International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 7, July 2014 **ISSN 2319 - 4847**

Analysis of Web Usage Mining

Swarna Latha Ampilli[1], Suresh Tippana[2]
1Senior Assistant Professor
Dadi Institute of Engineering & Technology, NH-5, Anakapalle
2 Assistant Professor
Dadi Institute of Engineering & Technology, NH-5, Anakapalle

ABSTRACT

With the growing popularity of the World Wide Web (Web), large volumes of data are gathered automatically by Web servers and collected in access log files. Analysis of server access data can provide significant and useful information. The web is an important source of information retrieval nowadays, and the users accessing the web are from different backgrounds. The usage information about users are recorded in web logs. Analyzing web log files to extract useful patterns is called web usage mining. Web usage mining approaches include clustering, association rule mining, sequential pattern mining etc.To facilitates web page access by users, and web recommendation model is needed.Web Usage Mining (WUM) integrates the techniques of two popular research fields - Data Mining and the Internet. By analyzing the potential rules hidden in web logs, WUM helps personalize the delivery of web content and improve web design, customer satisfaction and user navigation through pre-fetching and caching.
Keywords— web usage mining, web mining, pre-processing, pattern discovery, pattern analysis**.**

1. INTRODUCTION

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Based on the different emphasis and different ways to obtain information, web mining can divided into three major parts: Web Contents Mining, Web structure Mining and Web Usage Mining. Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Usage data. Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. We provide a brief overview of the three categories.[1]

1)Web Content Mining:-Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language .Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to Web content mining has been limited.

2.) Web Structure Mining:- The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an **Intra-Document Hyperlink:-** and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis, of which Desikan et al. provide an up-to-date survey.

Document Structure: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

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. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered: Web Server Data: The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

Application Server Data:- Commercial application servers such as Web logic Story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data:- New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined events. It must be noted however that many end applications require a combination of one or more of the techniques applied in the above the categories.

2. APPLICATIONS

An outcome of the excitement about the Web in the past few years has been that Web applications have been developed at a much faster rate in the industry than research in Web related technologies. Many of these were based on the use of Web mining concepts – even though the organizations that developed these applications, and invented the corresponding technologies, did not consider it as such. We describe some of the most successful applications in this section. Clearly, realizing that these applications use web mining is largely retrospective exercise.

Personalized Customer Experience in B2C E-commerce – Amazon.com:- Early on in the life of Amazon.com, its visionary CEO Jeff Bezos observed, 'In a traditional (brick-and-mortar) store, the main effort is in getting a customer to the store. Once a customer is in the store they are likely to make a purchase – since the cost of going to another store is high – and thus the marketing budget (focused on getting the customer to the store) is in general much higher than the in-store customer experience budget (which keeps the customer in the store). In the case of an on-line store, getting in or out requires exactly one click, and thus the main focus must be on customer experience in the store.' This fundamental observation has been the driving force behind Amazon's comprehensive approach to personalized customer experience, based on the mantra 'a personalized store for every customer' [M2001]. A host of Web mining techniques, e.g. associations between pages visited, click-path analysis, etc., are used to improve the customer's experience during a 'store visit'. Knowledge gained from Web mining is the key intelligence behind Amazon's features such as 'instant recommendations', 'purchase circles', 'wish-lists', etc..

Web Search – Google:- Google is one of the most popular and widely used search engines. It provides users access to information from almost 2.5 billion web pages that it has indexed on its server. The simplicity and the quickness of the search facility, makes it the most successful search engine. Earlier search engines concentrated on the Web content to return the relevant pages to a query. Google was the first to introduce the importance of the link structure in mining the information from the web. Page Rank, that measures an importance of a page, is the underlying technology in all Google search products. The Page Rank technology, that makes use of the structural information of the Web graph, is the key to returning quality results relevant to a query. Google has successfully used the data available from the Web content (the actual text and the hyper-text) and the Web graph to enhance its search capabilities and provide best results to the users. Google has expanded its search technology to provide site-specific search to enable users to search for information within a specific website. The 'Google Toolbar' is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. The full version of the toolbar, if installed, also sends the click-stream information of the user to Google. The usage statistics thus obtained would be used by Google to enhance the quality of its results. Google also provides advanced search capabilities to search images and look for pages that have been updated within a specific date range. Built on top of Netscape's Open Directory project, Google's web directory provides a fast and easy way to search within a certain topic or related topics. The Advertising Programs introduced by Google targets users by providing advertisements that are relevant to search query. This does not bother users with irrelevant ads and has increased the clicks for the advertising companies by four or five times. According to BtoB, a leading national marketing publication, Google was named a top 10 advertising property in the Media Power 50 that recognizes the most powerful and targeted business-to-business advertising outlets. One of the latest services offered by Google is,' Google News' It integrates news from the online versions of all newspapers and organizes them categorically to make it easier for users to read "the most relevant news". It seeks to provide information that is the latest by constantly retrieving pages that are being updated on a regular basis. The key feature of this news page, like any other Google service, is that it integrates information from various Web news sources through purely algorithmic means, and thus does not introduce any human bias or effort. However, the publishing industry is not very convinced about a fully automated approach to news distillations

Web-wide tracking – Doubleclick :- Web-wide tracking', i.e. tracking an individual across all sites (s) he visits is one of the most intriguing and controversial technologies. It can provide an understanding of an individual's lifestyle and habits to a level that is unprecedented – clearly of tremendous interest to marketers. A successful example of this is DoubleClick Inc.'s DART ad management technology. DoubleClick serves advertisements, which can be targeted on demographic or behavioral attributes, to the end-user on behalf of the client, i.e. the Web site using DoubleClick's service. Sites that use DoubleClick's service are part of 'The DoubleClick Network' and the browsing behavior of a user can be tracked across all sites in the network, using a cookie. This provides DoubleClick's ad targeting to be based on very sophisticated criteria.

Understanding Web communities – AOL:- One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base. A large portion of this customer base participates in various 'AOL communities', which are collections of users with similar interests. In addition to providing a forum for each such community to interact amongst themselves, AOL provides useful information, etc. as well. Over time, these communities have grown to be well-visited 'waterholes' for AOL users with shared interests.Applying Web mining to the data collected from community interactions provides AOL with a very good understanding of its communities, which it has used for targeted marketing through ads and e-mail solicitations. Recently, it has started the concept of 'community sponsorship', whereby an organization like Nike may sponsor a community called 'Young Athletic Twenty Something's'. In return, consumer

survey and new product development experts of the sponsoring organization get to participate in the community – usually without the knowledge of the other participants. The idea is to treat the community as a highly specialized focus group, understand its needs and opinions on new and existing products; and also test strategies for influencing opinions.

Understanding auction behavior – eBay :- As individuals in a society where we have many more things than we need, the allure of exchanging our 'useless stuff' for some cash – no matter how small – is quite powerful. This is evident from the success of flea markets, garage sales and estate sales. The genius of eBay's founders was to create an infrastructure that gave this urge a global reach, with the convenience of doing it from one's home PC [EBAYa]. In addition, it popularized auctions as a product selling/buying mechanism, which provides the thrill of gambling without the trouble of having to go to Las Vegas. All of this has made eBay as one of the most successful businesses of the Internet era.Unfortunately, the anonymity of the Web has also created a significant problem for eBay auctions, as it is impossible to distinguish real bids from fake ones. EBay is now using Web mining techniques to analyze bidding behavior to determine if a bid is fraudulent. Recent efforts are towards understanding participants' bidding behaviors/patterns to create a more efficient auction market.

Personalized Portal for the Web – My Yahoo:- Yahoo was the first to introduce the concept of a 'personalized portal', i.e. a Web site designed to have the look-and-feel as well as content personalized to the needs of an individual end-user. This has been an extremely popular concept and has led to the creation of other personalized portals, e.g. Yodlee for private information. Mining My Yahoo usage logs provides Yahoo valuable insight into an individual's Web usage habits, enabling Yahoo to provide compelling personalized content, which in turn has led to the tremendous popularity of the Yahoo Web site.

3.WEB MINING TECHNIQUES

Traditional data mining techniques can also be used for web mining, such as classification, clustering, association rule mining, and visualization. In web mining, classification algorithms can be used to classify users into different classes according to their browsing behavior, for example according to their browsing time. After classification, a useful classification rule like "30% of users browse product/food during the hours 8:00-10:00 PM" can be discovered. The difference between classification and clustering is that the classes in classification are predefined (supervised), but in clustering are not predefined (unsupervised). The criterion by which items are assigned to different clusters is the degree of similarity among them. The main purpose of Clustering is to maximize both the\similarity of the items in a cluster and the difference between clusters .The association rule technique can be used to indicate pages that are most often referenced together and to discover the direct or indirect relationships between web pages in users' browsing behavior. For example, an association rule in the web usage mining area could take the form "the people who view web page index.htm and also view product.htm the support=50% and the confidence=60%". Visualization is a special analytical technique in web mining that allows data and information to be understood or recognized by human eyes by using graphical and visualized means to represent data, information and analysis results .In web structure mining, it usually plays an important role in illustrating the structure of hypertexts and links in a website or the linking relationship between websites. For the other two types of web mining technique, visualization is also an ideal tool to model the data or information. For example, a graph (or map) can be used for web usage mining to present the traversal paths of users or a graph may show information about web usage. This approach enables the analyst to understand and efficiently interpret the results of web usage mining.

Association Rule:- After transactions are detected the pre-processing phase, frequent item-sets are discovered using the A-priori algorithm The support of item-set I is defined as the fraction of transactions that contain I and is denoted by $\sigma(I)$. A hyper graph is an extension of a graph where each hyper edge can connect more than two vertices. A hyper edge connects URLs within a frequent item-set. Each hyper edge is weighted by the averaged confidence of all the possible association rules formed on the basis of the frequent item-set that the hyper edge represents. The hyper edge weight can be perceived as a degree of similarity between URLs (vertices).

Sequential Pattern:- patterns are used to discover frequent sub sequences among large amount of sequential data. In web usage mining, sequential patterns are exploited to find sequential navigation patterns that appear in users sessions frequently. The typical sequential pattern has the form[15]:the 70% of users who first visited A.html and then visited B.html afterwards ,in the same session, have also accessed page C.html. Sequential patterns might appear syntactically similar to association pattern mining.

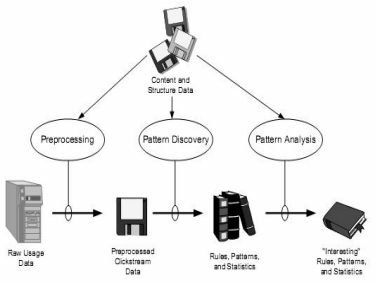
Clustering:- techniques look for groups of similar items among large amount of data based on a general idea of distance function which computes the similarity between groups. Clustering has been widely used in Web Usage Mining to group together similar sessions. Besides information from Web log files, customer profiles often need to be obtained from an on-line survey form when the transaction occurs. For example, you may be asked to answer the questions like age, gender, email account, mailing address, hobbies, etc. Those data will be stored in the company's customer profile database, and will be used for future data mining purpose.

4. WEB USAGE MINING

There are three main tasks for performing web usage mining and web usage analysis.

A.Pre-processing:- Pre-processing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

Usage Pre-processing:- Usage pre-processing is arguably the most difficult task in the Web Usage Mining process due to the incompleteness of the available data. Unless a client side tracking mechanism is used, only the IP address, agent, and server side click stream are available to identify users and server sessions. Some of the typically encountered problems are: Single IP address/Multiple Server Sessions – Internet service providers (ISPs) typically have a pool of proxy servers that users access the Web through. A single proxy server may have several users accessing a Web site, potentially over the same time period. Multiple IP address/Single Server Session - Some ISPsor privacy tools randomly assign each request from a user to one of several IP addresses



In this case, a single server session can have multiple IP addresses. Multiple IP address/Single User - A user that accesses the Web from different machines will have a different IP address from session to session. This makes tracking repeat visits from the same user difficult. Multiple Agent/Singe Users - Again, a user that uses more than one browser, even on the same machine, will appear as multiple users. Assuming each user has now been identified (through cookies, logins, or IP/agent/path analysis); the click-stream for each user must be divided into sessions. Since page requests from other servers are not typically available, it is difficult to know when a user has left a Web site. A thirty minute timeout is often used as the default method of breaking a user's click-stream into sessions. The thirty minute timeout is based on the results of [11]. When a session ID is embedded in each URI, the definition of a session is set by the Content server. While the exact content served as a result of each user action is often available from the request field in the server logs, it is sometimes necessary to have access to the content server information as well. Since content servers can maintain state variables for each active session, the information necessary to determine exactly what content is served by a user request is not always available in the URI. The final problem encountered when pre-processing usage data is that of inferring cached page references.

Content Preprocessing:- Content pre-processing consists of converting the text, image, scripts, and other files such as multimedia into forms that are useful for the Web Usage Mining process. Often, this consists of performing content mining such as classification or clustering. While applying data mining to the content of Web sites is an interesting area of research in its own right, in the context of Web Usage Mining the content of a site can be used to filter the input to, or output from the pattern discovery algorithms. For example, results of a classification algorithm could be used to limit the discovered patterns to those containing page views about a certain subject or class of products. In addition to classifying or clustering page views based on topics, page views can also be classified according to their intended use [50;30]. Page views can be intended to convey information (through text, graphics, or other multimedia), gather information from the user, allow navigation (through a list of hypertext links), or some combination uses. The intended use of a page view can also filter the sessions before or after pattern discovery. In order to run content mining algorithms on page views, the information must first be converted into a quantiable format. Some version of the vector space model [51] is typically used to accomplish this. Text files can be broken up into vectors of words. Keywords or text descriptions can be substituted for graphics or multimedia.

Structure Pre-processing:- The structure of a site is created by the hypertext links between page views. The structure can be obtained and pre-processed in the same manner as the content of a site. Again, dynamic content (and therefore links) pose more problems than static page views. A different site structure may have to be constructed for each server session.

B.Pattern Discovery:- Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Methods developed from other fields must take into consideration the different kinds of data abstractions and prior knowledge available for Web Mining. For example, in association rule discovery, the notion of a transaction for market-basket analysis does not take into consideration the order in which items are selected. However, in Web Usage Mining, a server session is an ordered sequence of pages requested by a user.

Statistical Analysis:- Statistical 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 (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path. Many Web analysis tools

produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This report may include limited low-level error analysis such as detecting unauthorized entry points or finding the most common invalid URI. Despite lacking in the depth of its analysis, this type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for marketing decisions.

Association Rules:- Association rule generation can be used to relate pages that are most often referenced together in a single server session. In the context of 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. For example, association rule discovery using the Apriori algorithm (or one of its variants) may reveal a correlation between users who visited a page containing electronic products to those who access a page about sporting equipment. Aside from being applicable for business and marketing applications, the presence or absence of such rules can help Web designers to restructure their Web site. The association rules may also serve as a heuristic for prefetching documents in order to reduce user-perceived latency when loading a page from a remote site.

Clustering:- Clustering is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for Internet search engines and Web assistance providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs.

Classification:- Classification is the task of mapping a data item into one of several predefined classes In the Web domain; one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc. For example, classification on server logs may lead to the discovery of interesting rules such as: 30% of users who placed an online order in /Product/Music are in the 18-25 age groups and live on the West Coast.

Sequential Patterns:- The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. Other types of temporal analysis that can be performed on sequential patterns include trend analysis, change point detection, or similarity analysis.

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. As an example, one may be interested to build a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (I e. from a casual visitor to a serious potential buyer). There are several probabilistic learning techniques that can be employed to model the browsing behavior of users. Such techniques include Hidden Markov Models and Bayesian Belief Networks. Modeling of Web usage patterns will not only provide a theoretical framework for analyzing the behavior of users but is potentially useful for predicting future Web resource consumption. Such information may help develop strategies to increase the sales of products offered by the Web site or improve the navigational convenience of users.

C.Pattern Analysis:- Pattern analysis is the last step in the overall Web Usage mining process as described in Figure 1. The motivation behind pattern analysis is to alter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to alter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

5.DATA PREPROCESSING IN WEB USAGE MINING

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. There are several preprocessing tasks that must be performed prior to applying data mining algorithms to the data collected from server logs. Data preprocessing is the process to convert the raw data into the data abstraction necessary for the further applying the data mining algorithm .There are several data preparation techniques that can be used to improve the performance of data preprocessing in order to identify unique users and user sessions. Generally, Web Usage Mining consists of three processes

data preprocessing, patterns discovery and patterns analysis.

ideally, the input for the Web Usage Mining process is a user session file that gives an exact account of who accessed the Web site, what pages were requested and in what order, and how long each page was viewed. A user session is the set of the page accesses that occur during a single visit to a Web site. However, because of the reasons we will discuss in the following, the information contained in a raw Web server log does not reliably represent a user session file before data preprocessing. Generally, data preprocessing consists of data cleaning,user identification, session identification and path completion,

Phases of Data Preprocessing in Web Usage Mining

Data Cleaning:- This task of data cleaning is to remove the irrelevant and redundant log entries for the mining process. There are three kinds of irrelevant or redundant data needed to clean: accessorial resources embedded in HTML file, robots' requests and error requests.

1) Accessorial Resources:- Because HTTP protocol is connectionless, a user's request to view a particular page often results in several log entries since graphics and scripts are down-loaded in addition to the HTML file. Since the main intent of Web Usage Mining is to get a picture of the user's behavior, it does not make sense to include file requests that the user did not explicitly request. Elimination of the

items deemed irrelevant can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes such as gif, jpeg, GIF, JPEG, jpg, JPG, css and map can be removed. In addition, common scripts such as the files requested with the suffixes of ".cgi" can also be removed.

2) Robots' requests:- Web robots (also called spiders) are software tools that scan a Web site to extract its content. Spiders automatically follow all the hyperlinks from a Web page. Search engines such as Google periodically use spiders to grab all the pages from a Web site to update their search indexes [8]. To remove robots' request, we can look for all hosts that have requested the page"robots.txt".

3) Error's requests:- Error's requests are useless for mining process. They can be removed by checking the status of request. For example, if the status is 404, it is shown that the requested resource is snot existence. This log entry in log can be removed then

User Identification:- A user is defined as the principal using a client to interactively retrieve and render resources or resource manifestations. User identification is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. The Web Usage Mining methods that rely on user cooperation are the easiest ways to deal with this problem. However, it's difficult because of security and privacy.

User Session Identification:- A user session means a delimited set of user clicks (click stream) across one or more Web servers. The goal of session identification is to divide the page accesses of each user into individual sessions. At present, the methods to identify user session include timeout mechanism [9] and maximal forward reference [10] mainly. The following is the rules we use to identify user session in our experiment:

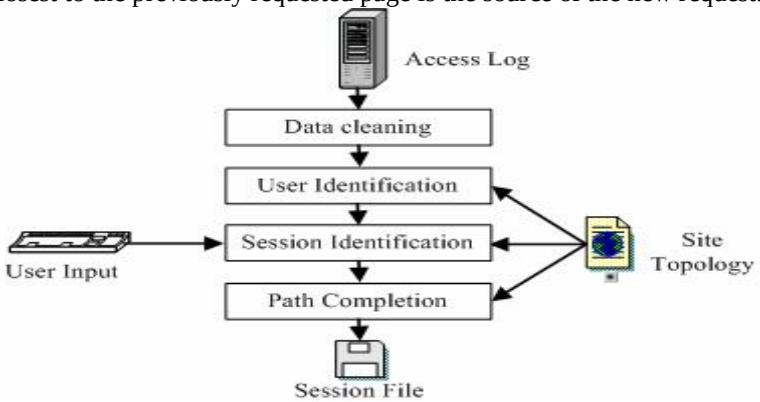
1) If there is a new user, there is a new session;

2) In one user session, if the refer page is null, there is a new session;

3) If the time between page requests exceeds a certain limit (30 or 25.5mintes), it is assumed that the

user is starting a new session.

Path Completion:- As the existence of local cache and proxy server, there are many important accesses that are not recorded in the access log. The task of path completion is to fill in these missing page references. Methods similar to those used for user identification can be used for path completion. If a page request is made that is not directly linked to the last page a user requested, the referrer log can be checked to see what page the request came from. If the page is in the user's recent request history, the assumption is that the user backtracked with the "back" button available on most browsers, calling up cached versions of the pages until a new page was requested. If the referrer log is not clear, the site topology can be used to the same effect. If more than one page in the user's history contains a link to the requested page, it is assumed that the page closest to the previously requested page is the source of the new request.



6.CONCLUSION

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper, we are discussing about web usage mining. Web usage mining is a kind of mining to server logs. Web usage mining places an important role in realizing, enhancing the usability of the website design, providing personalization server and other business making decision etc.. Our hope is that this overview provides a starting point for fruitful discussion

REFERENCES

[1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000, Vol. 1(2):1-12
[2] Li Chaofeng. Data Source Analysis on Web Usage Mining. Journal of south-central university for nationalities, 2005(4):82-85(in Chinese)
[3] Mobasher B., Dai H., Luo T, Nakagawa M. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery, 2002, 6(1): 61–82.
[4] Shahabi C., Kashani F.B.. A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking. Proc. WEBKDD 2001: Mining Web Log Data across All Customer Touch Points, LNCS 2356, Springer-Verlag, 2002: 113-144.
[5] Zhang Feng, Chang Huyou. Research and development in Web usage mining system-key issues and proposed solutions: a survey. Machine Learning and Cybernetics, 2002(2):986-990
[6] Berendt B., Mobasher B., Nakagawa M., Spiliopoulou M.. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. Proc. WEBKDD 2002: Mining Web Data for Discovery Usage Patterns and Profiles, LNCS 2703, Springer-Verlag, 2002:159-179
[7] Pirolli P., Pitkow J., Rao R... Silk from a sow's ear: Extracting usable structures from the Web. In: Proc. 1996 Conference on Human Factors in Computing Systems (CHI-96), Vancouver, British Columbia, Canada, 1996.
[8] Tanasa D., Trousse B.. Advanced data preprocessing for intersites Web usage mining. Intelligent Systems, IEEE, 2004(19): 59 – 65
[9] Catledge L., Pitkow J.. Characterizing browsing behaviors on the World Wide Web, Computer Networks and ISDN Systems ,1995,27(6):1065-1073.
[10] Chen M.S., Park J.S., Yu P.S.. Data mining for path traversal patterns in a web environment. In Proceedings of the 16th International Conference on Distributed Computing Systems, 1996:385-392.
[11] L. Catledge and J. Pitkow. Characterizing browsing behaviors on the World Wide Web. Computer Networks and ISDN Systems, 27(6), 1995.

AUTHORS

1] Swarna Latha Ampilli received the M.Tech Degree in Computer Science and Engineering from Andhra University. Currently working as Senior Assistant Professor in the Department of Information Technology at Dadi Institute of Engineering & Technology, Anakapalle, Visakhapatnam.

2] Suresh Tippana working as Assistant Professor At Dadi Institute of Engineering & Technology, Anakapalle, Visakhapatnam.