ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 4, June 2012

An Approach to Improve the Web Performance By Prefetching the Frequently Access Pages.

Ms. Seema
Dept of Computer Science and Engg
Manav Rachna College of Engineering
Faridabad, India

Ms. Priyanka Makkar
Dept of Computer Science and Engg
Manav Rachna College of Engineering
Faridabad, India

Abstract-World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. The growth of web is tremendous as approximately one million pages are added daily. Users' accesses are recorded in web logs. Because of the tremendous usage of web, the web log files are growing at a faster rate and the size is becoming huge. Web data mining is the application of data mining techniques in web data. Web Usage Mining applies mining techniques in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, customer profiling, prefetching, creating attractive web sites etc., The rapid growth in the amount of information and the number of users has lead to difficulty in providing effective search services for the web users and increased web latency; resulting in decreased web performance. Although web performance can be improved by caching, the benefit of using it is rather limited owing to filling the cache with documents without any prior knowledge .Web pre-fetching becomes an attractive solution wherein forthcoming page accesses of a client are predicted, based on log information. This paper proposes an approach for increasing web performance by analyzing and predicting user behavior both by collaborating information from user access log and website structure repository.

***Index Terms*—Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Data Cleaning, User Identification, Session Identification, Path Completion , Prefetching and Markov Model.**

I. INTRODUCTION

In this paper gives the description about Web Mining and their tasks. To extract the useful information over the Web using Web Mining technique. This chapter also gives the description about the web data that can be used in the Web Mining, architecture of Web Mining and their categories. The categories of the Web Mining are Web Content Mining, Web Structure Mining and Web Usage Mining. Web content mining is the process of extracting useful information from the content of Web documents. Web structure mining uses the hyperlink structure of the Web to yield useful information, including definitive pages specification, hyperlinked community's identification, Web pages categorization and Web site completeness evaluation. Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. This paper is an overview of these techniques. and in this paper we used the pre-fetching techniques Markov Model used in the identification of the next page to be accessed by the website user based on the sequence of previously accessed pages.

II. RELATED WORK

Padmanabhan et al [1] investigated ways of reducing retrieval latency. A scheme called prefetching was introduced in which clients in collaboration with servers prefetch web page that the user is likely to access soon, while he/she is viewing the currently displayed page. L Fan et al [2] investigate an approach to reduce web latency by prefetching between caching, proxies, and browsers. Research on predictive Web prefetching has involved the important issue of log file processing and the determination of user transactions (sessions). Pirolli and Pitkov [15] predict the next web page by discovering the longest repeating subsequence in the web sessions. Yang et al [7] studied different association rule based methods for web request prediction. Using association rules for web access prediction involves dealing with too many rules and it is not easy to find a suitable subset of rules to make accurate and reliable predictions. Padbanabham and Mogul [8] use N-hop Markov models predicted the next web page users will most likely access by matching the user's current access sequence with the user's historical web access sequences for improving prefetching strategies for web caches. Sarukkai [5] used first-order Markov models to model the sequence of pages requested by a user for predicting the next page accessed. Liu et al [19] integrated association rules and clustering for reducing the overhead associated with large database. Cadez et al [10] integrated clustering and first order.

III. WEB MINING

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services. Although Web mining puts down the roots deeply in data mining, it is not equivalent to data mining. The unstructured feature of Web data triggers more complexity of Web Mining. Web Mining is a technique to discover and analyze the useful information from the Web data. The Web involves three types of data: data on the Web (content), Web log data (usage) and Web structure data.

The classified the data type as content data, structure data, usage data, and user profile data [13].Web Mining can be divided into four subtasks.

- **Information Retrieval (IR) and Resource Discovery (RD):** Find all relevant information on the web. The goal of IR is to automatically find all relevant information, while at the same time filter out the no relevant ones. Search engines are a major tool people use to find web information. Search engines use keywords as the index to perform query. Users have more control in searching web content. Automated programs such as crawlers and robots are used to search the web. Such programs traverse the web to recursively retrieves all relevant information.
- **Information extraction (IE):** Automatically extract specific fragments of a document from web resources retrieved from the IR step. Building a uniform IE system is difficult because the web content is dynamic and diverse. Most IE systems use the "wrapper" technique to extract specific information for a particular site. Machine learning techniques are also used to learn the extraction rules.
- **Generalization:** Discover information pattern at retrieved web sites. The purpose of this task is to study user's behavior and interest. Data mining techniques such as clustering and association rules are utilized here. Several problems exist during this task. Because web data are heterogeneous, imprecise and vague, it is difficult to apply conventional clustering and association rule techniques directly on the raw web data.
- **Analysis/Validation:** analyze, interpret and validate the potential information from the information patterns. The objective of this task is to discover knowledge from the information provided by former tasks. Based on web data, we can build models to simulate and validate web information.[14]

Web Data: There are many kinds of data that can be used in the web mining.[3]

- **Content Data:** The real data in the Web pages, i.e. the data the Web page was designed to convey to the users .This usually consists of, but is not limited to, text and graphics. The content data in a site is the collection of objects and relationships that is conveyed to the user. For the most part, this data is comprised of combinations of textual materials and images. The data a source used to deliver or generates this data include static HTML/XML pages, multimedia files, dynamically generated page segments from script and collections of records from the operational databases. The site content data also includes semantic or structural meta-data embedded within the site or individual pages such as descriptive keywords, document attributes, semantic tags or HTTP variables.
- **Structure Data:** Data which describe the organization of content intra-page structure information includes the arrangements of various HTML or XML tags within a given page. This can be represented as a tree structure where the html tag becomes the root of the tree and the information of the inter page structure is hyperlinks connecting one page to another. The structure data represents the designer`s view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. The structure data also includes the intra-page structure of the content within a page.

  For Ex: both HTML and XML documents can be represented as tree structures over the space of tags in the page.

- **Usage Data:** Data that describes the pattern of usage of web pages such as IP address, page references and date and time accesses and various other information depending on the log format. The log data collected automatically by the Web and application servers represents the fine-grained navigational behavior of visitors. It is the primary source of data in Web usage mining. Each log entry (depending on the log format) may contain fields identifying the time and date of the request, the IP address of the client, the resources requested, possible parameters used in invoking a Web application, status of the request, HTTP method used, the user agent (browser and operating system type and version), the referring Web resource and if available, client-side cookies which uniquely identify a repeat visitor.

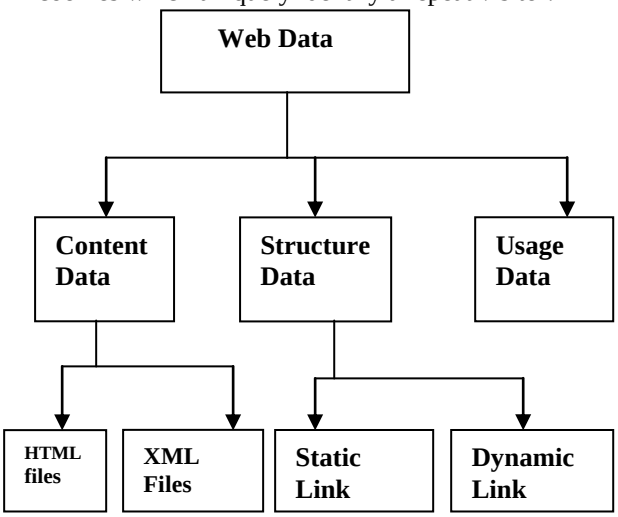


Figure No-1 Types of Web Data

Web mining is categorized into three areas. Each of interest based on part of Web to mine:

Web content mining describes the automatic search of information resource available online and involves mining web data contents. 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 characteristics of Web data force the Web content mining towards a more complicated approach.

Web Structure Mining: Web Structure of web graph consists of web pages as a nodes and hyperlinks as an edge connecting related pages. **Web structure mining is the process of discovering structure information from the web. Web structure mining tries to discover the model underlying the link structures of the web.** The model is based on the topology of the hyperlink with or without the link description. This model can be used to categorize the web pages and is useful to generate information such as similarity and relationships between web sites. And the link structure of the web contains important implied information,

and can help in filtering or ranking web pages. **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.**

Web Usage Mining is the application of the data mining techniques to discover interesting usage patterns from web usage data. This technique mines the data from the logs and provide user with relevant pages. Web usage mining process can be divided into three independent tasks: Preprocessing, pattern discovery and pattern analysis. Preprocessing is first phase of Web mining process. Usage, content and structure information contained in the various available data sources are converted for next step that is pattern discovery. Pattern discovery is based on methods and algorithms developed from several areas such as data mining, machine learning and pattern recognition

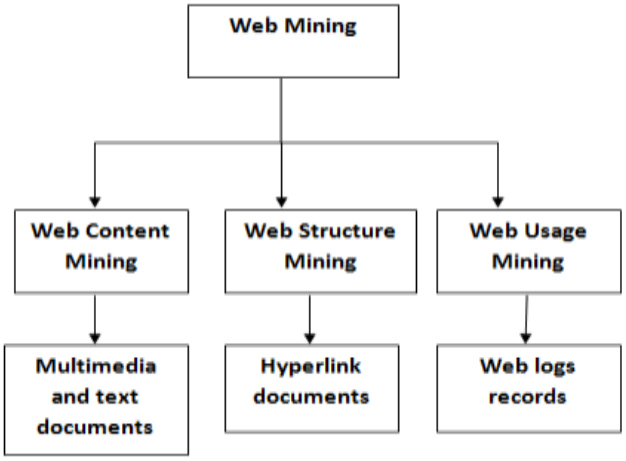


Figure No-2 Categories of Web Mining

Data Sources: The data sources used in Web Usage Mining may include web data repositories like:[18]

Web server logs: These are logs which maintain a history of page requests. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added.

Proxy server logs: A Web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of Web pages as well as the network traffic load at the server and client side. Proxy server logs contain the HTTP requests from multiple clients to multiple web servers.

Browser server logs: Various side browsers like Mozilla, internet explorer etc. can be modified or various java script and java applets can be used to collect client side data. This implementation of client side data collection requires user cooperation either in enabling the functionality of the java script and java applets or to voluntarily use the modified browser. Client-side collection scores over server-side collection because it reduces the session identification problems.

Format of server logs: Logs will contain access pattern of user and the data is stored in the Web access log files. A typical example of web log file is shown in Table 3.1. Each access to a Web page is recorded in the access log of the Web server that hosts it. The entries of a Web log file consist of fields that follow a predefined format such as Common Log Format (CLF), Extended Common Log Format (ECLF). A CLF file is created by the Web server to keep track of the requests that occur on a web site. ECLF is supported by Apache and Netscape (W3C).

The common log format appears exactly as follows:

Host/iprfcname[DD/MM/YYYY:HH:MM:SS-0000]"METHOD/PATH HTTP/1.0" code bytes

Host/ip: If reverse DNS works and DNS lookup is enabled, the host name of the client is dropped in otherwise the IP number displays.

RFC name: You can retrieve a name from the remote server for the user. If no value is present, a"-" is

Date stamp: The format is day, month, year, hour in 24-hour clock, minute, second.

Retrieval: Method is GET, PUT, POST and HEAD; path and file retrieved.

Bytes: Number of bytes in the file retrieved.

One record in web access log is written as:[6]

219.144.222.253—[16/Sep/2008:16:32:56+0800]"GET/images/2_r3_g2.jpgHTTP/1.1"200418http://230.117.16.223:8090/index.html "Mozilla/4.0(compatible; MSCE6.0; windows NT5.1)".

Format of Web Access Log as shown below:

Field	Meaning
219.144.222.253	User IP address
16/Aug/2004:15:36:11	Date and time of request
GET	The method of request
images1_r3_c2.jpg	The URL of current request(URI)
HTTP/1.1	The version of transport protocol(version)
200	The HTTP status code return to client (status)
418	The content length of page transferred (Byte)
http://202.117.16.119:8089/index.html"	The URL requested just before (Refer URI)
Mozilla/4.0(compatible;MSCE6.0;Windows NT5.1"	Browser & operating System

Figure No-3 Format of Web Server Log

Phases of Web usage mining: In general, the process of web usage mining can be divided into three parts namely preprocessing, pattern discovery, and pattern analysis [17]

- **Preprocessing:** Performs a series of preprocessing of web log file covering data cleaning, user identification, session identification, path completion and transaction identification. Data preprocessing is predominantly significant phase in Web usage mining due to the characteristics of Web data and its association to other related data collected from multiple sources. This phase is often the most time-consuming and computationally intensive step in Web usage mining.

- **Pattern Discovery:** Application of various data mining techniques to preprocessed data like statistical analysis, association, clustering, and pattern matching and so on.
- **Pattern Analysis:** Once patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

Data Preprocessing: The main sources of weblog files in web usage mining are 1.Web Servers 2.Web proxy Servers 3.Client browsers. Here Server logs supply the best usage data than the remaining log files. If at all there are multiple weblog files, they should be converted into a joint weblog file. This joint file will undergo further operations. In this paper a server log file of 7.8 MB is considered for the purpose of analysis.

- Data Collection(Data Fusion)
- Data cleaning
- User identification
- Session identification
- Path completion
- Formatting

Data Collection: Data Collection refers to the techniques that combine data from multiple sources to retrieve additional information with respect to identification of users and sessions. The primary data sources used in web usage mining
are the server log includes web server access logs and application server logs. An additional data may be available from client side and Proxy server.

Server level Collection: The server stores data regarding request performed by the client, it gives the browsing behavior of the client. All the records are recorded into the server log file. Only information requested by GET is recorded but POST method request is not available on server log.

Client Level Collection: It is the client itself which sends to a repository. Client-side data collection can be implemented by using a remote agent or by modifying the source code of an existing browser to enhance its data collection capabilities. Client level collection overcomes the problem of cache data which is not available at server level.

Proxy Level Collection: Proxy level collection sits in between server level and client level. Due to proxy the page load time gets reduce, so user experience high performance.

Data Cleaning: The purpose of data cleaning is to remove irrelevant, noisy data from the server logs and may not be useful for analysis purposes. For example: log entries with file name suffixes such as .gif, .jpeg can be removed.

Algorithm for data cleaning [16]**:**An algorithm for data cleaning is depicted as shown below.

Input: log files (LF)
Output: Summarized log file (SLF)

```
Begin
1. Read records in LF
2. for each record in LF
3. Read fields (Status code, method)
4. If status code=" 200" and method=" GET" Then
5. Get IP_address and URL
6. If URL_suffix is= {.gif, .js, .jpeg, .jpg, .css} Then
7. Remove URL_suffix
8. Save IP_address and URL
End if
Else
9. Next record
End if
End
```

User Identification: This step requires the identification of unique users. User identification deals with associating page references with the different users.

- **User Agent:** User agent plays an important role in user identification. It refers to the browser used by the client. A change in the browser or the operating system under the same IP address represents a different user.
- **Referring URL:** IP address is generally not sufficient for user identification. This is due to the fact that proxy servers may assign the same IP address to multiple users or the same user may be assigned multiple IP addresses by the proxy server. Another determines the navigation paths for each user.

Session Identification: Session identification splits all the pages accessed by a user into different sessions. Users may have visited the pages for long periods of time. It is necessary
to divide the log entries of a user into multiple sessions through a time. Where if the time between page requests
exceeds a certain limit, it is assumed that the user has started a new session. Two methods depend on time and one on navigation in web topology.

Time Oriented Heuristics[4]: The simplest methods are time oriented in which one method based on total session time and the other based on single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes to 24 hours while 30minutes is the default timeout. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes then the second entry is assumed as a new session. Time based methods are not reliable because users may involve in some other activities after opening the web page and factors such as busy communication line, loading time of components in web page, content size of web pages are not considered.

Navigation-Oriented Heuristics: uses web topology in graph format. It considers webpage connectivity, however it is not necessary to have hyperlink between two consecutive page requests. If a web page is not connected with previously visited page in a session, then it is considered as a different session. If no referrer is found then it is a first page of a new session.

Path completion: Path completion refers to the important page accesses that are missing in the access and log due to browser and proxy server caching. Similar to user identification, the heuristic assumes that if a page is requested that is not directly linked to the previous page

accessed by the same user, the referrer log can be referred to see from which page the request came. If the page is in the user`s recent click history, it is assumed that the user browsed back with the "back" button using cached sessions of the pages. Hence each session reflects the full path, including the pages that have been backtracked.

Formatting: The resultant file may not be in a suitable format to be used for any mining tasks. Data should be formatted according to the type of mining tasks undertaken. Information which is viewed irrelevant or unnecessary for the analysis may not be included in the resultant session file.

Pattern Discovery [9]: Web Usage mining can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. The following are the pattern discovery methods.

Statistical Analysis: Statistical Analysis techniques are the most common method to extract knowledge about visitors to a web site. **By analyzing the session file, one can perform different kinds of descriptive statistical analyses on variables such as pages views, viewing time and length of a navigational path.** Statistical analysis of pre-processed session data constitutes the most common form of analysis. In this case, data is aggregated by predetermined units such as days, sessions, visitors or domains. Standard statistical techniques can be used on this data to gain knowledge about visitor behavior. This is the approach taken by most commercial tools available for web log analysis. Reports based on this type of analysis may include information about most frequently accessed pages.

Clustering: Clustering is a technique to **group together a set of items having similar characteristics.** In web usage mining there are two kinds of interesting clusters to be discovered: **usage clusters and page clusters.** Clustering of pages will discover groups of pages having related content. Clustering analysis allows one to group together clients or data items that have similar characteristics. Clustering of client information or data items on Web transaction logs, can facilitate the development and execution of future marketing strategies, both online and off-line, such as automated return mail to clients falling within a certain cluster, or dynamically changing a particular site for a client, on a return visit, based on past classification of that client. For web usage mining, clustering techniques are mainly used to discover two kinds of useful clusters, namely user clusters and page clusters.

Association Rules: Association rule generation can be used to relate pages that are most referenced together in a single server session. In web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. **Association rule mining discovery techniques [2], [8] are generally applied to databases of transactions where each transaction consists of a set of items.** In such a framework the problem is to discover all associations and correlations among data items where the presence of one set of items in a transaction implies (with a certain degree of confidence) the presence of other items. In context of web usage mining, this problem amounts to discovering the correlations among references to various files available on the server by a given client. Each transaction is comprised of a set of URLs accessed by a client in one visit to the server.

Classification: Classification is the task of mapping a data item into one of several predefined classes. In web domain developed the user profile belonging to a particular class and category. Classification can be done by using supervised inductive learning algorithm such as decision tree classifier, support vector machines. In Web mining, classification techniques allow one to develop a profile for clients who access particular server files based on demographic information available on those clients, or based on their access patterns.

Sequential Patterns: The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes [9]. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes.

Dependency Modeling: Dependency modeling is another useful pattern discovery task in Web Mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the Web domain. Such techniques include Hidden Markov Models and Bayesian Belief Networks. **The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the Web domain.** The modeling technique provides a theoretical framework for analyzing the behavior of users, and is potentially useful for predicting future Web resources consumption.

Markov Model:[20] Web access prediction using Markov model and association rules. **Markov models are becoming very commonly used in the identification of the next page to be accessed by the Web site user based on the sequence of previously accessed pages.** Let P= {p1, p2,……., pm} be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited l pages, then prob (pi|W) is the probability that the user visits pages pi next. Markov models have been Markov models used for studying and understanding stochastic processes, and were shown to be well-suited for modeling and predicting a user's browsing behavior on a web-site. **Markov models have been widely used to model user navigation on the web and predicting the action a user will take next given the sequence of actions he or she has already performed.**

The fundamental assumption of predictions based on Markov models is that the next state is dependent on the previous k states. The longer the k is the more accurate the predictions are. However, longer k causes the following two problems: the coverage of model is limited and leaves many states uncovered and the complexity of the model becomes unmanageable. Therefore, the following are three modified Markov models for Predicting Web page access.

- **All kth Markov model:** This model is to tackle the problem of low coverage of a high order Markov model. For each test instance, the highest order Markov model that covers the instance is used to predict the instance.

For example, if we build an all 4-Markov model including 1, 2, 3, and 4 for test instance we try to use 4- markov model

to make prediction. If the 4-markov model does not contain the corresponding states, we then use the 3 markov model and so forth.

- **Frequency pruned Markov model:** though all kth order Markov models result in low coverage they exacerbate the problem of complexity since the states of all Markov models are added up. Note that many states have low statistically predictive reliability since their occurrence frequencies are very low. The removal of these low frequency states affects the accuracy of a Markov model. However, the number states of the pruned Markov model will be significantly reduced.
- **Accuracy pruned Markov model:** Frequency pruned Markov model does not capture factors that affect the accuracy of states. A high frequent state may not present accurate prediction. When we use a means to estimate the predictive accuracy of states, states with low predictive accuracy can be eliminated. One way to estimate the predictive accuracy using conditional probability is called confidence pruning. Another way to estimate the predictive accuracy is to count (estimated) errors involved, called error pruning.

Pattern Analysis: Pattern Analysis is the final stage of WUM (Web Usage Mining), which involves the validation and interpretation of the mined pattern.

- **Validation:** To eliminate the irrelevant rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process.
- **Interpretation:** The output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations.

This is the final step in the Web Usage Mining process**. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL (Structured Query Language) processing and OLAP (Online Analytical Processing) can be used.** The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase.

The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations.

OLAP (Online Analytical Processing):- OLAP provides a more integrated framework for analysis with a higher degree of flexibility. The data sources for OLAP analysis are usually a multidimensional data warehouse which integrates usage, content and e-commerce data at different levels of aggregation for each dimension. OLAP tools allow changes in aggregation for each dimension during the analysis. Analysis dimensions in such a structure can be based on various fields available in the log files, and may include time duration, domain, requested resources, user agent, and referrers.

IV. PROPOSED WORK

Proposed Architecture: This section describes the architecture of the proposed system shown in Figure 4. Following subsection describes various components of the proposed system.

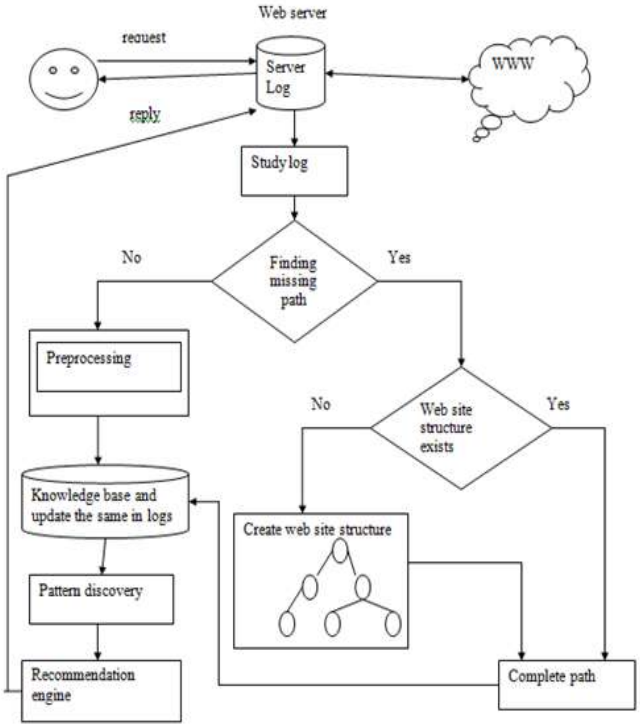


Figure No-4 Proposed System Architecture

Preprocessing: It is important to understand that the quality data is a key issue when we are going to mining from it. Nearly 80% of mining efforts often spend to improve the quality of data. The data which is obtained from the logs may be incomplete, noisy and inconsistent. Data preprocessing involves the data cleaning, user identification and session identification. In data cleaning removing all the data tracked in web logs that are useless for mining purpose. For e.g.:(.jpg, .gif and .css).

Knowledge Base: The knowledge base is a repository of extracted rules which have been derived by applying **data mining techniques.** The attributes it contains are Number of users, the web page they access and the time of accessing the web pages.

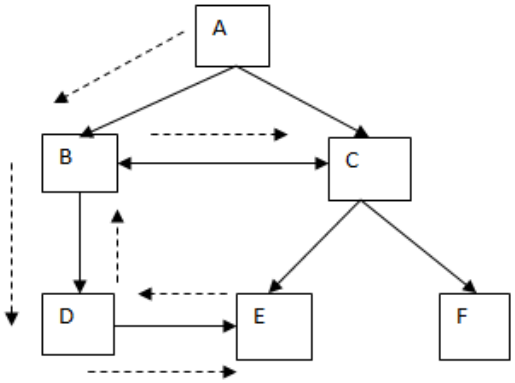
Various algorithms and techniques like[12] **Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.**

Recommendation Engine: Whenever user request for URL, Prediction Pre-fetching Engine (PPE) sends the URL to the recommendation engine, which in turn does prediction based on rules from knowledge base. Finally based on this, it pre-fetch the web page in server cache before he/she explicitly request for, thus decreasing access time of retrieving that page and improving web server performance.

Algorithm for recommendation engine
1. Begin
2. for each new query or for each coming user session
3. Make predictions based on KB (Knowledge Base)
4. Otherwise find its closest cluster
5. Use corresponding Markov model to make prediction
6. If the predictions are made by states that do not belong to a majority class
7. Use association rules to make a revised prediction
8. End If
9. Pre fetches pages
10. End For
11. End For
12. End

Path Completion: Path completion is depends on mostly URL and REF URL fields in server log file. It is also graph model. Graph model represents some relation defined on Web pages (or web), and each tree of the graph represents a web site. . Each node in the tree represents a web page (html document), and edges between trees represent the links between web sites, while the edges between nodes inside a same tree represent links between documents at a web site. Client or Proxy side caching can often result in missing access references those pages or object that have been cached. For instance, if a user goes back to a page A during the same session, the second access to A will likely result in viewing the previously downloaded version of A that was cached on the client-side, and therefore no request is made to the server. This results in the second reference to A not being recorded on the server logs.

For Example: **A simple example of missing references is given in figure 5.3. On the left, a graph representing the linkage structure of the site is given. The dotted arrows represent the navigational path followed by a hypothetical user. After reaching page E, the user has backtracked (e.g., using browser's "back" button) to page D and then B from which she has navigated to page C. The back references to D and B not appear in the log file because these pages where cached on the client-side (thus no explicit server request was made for thesepages). The log files shows that after a request for E, the next request by the user is for page C with a referrer B.**



User's actual navigation path:

A->B->D->E->D->B->C

What the server log shows:

URL	Referrer
A	----
B	A
D	B
E	D
C	B

Figure No-5 Example of Path Completion

V. EXPERIMENTAL ANALYSIS

In this we draw the web site structure and after that we create the matrix to represent the web site structure. According to matrix we complete the path in the user sessions. And after that using the Markov Model we predicting the next page to be accessed by the web site user based on the sequence of previously accessed pages. The fundamental assumption of prediction based on Markov models is that the next state is dependent on the previous k states.

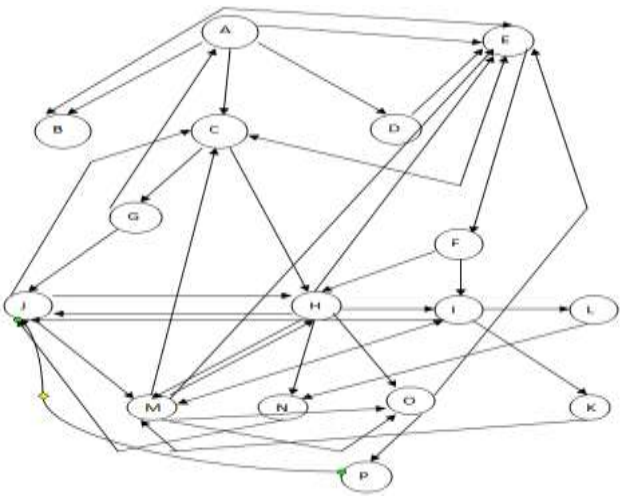


Figure No-6 Structure of Web Site

Draw Matrix According to the Web Site Structure

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0
D	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
E	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
G	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
H	0	0	0	0	1	0	0	0	1	1	0	0	1	1	1	0
I	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
J	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
M	0	0	1	0	1	0	0	1	1	0	0	0	0	0	1	0
N	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
O	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
P	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Table No- 1 Matrix for Web site structure

User Sessions before path completion: We examines the 7 user sessions in **Figure no. 7**. These user sessions are shows before path completion. We complete the path using matrix that shows in **Table No-1**.

Session 1: A->C->E->F->H->O->P->M
Session 2: A->D->B->E->F->H->M->I->J
Session 3: A->B->D->C->E->F->I->L->N->J->H
Session4:A->C->B->E->F->I>K->M->O->E->F->I->M->C->G
Session 5: A->C->G->J->M->E->F->H->N->J->E
Session 6: A->B->D->C->E->F->H->I->J->H->E->F->I->J->C
Session 7: A->C->H->J->M->H->M->C->A

Figure No- 7 User session before path completion

User sessions after path completion: We complete the path in the user sessions using matrix shown in **Table no-1** and user sessions after path completion shown in **Figure no-8**.

Session 1: A->C->E->F->H->O->P->J->M
Session 2: A->D->E->B->E->F->H->M->I->J
Session 3: A->B->D->E->C->E->F->I->L->N->J->H
Session 4:
A->C->E->B->E->F->I>K->M->O->E->F->I->M->C->G
Session 5: A->C->G->J->M->E->F->H->N->J->H->E
Session 6:
A->B->D->E->C->E->F->H->I->J->H->E->

Figure No- 8 User session after path completion

Applying the Ist order Markov Model to above user sessions and calculating the frequencies of accessed pages, **TableNo-2 lists the page views with their frequencies.**

Page	Frequency
A	8
B	4
C	9
D	3
E	13
F	8
G	3
H	9
I	6
J	8
K	1
L	1
M	7
N	2
O	2
P	1

Table No-2 Page Views with their frequencies

A 100% support results in a very large number of rules and is rather cumbersome. Therefore, assuming that the **minimum support is 3: K, L, N, O and P are removed** from the user sessions. **Figure No-9** lists the user sessions that passes the frequency and support tests.

Session 1: A->C->E->F->H->J->M
Session 2: A->D->E->B->E->F->H->M->I->J
Session 3: A->B->D->E->C->E->F->I->J->H
Session 4:
A->C->E->B->E->F->I->M->E->F->I->M->C->G
Session 5: A->C->G->J->M->E->F->H->J->H->E
Session 6: A->B->D->E->C->E->F->H->I->J->H->E->F->I->J->C

Figure No- 9 User sessions after frequency and support pruning

	A	B	C	D	E	F	G	H	I	J	M
A	0	2	4	1	0	0	0	0	0	0	0
B	0	0	0	2	2	0	0	0	0	0	0
C	0	0	0	0	4	0	3	1	0	0	0
D	0	0	0	0	3	0	0	0	0	0	0
E	0	2	2	0	0	8	0	0	0	0	0
F	0	0	0	0	0	0	0	4	3	0	0
G	1	0	0	0	0	0	0	0	0	1	0
H	0	0	0	0	0	0	0	0	0	1	1
I	0	0	0	0	0	0	0	0	0	3	1
J	0	0	1	0	0	0	0	3	0	0	2
M	0	0	1	0	1	0	0	1	1	0	0

Table No-3 Using 2nd Markov Model calculate the frequencies of each state

Applying the **2nd order Markov Model** to the above user sessions we notice the **most frequent state is (E,F)** and it appeared 8 times as follows:

Pl+1= argmax {P (H | E, F)} = H OR I

Obviously, this information alone does not provide us with correct prediction of the next page to be accessed by the user as we have high frequencies for pages, H and I. To break the tie and find out which page would lead to the most accurate prediction, we have to look at previous pages in history. This is where we use subsequence association rules as it shows in **Table No- 4 below.**

Previous node of most frequent node	Most frequent node	Post node of most frequent node
A,C	EF	H
A,D,E,B	EF	H
A,B,D,E	EF	I
A,C,E,B	EF	I
M	EF	I
A,C,G,J,M	EF	H
A,B,D,E,C	EF	H
I,J,H	EF	I

Table No-4 User sessions history

Table No-5 and 6 summaries the results of applying subsequences association rules to the training data. **Table No-5** shows **that G->H has the highest confidence of 100%.** While **Table No-6** shows that **H->I has the highest confidence of 100%.**

A->H	AH/A	4/6	66%
B->H	BH/B	2/4	50%
C->H	CH/C	2/4	50%
D->H	DH/D	1/3	33%
E->H	EH/E	2/4	50%
G->H	GH/G	1/1	100%
I->H	IH/I	0/1	0%
J->H	JH/J	1/2	50%
M->H	MH/M	1/2	50%

Table No- 5 Confidence of accessing page H using subsequence association rules

A->I	AI/A	4/6	66%
B->I	BI/B	2/4	50%
C->I	CI/C	1/4	25%
D->I	DI/D	0/3	0%
E->I	EI/E	2/4	50%
G->I	GI/G	0/1	0%
H->I	HI/I	1/1	100%
J->I	JI/J	½	50%
M->I	MI/M	1/2	50%

Table No- 6 Confidence of accessing page I using subsequence association rules

Using Markov Models, we can determine that there is a 50/50 chance that the next page to be accessed by the user after accessing the pages E and F could be either H or I. **Whereas subsequence association rules take this result a step further by determining that if the user accesses page G before pages E and F then there is a 100% confidence that the user will access page H next.**

Whereas, if the user visits page H before visiting pages E and F, then there is a 100% confidence that the user will access page I next.

VI. REFERENCES

[1] Venkata N. Padmanbhan. "Improving World Wide Web Latency", Technical Report, Computer Science Division, University of California, Berkeley, CA, May, 1995.

[2] Fan L., Cao P., and Jacobson Q., "Web prefetching between Low-Bandwidth Clients and proxies: potential and performance." In Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems. May 1999.

[3] Bam shad Mobahser,"Web Data Mining: Exploring Hyperlinks, Contents and Usage Data", "Department of Computer Science".

[4] V. Chitraa,"A Survey on Preprocessing Method for Web Usage Data", "CMS College of Science and Commerce", Coimbatore.(IJCSI) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010

[5] Sarukkai, R.R., "Link prediction and path analysis using Markov chain", proc. of the 9th International World Wide Web Conference on Computer networks, 2000.

[6] Research on Path Completion Technique in Web Usage Mining," A School of Electronics and Information Engineering, "Xi'an Jiao tong University", China International Symposium on Computer Science and Computational Technology . 2008

[7] Yang, Q., Li, T., Wang, K., "Building Association Rules Based Sequential Classifiers for Web Document Prediction", journal of Data Mining and Knowledge Discovery, Netherland: Kluwer Academic Publisher, 2004.

[8] V. Padmanabhan and J. Mogul, "Using Predictive prefetching to improve World Wide Web latency", *ACM SIGCOMM Computer Comm. Rev., Vol. 26,no.3,* July 1996.

[9] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In International Conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, 1997.

[10] Cadez I., Heckerman D., Meek. Smyth P., and Whire S., "Visualization of Navigation Patterns on a website using Model Based Clustering", March, 2002.

[11]Papoulis, A., Probability, Random Variables, and Stochastic Processes, NY: McGraw Hill, 1991.

[12] Mrs. Bharati M. Ramageri," Data Mining Techniques and Applications"**,** "Department of Computer Application, Yamunanagar, Nigdi Pune, Maharashtra", Vol. 1 No. 4 301-305.

[13]Srivastava,"Research paper on web mining"," Department of computer science", University of Minnesota.

[14] Jin,M.S.,"Design and implementation of web mining Research Support System", "Department of Computer Science", University of Notre Dame.

[15] Pitkov, J. and Piroli, P. "Mining Longest repeating Subsequences to predict World Wide Web surfing, Proc. USENIX Symp. On Internet Technologies and Systems, 1999.

[16] Sanjay Bapu Thakare," A Effective and Complete Preprocessing for Web Usage Mining", "Department of Information and Technology",(IJCSE) International Journal on Computer Science and Engineering, Vol-2, No-3,2010,848-851.

[17] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pan-Ning Tan: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, Vol. 1, Issue 2, Jan. 2000, pp. 12-23

[18] Mohd Helmy Abd Wahab,"Discovering web servers logs patterns using generalized association rules algorithm", "Department of Computer Science and Engg", Malaysia.

[19] Liu, F. Lu Z. "Mining Association rules using clustering", Intelligent Data Analysis, 2001.

[20] Sir porn Chimphlee ,"Using Markov Model and Association Rules for Web Access Prediction**",**" Faculty of Science and Technology"," Suan Dusit Raja hat University", Bangkok, Thailand.

AUTHOR PROFILE

Seema received B.E. degree in Computer Science & Engineering with Hons. from Maharshi Dayanand University in 2010 and is pursuing M.Tech.in Computer Engineering from Manav Rachna College of Engineering, Faridabad.

Priyanka Makkar received B.E. degree in Computer Science & Engineering with Hons. from Maharshi Dayanand University in 2007 and received M.Tech. in Information Technology YMCA University of Science & Technology. Presently, she is working as a Assistant Professor in Computer Engineering department in Manav Rachna College of Engineering, Faridabad. Her areas of interests are Web Mining and Search Engines.