, first get the basic ideas of the DHP algorithm. DHP utilizes a hashing technique and so is very efficient in the generation of candidate item-set ($C_k$). Also DHP assists with pruning techniques which reduces the database size considerably. Lk denotes the set of all large k-references and ck is a set of candidate k-references, ck is usually a superset of Lk**.** By scanning through the database $D_F$**,** FS gets L1 and makes a hash table (H2**)** to count the number of occurrences of each 2-reference. Similarly to DHP, starting with k = 2 FS generates ck based on the hash table count obtained in the previous pass determines the set of large k-references reduces the size of database for the next pass and makes a hash table to determine the candidate (k + 1)-references. Database size can decrease significantly with each pass. While devising the algorithm it was found that it is better to obtain the Ck from Lk-1*Lk-1 rather than using hash able to generate Lk. To

count the occurrences of each A-reference in ck to determine Lkj we need to scan through a trimmed version of database DF**.**

From the set of maximal forward references it is determined among k-references in Ck large k-references. After the scan of the entire database, those k-references in ck with count exceeding the threshold become Lk**.** If Lk is non-empty, the iteration continues for the next pass which means pass k + 1. Same as in DHP, every time when the database is scanned the database is trimmed by FS to improve the efficiency of future scans.

Algorithm SS is similar to algorithm FS in that it also employs hashing and pruning techniques to reduce both CPU and I/O costs but is different from the latter. In algorithm SS by properly utilizing the information in candidate references in prior passes it is able to avoid database scans in some passes thus further reducing the disk I/O cost. Recall that algorithm FS generates a small number of candidate 2-references by using a hashing technique. In fact, this small C2 can be used to generate the candidate 3-references.

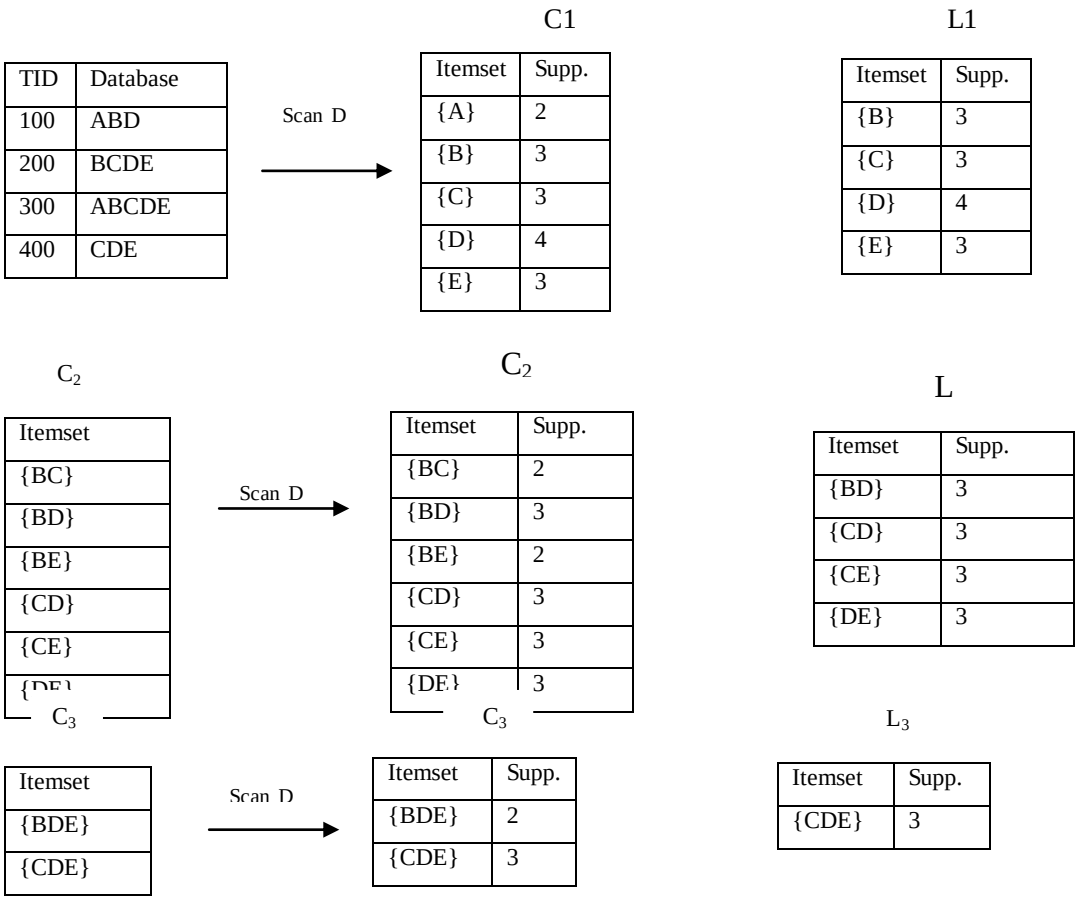


**Figure 9: Example Execution of Full Scan [12]**

### 2.5.3 Pattern Extraction

The existing research work done by different authors can be categorized into

a) Association Rule Mining (ARM)
b) Clustering
c) Classification

**1. Research Work on Association Rule Mining**

Ming-Syan [12] proposed a new data mining algorithm that involves mining path traversal patterns in a distributed information-providing environment where documents or objects are linked together to facilitate interactive access. Their solution procedure consists of two steps. First, derive an algorithm to convert the original sequence of log data into a set of maximal forward references. Second derive another algorithm to determine the frequent traversal patterns i.e. large reference sequences from the maximal forward references obtained. Jianhan Zhu [68] applied the Markov chains to model user navigational behaviour. They proposed a method for constructing a Markov model of a web site based on past visitor behaviour. Then the Markov model is used to make link predictions that assist new users to navigate the Web site. WANG Tong [56] offers an improved algorithm based on the original AprioriAll algorithm. The new algorithm adds the property of the User-id during the every step of producing the candidate set and every step of scanning the database by which to decide whether an item in the candidate set should be put into the large set which will be used to produce next candidate set. Hengshan Wang [59] introduced two prevalent data mining algorithms – Fpgrowth and PrefixSpan into WUM. Maximum Forward Path (MFP) is also used in the web usage mining model during sequential pattern mining along with PrefixSpan so as to reduce the interference of "false visit" caused by browser cache and raise the of mining frequent traversal paths. Sandeep Singh Rawat [49] proposed a custom-built apriori algorithm which is based on the old Apriori algorithm, to find the effective pattern analysis.

**Table 2: Algorithm on Association Rule Mining [45]**

| Algorithm Used | Authors | Year |
|---|---|---|
| **Maximalforward references** | Ming-Syan Chen,Jong Soo Park,Philip S.Yu | 1998 |
| **Markov Chains** | Jianhan Zhu,Jun Hong and John G. Hughes | 2002 |
| **Improved AprioriAll** | Wang tong,HE Pi-lian | 2005 |
| **Fpgrowth and prefixspan** | Hengshan Wang,Cheng Yang,Hua Zeng | 2006 |
| **Custom Built APRIORI Algorithm** | Sandeep Singh Rawa, Lakshami Rajamani | 2010 |

**2.Research Work on Clustering**

Paola Britos [8] described the capacity of use of Self Organized Maps, kind of artificial neural network, in the process of Web Usage Mining to detect user's patterns. The process details the transformations necessaries to modify the data storage in the Web Servers Log files to an input of Self Organized Maps. Mehrdad Jalali [31] presented an approach which is based on the graph partitioning for modelling user navigation patterns. In order to mining user navigation patterns, they establish an undirected graph based on connectivity between each pair of the web pages and also proposed novel formula for assigning weights to edges of the graph. Kobra Etminani [20] applied ant-based clustering algorithm to pre-processed logs to extract frequent patterns for pattern discovery and then it is displayed in an interpretable format. N. Sujatha [54] has proposed a new framework to improve the web sessions' cluster quality from k-means clustering using Genetic Algorithm (GA). Initially a modified k-means algorithm is used to cluster the user sessions. The refined initial starting condition allows the iterative algorithm to converge to a "better" local minimum. And in the second step, they have proposed a GA based refinement algorithm to improve the cluster quality.

**Table 3: Algorithm on Clustering [45]**

| Algorithm Used | Authors | Year |
|---|---|---|
| **Self Organized Maps** | Paola Britos,Damian Martinelli,Hernan merlino,Ramon Garcia-Martinez | 2007 |
| **Graph Partitioning** | Mehrdad Jalali,Norwati Mustapa,Ali Mamat,Md. NAsir B Sulaiman | 2008 |
| **Ant-based** | Kobra Etminani,Mohammad-R. Akbarzadeh-T,Noorali raeeji yaneshsari | 2009 |
| **K-mean with genetic Algorithm** | N. Sujatha, K. Iyakutty | 2010 |

**3. Research Work on Classification**

Mahdi khosravi [34] proposed a novel approach for dynamic mining of user's interest navigation patterns using naive Bayesian method.

**Table 4: Algorithm on Classification [45]**

| Algorithm Used | Authors | Year |
|---|---|---|
| **Naive Bayesian** | Mahdi khosravi, Mohammad J. Tarokh | 2010 |

## 2.6 Web Usage Mining Applications

### 2.6.1 Pre-fetching and caching:

The results produced by web usage mining can be exploited to improve the performance of web servers and web-based applications. Typically, web usage mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time.

### 2.6.2 Support to the design:

Usability is one of the major issues in the design and implementation of web sites. The results produced by web usage mining techniques can provide guidelines for improving the design of web applications. Some studies use stratograms to evaluate the organization and the efficiency of web sites from the user's viewpoint and some exploit web usage mining techniques to suggest proper modifications to web sites. Adaptive web sites represent a further step. In this case, the content and the structure of the web site can be dynamically reorganized according to the data mined from the user's behaviour.

### 2.6.3 E-commerce:

Mining business intelligence from web usage data is dramatically important for ecommerce web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of web usage mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales and customer departure.

### 2.6.4 Personalization of web content

Web usage mining techniques can be used to provide personalized web user experience. For instance it is possible to anticipate the user behaviour in real time by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users. Personalized Site Maps are an example of recommendation system for links proposed an adaptive technique to reorganize the product catalogue according to the forecasted user profile [23].

## 2.7 Software

There are many commercial tools which perform analysis on log data collected from web servers. Most of these tools are based on statistical analysis techniques, while only a few products actually exploit data mining techniques. With respect to web mining commercial tools it is worth noting that since the review the number of existing products almost doubled. Companies which sold web usage mining products in the past have disappeared (e.g., Andromeda's Aria); others have been bought by other companies. In most cases, web usage mining tools have become part of integrated Customer Relation Management (CRM)

solutions for ecommerce. In some cases, web usage mining tools are simple web log analyzers. One of the software developed in a research environment, WUM, appears to be at an interesting maturity level; WUM has currently reached the version 7.0. All these products provide data mining tools for user profiling. Lumios Recognition is a complete e-business solution for Web analytics. Recognition provides different approaches to track users behaviour: (i) cookies and web logs from the web server side, (ii) Javascript and Java Applets from the client side, and even more advanced techniques including (iii) packet sniffer, proxy server logging. NetTracker, by Sane Solutions is a family of tools for web analytics. The advanced editions are suited for e-business analysis and allow integrations with CRM solutions. Sane also offers a service for remote web analysis based on NetTracker solution. Elytics AnalysisSuite integrates web log data with data from the client side and combines them with different user metrics. E.piphany E.6 is a complete CRM solution that includes tools for web log and transaction log analysis. NetIQ's WebTrend's Log Analyzer Series provides web traffic reporting for the small business. Thanks to the join with NetGenesis, SPSS offers a complete web analytic solution that integrates NetGenesis web analytics and SPSS Clementine in SPSS Predictive web Analytics. WebSideStory offers HitBox, a suite of products that includes different solutions for tracking user's behaviour, ranging from simple statistics to very complete solutions for big companies including e-business oriented analysis. IBM provides a family of on-line services called Surfaid Analytics that allow the filtering and integration of web log, commerce server, registration and other data into a database of profiles. These profiles are subsequently used for producing standardized reports as well as performing OLAP operations. Data can be collected both from web logs and using Javascript based on techniques developed by IBM. Quest's Funnel web analyzer is available both with a free license and with a commercial license. The former version offers basic statistical reports and the latter one offer advanced features based on data mining techniques. WebHound is the Web analytic tool developed by SAS. It extracts information from Web logs and performs click-stream analysis. SAS offers also an on-line web analytic service, SAS IntelliVisor, developed for e-business oriented analysis of web sites. Different versions of the service for three different market areas are available: Financial Services, Pharmaceuticals, and Retail. Megaputer WebAnalyst is a complete application server that integrates with existing web servers to provide advanced web usage analysis. The application server acts as a sort of proxy between the clients and the actual web server. The application server instead of the web server intercepts the page requests and after a processing step it forwards the page requests to the actual web server. The page forwarded from the server is again intercepted by the

application server before being sent to the client who issued the page request. Prudsys ECOMMINER is oriented toward e-business and integrates data from web logs with server side transactional data. It has been developed for two specific e-commerce platforms: InterShop and Logisma Business Webstore [23].

# Chapter 3 Problem Formulation

As the popularity and vastness of World Wide Web is increasing with each day so is the importance of web mining. With the increasing number of websites and web users, web data is being collected and stored by the server as web server data composed with different fields. For performing this task it is necessary to collect a good amount of data for analysis before coming to any productive conclusion. As the web server data tends to be too large there is a need to devise an efficient algorithm to first extract useful data and then mine it to get patterns which are helpful for the website. Many different algorithms have been proposed and developed to increase the efficiency of mining frequent patterns with different strength and weaknesses for different sizes of data.

## 3.1 Motivation

Studies of pattern mining have been acknowledged in an environment which provides us with interactive access say World Wide Web because of the broad applications it has in this field.
By the analysis of the web data one can extract various information like user surfing behaviour which can help in user profiling, web site designs and making better business and marketing decision making our website more popular and user friendly. As already known the size of the web data is too large and extracting useful information from this data requires devising of an algorithm which is efficient, scalable and can detect patterns which can be used in various ways.

## 3.2 Gap Analysis

- The existing algorithms produce the frequent patterns on the basis of minimum support
- The apriori variations algorithms like DHP tries to reduce candidate item-set and prune the database size in each step.
- DHP works well at early stages and performances deteriorates in later stage and also results in I/O overhead.
- The apriori performs better with small database size.
- DHP requires big memory space and repeated database scans due to its candidate generation and test strategy

## 3.3 Problem Statement

Let P = {p1; p2; .....; pn} be the complete set of web pages. Let W be the web log to be analysed. Let $D_F$ be the task relevant database which contains the maximal forward references obtained from algorithm MF. A maximal forward reference is the maximum path covered by the user for a particular website. After obtaining the maximal forward references, the proposed algorithm is applied so as to obtain the frequently occurring patterns for the website known as the large reference. The details will be given in the next section.

The aim of this study is to

(1) Examine the features of the web log data

(2) Obtain the maximal forward references

(3) Obtain the large references.

As these information giving services are expanding with each day with fierce competition to face with analysis of user access patterns is a necessity so as to come to qualitative conclusions with meaningful patterns to end with and thus providing services which help in achieving user satisfaction. The web has a natural graph structure: pages in general are linked via hyperlinks. Each page can be represented by a node and the hyperlinks can be represented by the arrows linking the nodes as shown in figure 10 where the red circles represent the nodes and the arrows represent the link between the pages. When a user surfs the web, s/he may move forward along the graph via selecting a hyperlink in the current page. S/he may also move backward to any page visited earlier in the same session via selecting a backward icon. A forward reference may be understood as the user looking for her/his desired information. A backward reference may mean that the user has found her/his desired information and is going to looking for something else. A sequence of consecutive forward references may indicate the information for which the user is looking. A maximal forward reference is defined as the longest consecutive sequence of forward references before the first backward reference is made to visit some previously visited page in the same session. Thus, the last reference in a maximal forward sequence indicates a content page that is desired by the user. Under such understanding, when a user searches for desired information, his/her information needs can be modelled by the set of maximal forward references that occurred during his/her search process.

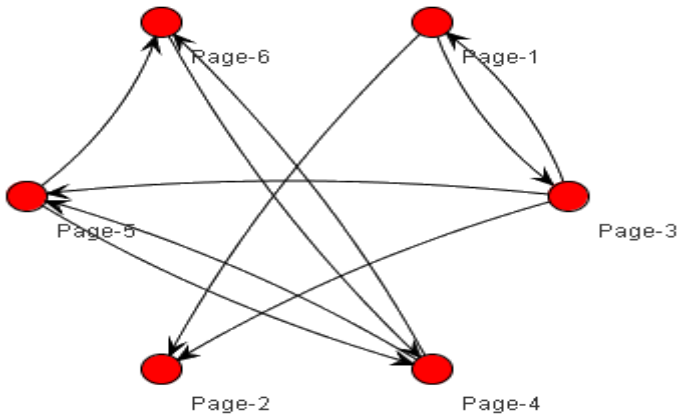


**Figure 10: web graph**

Web logs are the source data of web usage mining. Usually a web server registers in each web log entry the IP address from which the request originated, the URL requested, the date

and time of accesses, the page references and the size of request data. To analyse web logs, the first stage is to divide web log records into sessions where a session is a set of page references of one source site during one logical period. Practically a session can be seen as a user starting visiting a web site, performing work and then leaving the web sites.

### 4.1 Algorithm for User Access Pattern

Now, moving on to the explanation of the algorithm (MF) used to obtain the maximal forward references described in the next section with help of an example explained below. As we said that the web pages also known as the nodes are here represented using alphabets which are used to represent the user accessed path.

Suppose the log contains various user traversed path say : {A,B,C,D,E,F,G,H,I,J} as in Figure 12 . Now, after applying the algorithm MF we get the following maximal forward references as output that have been accessed by the user {ABCD, ABEFG, ABEH, AIJK}.After obtaining the maximum traversed patterns the frequently occurring substrings from the maximal forward references are obtained. These frequently occurring substrings are known as large reference. A large reference sequence is a traversal pattern which appears number of times.

The overall procedure is as follows:

Step 1: We determine the maximal forward reference using algorithm MF.

Step 2: After this we extract the large reference sequence using the algorithm RF.

After performing the above two steps the large references which are frequently occurring sequences are obtained.

## 4.2 Algorithm for Maximal Forward Reference

In this section the maximal forward reference using the algorithm MF are obtained. Maximal forward references are those strings which tell us the maximum length till which the user accessed a particular website before it traversed back to a previously accessed webpage. For this we have considered an already proposed algorithm MF. A server log consist of various fields as Host,rfc931,username,date:time,request,statuscode,bytes,referrer,user_agent. Now the pair of source and destination field from the server log is considered and this pair is sorted on the account of user id's for each user where we order the (source,destination) pair by time so as to obtain the maximal forward references. The output obtained from MF is stored in a

database $D_F$ as this output will act as input to the algorithm for reference scan giving us large reference sequence.

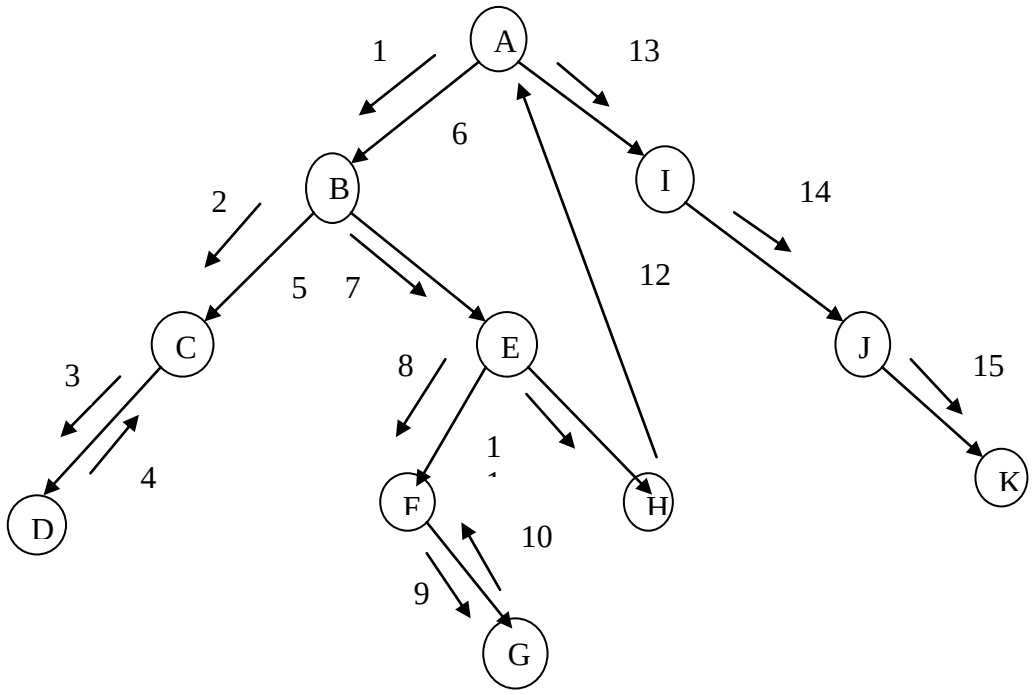


**Figure 11: User Traversal Path**

### 4.2.1 Finding Maximal Forward References

Algorithm *MF:*

Step 1: Set *i* = 1 and string *Y* to null for initialization,
where string *Y* is used to store the current
forward reference path.
Also, set the flag *F* = 1
to indicate a forward traversal.

Step 2: Let *A* = *si* and *B* = *di.*
If *A* is equal to null then
/* this is the beginning of a new traversal */
Begin
Write out the current string *Y* (if not null)
to the database *DF;*
Set string *Y* = *B;*
Go to Step 5.
end

Step 3: If *B* is equal to some reference (say the j-th
reference) in string *Y* then
/* this is a cross-referencing back to a previous
reference */
begin
If *F* is equal to 1 then write out string *Y* to
database $D_F$;
Discard all the references after the j-th one in string Y
*F =0 ;*
Go to Step 5.
end

Step 4: Otherwise, append B to the end of string Y.
/* we are continuing a forward traversal */
If F is equal to 0, set F = 1.

Step 5: Set i = i+l. If the sequence is not completed scanned then go to Step 2.

**Table 5: An Example Execution of MF**

| **Sequence** | **String Y** | **Database $D_F$** |
|---|---|---|
| 1 | A | - |
| 2 | AB | - |
| 3 | ABC | - |
| 4 | ABCD | - |
| 5 | ABC | ABCD |
| 6 | AB | - |
| 7 | ABE | - |
| 8 | ABEF | - |
| 9 | ABEFG | - |
| 10 | ABEF | ABEFG |
| 11 | ABEH | - |
| 12 | A | ABEH |
| 13 | AI | - |
| 14 | AIJ | - |
| 15 | AIJK | - |

Now, analyzing the above given table as per the traversal pattern in figure 1, it can be seen that there is a backward traversal in sequence 5 resulting in ABCD as the first maximal forward reference written to database (as in step 3). The user traversal graph is being scanned till the backward reference is encountered and then the final sequence is obtained which is

written to string Y and ultimately fed and stored to the database $D_{F.}$ The final sequence from the figure 1 is {ABCD, ABEFG, ABEH, AIJK}.

## 4.3 Algorithm for Finding Large Reference

After obtaining the maximal forward references in the database $D_{F,}$ the large references are derived which are the most frequently occurring patterns that are traversed by the user many a times. For a sequence to qualify as large reference sequence it needs to have a minimum support which is a user defined value changing with the size of the database.

### 4.3.1 Finding Large Reference

The website is built in a hierarchical structure starting with the initial webpage and then the rest of the webpage's are connected via the hyperlink. The information mostly lies on the last level of the tree structure with the previous levels containing the folders and the sub folders. Keeping this fact in mind an algorithm is devised known as reference scan where starting from the last level the move towards the upper levels is made so as to get large reference sequences. The algorithm is reducing the database scan and is making search more efficient and fast. The proposed algorithm is presented below and is illustrated with an example.

**Algorithm Reference Scan (RS)**

Input – Array of structure of TID, seq, minsupp

Output –large reference.

1. Find max i.e. maximum length of seq from inputs.
2. Repeat for flength from max to 1
3. Find number_of_sequences from flength
4. if(number_of_sequences>minimumsupport)
5. continue;
6. else
7. Create subsets of all web-pages in forward direction of length equal to flength.
8. Compare all subsets with all input to find occurrence of subsets
9. Get subset with maximum occurrence and occurrence>minimum support
10. End if-else
11. End loop

#### 4.3.1.1 Illustrative Example

The algorithm proposed in this section consists of various steps which will be explained one by one. Now known that web has a tree like structure with pages being represented as nodes denoted by alphabets and the hyperlinks represented by arrows. The algorithm starts by taking an example consisting of two fields, transaction id (tid) and the nodes that have been accessed by user during that transaction.

**Table 6: Example execution of RS**

| Tid | Database |
|---|---|
| 100 | ABCD |
| 200 | ABEFG |
| 300 | ABEH |
| 400 | AIJK |

The above table is the database output consisting of maximal forward reference obtained from algorithm MF which acts as input as an example execution explanation for the algorithm. This table contains the user accessed path as :{ A, B, C, D, E, F, G, H, I, J, K}

**Step 1**

In the first step of the algorithm the number of nodes denoted by alphabets here that are accessed during each transaction by the user are counted say for instance in the first transaction with Tid 100 there are 4 nodes that are being accessed, in next transaction with Tid 200 there are 5 nodes again, so on and forth. After recording the Flength of each transaction next step is executed.

**Table 7: flength count**

| TID | Database | Flength |
|---|---|---|
| 100 | ABCD | 4 |
| 200 | ABEFG | 5 |
| 300 | ABEH | 4 |
| 400 | AIJK | 4 |

**Step 2**

After step 1 where flength value is obtained for each transaction the next involves obtaining the number of occurrences of each flength implying counting how many times the string with flength value 1 occurs as it can be seen that there is no transaction containing only 1 node so

its value is 0. Now, check the occurrence of flength value 2 and as it can be seen in table 4 there is only 1 such transaction with tid 400 that contains 2 nodes. After this count the occurrences of flength value 3 which as can be seen in table 4 is two transactions containing 3 nodes till the value of flength reaches 11. As in all there are 11 nodes that can be accessed by the user.

**Table 8: Occurrence of flength**

| **Count(Flength)** | **Value** |
|---|---|
| Count(1) | 0 |
| Count(2) | 0 |
| Count(3) | 0 |
| Count(4) | 3 |
| Count(5) | 1 |
| Count(6) | 0 |
| Count(7) | 0 |
| Count(8) | 0 |
| Count(9) | 0 |
| Count(10) | 0 |
| Count(11) | 0 |

**Step 3**

As the occurrence of different values of flength has been obtained in the above step, now the flength value count will be compared with minimum support value. A minimum support value is the one for which the count of flength has to be greater so as to qualify to be analysed further. The minimum support value keeps on increasing and is user defined. First of all the maximum flength value is taken and its occurrence is also considered and compared with minimum support value. Now starting with 11 its occurrence is 0 and is compared with minimum support value 1 (assumed). After this the count for occurrence of flength value keeps on reducing and each statement is false until we reach count for flength value 5 which is 1,after comparing this value with minimum support the value is still false. Now reduce the count, occurrence of value 4 is 3 which is higher than the minimum support.

**Step 4**

After obtaining the flength value which has more occurrences than the minimum support, create the subsets for the same order say in this case subset of the order 3 will be created in forward direction of the traversal: {A, B, C, D, E ,F, G, H, I, J, K}. the subsets will be {ABC,ABD,ABE,ABG,ABH,ABI,ABJ,ABK,ACD,ACE,ADE,AEG,AGH,AHI,AIJ,AJK,ABF,AEF,AFG,ACF,ADF,AFH,AFI,AFJ,AFK,BCD,BCE,BCG,BCH,BCI,BCJ,BCK,BDE,BEG,BGH,BHI,BIJ,BJK,BDE,BDG.BDH,BDI,BDK,BDJ,BEH,BEI,BEJ,BEK,BGH,BGI,BGJ,BGK,BHJ,BHK,BCF,BFG,BEF,BDF,CDE,CDG,CDH,CDI,CDJ,CDK,CEG,CGH,CHI,CIJ,CJK,CDF,CEF,CFG,CFH,CFI,CFJ,CFK,DEG,DEH,DEI,DEJ,DEK,DGH,DHI,DIJ,DEF,DFG,DFH,DFI,DFJ,DHK,DJK,EHI,EIJ,EJK,GIJ,GJK,EGH,EGI,EGJ,EGK,EFG,EFH,EFH,EFI,EFJ,EFK,GHI,GHJ,GHK,HIJ,HJK,CEG,CEH,CEI,CEJ,CEK,CGI,CGJ,CGK,IJK}.

**Step 5**

Each of the traversal patterns subset created above will be taken and matched to the database one by one. Also the occurrence of each and every subset will be counted in the database and the subset with a value greater than minimum support qualifies to be the final useful pattern for the website user and the owner. Now after comparing the above given subsets with our database in table the pattern ABE occurs twice and is the final pattern and the result.

**Tabular form comparison between the new proposed algorithm and the previous algorithm:**

**Table 9: Comparison of Proposed and Existing Algorithm**

| FACTORS | Full/Selective scan | Reference scan |
|---|---|---|
| **Consecutive Patterns** | This algorithm only works on consecutive patterns | This algorithm also considers links which are not consecutive. |
| **Database Scan** | With each pass database scan need to be done | Database scan needs to be performed at step 4 where the subsets are compared with the database. |
| **Database size** | It works well only on small size database; with large size of database it is not that efficient. | It works well both on small and large databases but its performance reduces with large size databases |
| **I/O Overhead** | This algorithm incurs us high I/O costs | This algorithm does not bring us high I/O overheads as there are less database scans. |
| **Database** | As this algorithm uses DHP technique with each pass the size of the database reduces | There is no reduction in the size of database. |

# Chapter 6 Conclusion

The web is a very essential means to carry out business and commerce. So the design of web pages is highly essential for the system managers and web creators. These characteristics have huge impact on the number of users who access the page. Therefore there is a need for algorithm that examines the data of server log file for identifying the navigation pattern as it is essential for good understanding of the data preparation technique and pattern discovery method. Web usage mining systems will offer those techniques stated. Several techniques have been proposed by different researchers for pattern mining with its own merits and demerits. This chapter summarizes the work done and the future scope.

## 6.1 Conclusions

In this thesis various frequent pattern mining algorithm have been seen and finally so as to reduce the problem of repeated database scan a new approach has been proposed. For extracting useful user accessed patterns the maximal forward reference have been obtained by using algorithm MF. A maximum forward reference is the maximum point till which a user can access a website and as soon as the user gets a backward reference it is considered that the user has obtained the required content. In the second step the new approach is applied and the large references are obtained on the basis of user defined minimum support. A large reference is a frequently occurring pattern in the database. The new approach works on the basis that the information for website lies mainly on the leaf node if the web is represented in the form of a tree structure. The new approach thus starts by doing backward scan so as to obtain the useful and meaningful patterns. The procedure has been explained with examples and then the comparison between the apriori based algorithm and the new approach.

## 6.2 Future Scope

There is scope of research done in the work that has been done in the thesis:

1. This algorithm has been applied to small and large databases and can be applied to very large databases
2. This algorithm has been applied to environment providing us interactive access like World Wide Web but this algorithm can be applied to transactional database and check its performance there.

# References


[1] About Web, available:http://www.technicalsymposium.com/web_mining_notes.html.

[2] Ahmed C. F., Tanbeer S. K., Jeong B.S., Lee, Y.K., "Efficient mining of utility-based web path traversal patterns", *in Proceedings of 11th International Conference on Advanced Communication Technology (ICACT'09)*, pp. 2215-2218, 2009.

[3] Available: http://www.internetworldstats.com/stats.html.

[4] Available: http://www.domaintools.com/internet-statistics/2.

[5] Bayir M.A., "a new reactive method for processing web usage data", M.S. thesis, The graduate school of natural and applied sciences, Middle east technical university, 2006.

[6] Baeza-Yates R. and Berthier e.,"Modern Information Retrieval", Addison-Wesley London publishing company, 1999.

[7] Singh B. and Singh H.K., "Web data mining research: a survey", *in Proceedings of IEEE International Conference on Computational Intelligence and Computer Research (ICCIC)*, pp.1-10, December 2004.

[8] Britos P., Martinelli D., Merlino H. and García-Martínez R, "Web Usage Mining Using Self Organized Maps", *in International Journal of Computer Science and Network Security,* vol.7 ,no.6, June 2007.

[9] Brin, S. and Page,L, "The Anatomy of a Large-scale Hyper-textual Web Search Engine", *in Proceedings of the Seventh International World Wide Web Conference*, 1998.

[10] Büchner A.G., Baumgarten M., Anand S.S., Mulvenna M.D. and Hughes J. G., "Navigation Pattern Discovery from Internet Data", *in Proceedings of WebKDD*, pp. 74-91, 1999

[11] Chakrabarti S., Dom B., and Indyk P., “Enhanced hypertext categorization using hyperlinks,” *in Proceedings of the 1998 ACM SIGMOD international conference on management of data*, vol. 3, no. 1, pp. 307-318, June 1998.

[12] Chen M.-S., Park J. S. and Yu P. S., “Efficient Data Mining for Path Traversal Patterns”, *in Proceedings of IEEE transactions on knowledge and data engineering*, Vol. 10, no. 2, pp. 209-221, April 1998.

[13] Clifton C., “Encyclopaedia Britannica: Definition of Data Mining", Available: http://www.britannica.com/EBchecked/topic/1056150/data-mining, May 2012.

[14] Cooley R., Mobasher B., and Srivastava J., “Data preparation for mining world wide web browsing patterns,” *Knowledge and Information Systems*, vol. 1, no. 1, pp. 5-32, 1999.

[15] Cooley R., Srivastava J., and Mobasher B., “Web mining: Information and pattern discovery on the World Wide Web”, *in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’97)*, 1997.

[16] Cohen E., Krishnamurthy B., and Rexford J.,” Improving end-to-end performance of the web using server volumes and proxy filters”, *in Proceedings of the ACM SIGCOMM '98 conference on Applications, technologies, architectures, and protocols for computer communication*, pp. 241-253, 1998.

[17] Cowie J. and Lehnert W.," Information extraction", *in Communications of the ACM*, vol. 39, no. 1, January 1996.

[18] DeRose S., “What do those weird XML types want, anyway?”, Scotland, September 1999.

[19] Eirinaki M. and Vazirgiannis M., “Web Mining for web personalization*”, in Journal of ACM Transactions on Internet Technology* ,vol. 3 , no. 1, pp. 1–27,2003.

[20] Etminani K., Akbarzadeh-T. M.-R. and Yanehsari N. R., "Web Usage Mining: user's navigational patterns extraction from web logs using Ant-based Clustering Method", in *Proceedings of Conference of European Society for Fuzzy Logic and Technology*, vol. 1, no. 1, 2009.

[21] Etzioni O., "The World Wide Web: Quagmire or Gold Mine", *in Communications of the ACM*, vol. 39, no. 11, pp. 65-68, 1996.

[22] Feldman R. and Dagan I., "Knowledge discovery in textual databases", *in the Proceedings of the first international conference on knowledge discovery and data mining*, 1995.

[23] Facca F.M. and Lanzi P. L. "Mining interesting knowledge from weblogs: a survey", *in Journal Data & Knowledge Engineering*, vol. 53, no. 3, pp. 225 - 241 ,June 2005.

[24] Getoor L., "Link Mining: A New Data Mining Challenge", *in SIGKDD Explorations*, vol. 4, no. 2, 2003.

[25] Goldman R., McHugh J. and Widom J., "From semi-structured data to XML: Migrating the Lore data model and query language", *in Proceedings of the 2nd International Workshop on the Web and Databases*, pp, 25-30, 1999.

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

[27] Han H. and Elmasri R., "Learning rules for conceptual structure on the web", *in Journal of Intelligent Information System*, vol. 22, no. 3, pp. 237-256, 2004.

[28] Han J., Kamber M., and Pei J., "Data Mining: Concepts and Techniques", 3rd edition, USA, June 2011.

[29] Hearst M. A.,"Untangling text data mining", *in Proceedings of ACL'99 the 37th annual meeting of the association for computational linguistics*, pp. 3-10, 1999.

[30] Hou J. and Zhang Y., "Effectively finding relevant web pages from linkage information." *in Proceedings of IEEE Transaction Knowledge Data Engineering*, vol. 15, no. 4, pp. 940-951,2003.

[31] Jalali M., Mustapha N., Mamat A. and Sulaiman Md. N. B., "web user navigation pattern mining approach based on graph partitioning algorithm", *in Journal of Theoretical and Applied Information Technology*, vol. 11,no. 4, 2006.

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

[33] Kantardzic M., "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, pp. 339-343, 2003.

[34] Khosravi M. and Tarokh M. J., "Dynamic Mining of Users Interest Navigation Patterns Using Naive Bayesian Method", *in the Proceedings of the IEEE sixth International Conference on Intelligent Computer Communication and Processing* ,pp. 119-122 ,2010.

[35] Kleinberg J. M., Kumar R., Raghavan P., Rajagopalan S., and.Tomkins A. S, "The Web as a graph: Measurements, models and methods", Lecture Notes in Computer Science, vol. 1627, pp. 1-18, 1999.

[36] Kleinberg J.M.," Authoritative sources in a hyperlinked environment", *in Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, vol. 46,no.5,pp. 668-677 ,1998.

[37] Kosala R. and Blockeel H., 2000, "Web Mining Research: A Survey", *in ACM SIGKDD*, vol. 2, no. 1, pp. 1-15, June 2000.

[38] Lim E.P., Madria S.K., Bhowmick S.S. and Ng W.K, "Research issues in Web data mining", *in the Proceedings of First International Conference on Data Warehousing and Knowledge Discovery*, pp. 303-312, 1999.

[39] Liu B. and Chang K., “Editorial: Special Issue on Web Content Mining”, *in SIGKDD Explorations special issue on Web Content Mining*, vol. 6, no. 2, December 2004.

[40] Masseglia F., Poncelet P. and Cicchetti R., “Webtool: An Integrated framework for data mining”, *in the Proceedings of the Ninth International Conference on Database and Expert System Application (DEXA’99)*, pp.892-901, 1999.

[41] Mortazavi-A’sl B., “Discovering and mining user web-page traversal patterns”, Master’s thesis, Simon Fraser University, 2001.

[42] Muslea I., Minton S. and Knoblock C.," Wrapper induction for semi-structured web-based information sources", *in Journal of Autonomous Agents and Multi-Agent Systems* ,vol. 4,no. 1-2,pp. 93-114, June 2001.

[43] Page L., Brin S., Motwani R., and Winograd T.,” The Pagerank citation ranking: Bring order to the web”, Technical report, Stanford University, 1998.

[44] Park J.S., Chen M.S. and Yu P.S.,” Data Mining for Path Traversal Patterns in a Web Environment”, *in Proceedings of 16th international conference on Distributed Computing Systems*, pp. 385-392, 1996.

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

[46] Pazienza M. T., "Information Extraction: A multidisciplinary approach to an emerging information technology", *in Proceedings of SCIE 1997 International Summer School*, vol.1299, London, 1997.

[47] Pei J., Han J., Mortazavi-Asl B., and Zhu H., “Mining access patterns efficiently from web logs,” *in Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, London, pp.396-407,2000.

[48] Rajman M. and Besancon R., "Text mining-knowledge extraction from unstructured textual data", *in the Proceedings of 6th Conference of International Federation of Classification Societies*, 1998.

[49] Rawat S.S. and Rajamani L., "Discovering potential user browsing behaviours using custom-built apriori algorithm", *in International Journal of Computer Science & Information Technology*, vol. 2, no. 4, August 2010.

[50] Rijsbergen C.J.V.," Information Retrieval", London, Butter worths, 1979.

[51] Srivastava J., Cooley R., Deshpande M. and Tan P.-N., "Web usage mining: discovery and applications of usage patterns from web data", *in SIGKDD Explorations*, vol.1, no. 2, pp. 12–23, 2000.

[52] Srivastava G., Sharma K., Kumar V.," Web Mining: Today and Tomorrow", *in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT)*, pp.399-403, April 2011.

[53] Spiliopoulou M., and Faulstich L. C.,"Wum: A web utilization miner", *in the Proceedings of EDBT Workshop on the Web and Data Bases (WebDB'98)*", Springer Verlag, pp. 109-115, 1999.

[54] Sujatha N. and Iyakutty K., "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", *in Proceedings of European Journal of Scientific Research* ,vol.42 ,no.3 pp.464-476,2010.

[55] Tan A. H.," Text mining: the state of the text and the challenges", *in Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases*, 1999.

[56] Tong W. and Pi-lian H., "Web Log Mining by an Improved AprioriAll Algorithm", *in Proceedings of The Second World Enformatika Conference*, 2005.

[57] Types of web data, Available at http://ptucse.loremate.com/dw/node/7.

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

[59] Wang H., Yang C. and Zeng H., "Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan", *in Communications of the IIMA*, vol. 6 ,no. 2,2006.

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

[61] Wilks Y.," Information extraction as a core language technology", *in Proceedings of SCIE '97 International Summer School* , vol.1299, pp. 1-9, London,1997.

[62] Wu Y.-H. and Chen A.L.P., "Prediction of web page accesses by proxy server log", *World Wide Web* ,vol.5 no. 1,pp. 67–88 ,2002.

[63] Yao H. , Hamilton H. J. and Butz C. J," A Foundational Approach to Mining Itemset Utilities from Databases", *in the Proceedings of the 4th SIAM International Conference on Data Mining*, pp. 482-486, 2004.

[64] Yen S.J., Lee Y.S. and Hsieh M.C.,"An efficient incremental algorithm for mining Web traversal patterns", *in the Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05)*, pp. 274-281, 2005.

[65] Zhang Q. and Segall R.S., "Web Mining: A Survey of Current Research, Techniques, and Software", *in International Journal of Information Technology and Decision Making*, vol. 07, no. 4, pp. 683-720, January 2008.

[66] Zhixiang C., Fowler R.H. and Fu A.W.-C.," Linear Time Algorithms for Finding Maximal Forward References", *in the proceedings of Information Technology: Coding and Computing*, pp. 160-164,2003.

[67] Zhou L., Liu Y., Wang J., Shi, Y, "Utility-based Web Path Traversal Pattern Mining", *in the Proceedings of Seventh IEEE International Conference on Data Mining Workshops*, pp. 373-378,2007.

[68] Zhu J., Hong J. and Hughes J. G., "Using Markov Chains for Link Prediction in Adaptive Web Sites", *in Proceedings of the First International Conference on Computing in an Imperfect World*, pp. 60–73, 2002.

.

# List of Publications

[1] Kaur Chintandeep, Aggarwal Rinkle rani, "Web Mining tasks and types: A survey", in International Journal of Research in IT and management ISSN 2231-4334, vol. 2, no. 2, February 2012.

[2] Kaur Chintandeep, Aggarwal Rinkle rani, "Pattern Discovery Techniques: A survey", in International Conference on Advanced Computing Technologies, June 2012.

[3] Kaur Chinatandeep, Aggarwal Rinkle rani, "Web Mining tasks and types: A survey", in International Conference on Competitiveness & Innovativeness in Engineering, Management and Information Technology (ICCIEMI-2012), pp.40, February 2012.

[4] Kaur Chintandeep, Aggarwal Rinkle rani, "Reference Scan Algorithm for Path Traversal Patterns", in International Journal of Computer Applications, vol. 48, no.2, June 2012.