Introduction

CHAPTER 1
INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They search databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

1.1 WEB MINING

As a result of the quick development of the World Wide Web, the utilization of computerized Web mining strategies to find helpful, applicable data has turned into an undeniably essential exploration range. One imperative subarea is the Web utilization mining, wherein one endeavors to find the pattern of Web use from Web log information. Despite the fact that Web log information is generally noisy and to a great degree vague, there remains a potential for finding helpful structure in the associations between the website and its clients. Such information can be considered to produce inferences about website plan, test models of Web destinations or their adjustments, and test theories about the impacts of various design variables on Web-client behavior.

As a rule, Weblogs record clients' request to a Web server. A request is recorded in a log record section, which contains distinctive sorts of data, including the IP location of the PC making the request, the client access date and time, the report or picture asked for and so on. Depending upon the prominence of the webpage, a Web log can record thousands or a huge number of solicitations consistently. To discover valuable examples, (for example, affiliation rules or successive examples) from this large amount of data, demands (or log sections) should be assembled into utilization sessions (Monika Yadav Mr. Pradeep Mittal, 2013).

A session is characterized as a gathering of the request made by a single client for a solitary route reason. A client may have a single session or numerous sessions amid a time

frame. Just once these nuclear sessions have been distinguished can regular utilization designs among sessions be found by Web usage mining algorithm. The most regularly utilized session distinguishing proof strategy is called timeout, in which a client session is typically characterized as a grouping of solicitations from the same client such that no two back to back requests are isolated by an interim more than a predefined limit. This session recognizable proof technique experiences the issue that it is hard to set the time edge. Diverse clients may have distinctive route practices, and their time interval between sessions might be essentially diverse. Notwithstanding for the same client, intervals between sessions may fluctuate. A dynamic session recognizable proof technique that depends on the connection of solicitations would appear to be a great deal more proper.

Web structure mining, one of three classifications of web mining for information, is an area used to recognize the relationship between Web pages visited. This sort of mining can be utilized to uncover the structure of website pages; this would be useful for navigation and make it conceivable to incorporate Web page schemes.

The pattern of the navigation for a group of users having similar usage of the pattern is found by the clustering process. With the procedure of learning navigation designs, the data suppliers would be happy to view the change of the adequacy on their Web destinations, which results in adjusting the Website plan or by biasing the client's conduct towards fulfilling the objectives of the website. Online part utilizes the found examples to give customized substance to clients, in light of their current navigational action.

The potential key points in every area into mining objective is the prediction of the client's conduct inside the webpage, examination amongst expected and real Website utilization, alteration of the webpage according to the interests of its clients. There are no unmistakable qualifications between the Web utilization mining and other two classes. The web is the quickest developing field with expansion in the number of information accessible on it and additionally in the number of clients. Prediction frameworks give straightforwardness to the client to getting to as it were required data. The data in the weblogs get prioritized by the prediction mechanism. The forecasting of the websites can be done from the data present in the history. The relationship between the input variables and the output are analyzed by the mechanism of prediction.

The web mining also supports providing of suggestions to the users. When the user enters a query for search then the mining provides the suggested pages related to the user. The similarity search with high value gets recommended for the user search. The suggestions of the log files help to support the query of the user search.

1.2 TYPES OF WEB MINING

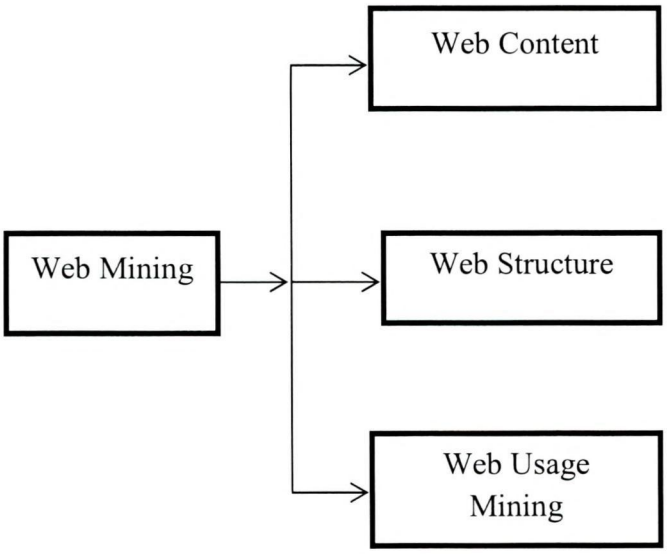

Figure 1.1 Web mining types

1.2.1 Web Content Mining

Web content mining is the mining, extraction and reconciliation of helpful information, data and learning from Web page substance. Content mining is the checking and mining of content, pictures and charts of a Web page to decide the importance of the substance to the query search. This filtering is finished after the grouping of site pages through structure mining and gives the outcomes based upon the level of significance to the recommended inquiry. With the enormous measure of data that is accessible on the World Wide Web, content mining gives the outcomes records to web search altogether of most elevated pertinence to the keywords in the inquiry (R. Kosala, H. Blockeel,2000).

1.2.2 Web Structure Mining

Web Structure Mining (Exploiting Hyperlink Structure) Web structure mining is a process of picking up information from linkages of web pages. The structure information is discoverable by the procurement of web structure pattern through database procedures for Web pages. This association permits a web search engine to draw information identifying

with an inquiry question straightforwardly to the connecting Web page from the Web website the substance rests upon. This completion happens through utilization of examining the Web sites, recovering the landing page, then, connecting the data through reference connections to deliver the particular page containing the sought information.(Jaideep Srivastava ,et al). Structure mining utilizes minimize two fundamental issues of the World Wide Web because of its tremendous measure of data. The first of these issues is unimportant list items. The pertinence of pursuit data gets to be confused because of the issue that search engines frequently take into consideration low exactness criteria. The second of these issues is the failure to record the immeasurable sum if data gave on the Web. This causes a low measure of review with substance mining. This minimization comes to some degree with the capacity of finding the model fundamental the Web hyperlink structure gave by Web structure mining.

The fundamental reason for structure mining is to extricate beforehand obscure connections between Web pages. This structure information digging gives use to a business to connect the data of its own Web website to empower route and bunch data into sitemaps. This permits its clients the capacity to get to the fancied data through watchword affiliation and substance mining. Hyperlink progression is likewise resolved to weigh the related data inside the locales to the relationship of competitor connections and association through search engines and unknown co-links (J. Srivastava et al, 1997). This empowers grouping of associated Web pages to build up the relationship of these pages.

- Extracting pattern from the hyperlink.
- Mining the structure of the document.

1.2.3 Web Usage Mining

Web usage mining is the way of separating valuable data from server logs i.e. clients history. Web usage mining is the procedure of discovering what clients are searching for on the Internet. A few clients may take a look at just literary information, while a few others may be keen on mixed media information. Web utilization mining process includes the log time of pages. This technology is essentially focused upon the utilization of the web advancements which could help for improvement. The world's biggest entryway like yahoo, msn and so on., has a great deal of requirements about the experiences from their clients' web visits. Without this use reports, it will be hard to structure their adaptation endeavors. Utilization mining has the direct effect on businesses (B. Masand, et al. 2002). This is the

movement that includes the programmed revelation of client access designs from one or more Web servers. As additional associations depend on the Internet and the World Wide Web to direct business, the customary procedures and systems for market investigation should be returned to in this connection. Associations regularly create and gather vast volumes of information in their day to day operations. A large portion of this data is normally produced naturally by Web servers and gathered in server access logs. Different wellsprings of client data incorporate a referrer log which contains data about the alluding pages for each page reference, and client enlistment or overview information accumulated through tools like CGI scripts (M. Spiliopoulou et al, 1999).

Breaking down such information can help these associations to decide the lifetime estimation of clients, cross advertising procedures crosswise over items, and adequacy of limited time battles, in addition to other things. Investigation of server access logs and client enlistment information can likewise give significant data on the most proficient method to better structure a webpage with a specific end goal to make a more viable nearness for the association. In associations utilizing intranet innovations, such investigation can reveal insight into the additional viable administration of workgroup correspondence and authoritative foundation. At long last, for associations that offer to publicize on the World Wide Web, breaking down client access designs encourages in focusing on advertisements to particular gatherings of users (R. Kohavi, 2001).

- Web server data
- Application server data
- Application level data

Web mining extracts the knowledge which is relevant to the search and the new information about the usage of logs. The pattern of the usage of the particular user gets stored in the web log files. The structure of the web, content present on the web helps to extract the useful patterns for a particular user. This information stored in the web log files which is present in three types. Each time when the user requests the server they will store as the log file (L.K. Joshila Grace, 2011).

i. Web Server Log

A server log is a log file that is automatically created and maintained by a server consisting of a list of activities performed. This is an automatic process of saving the data

about the user process. The substance of the record will be the same as it is talked about in the past subject. The server which gathers the individual data of the client must have a secured storage. The web server log helps to examine the traffic present in the day to day life about the requirement of that particular website and it can be managed in an effective manner. This server log doesn't collect the particular information.

ii. Browser Log

This sort of log documents can be made to live in the customer's program window itself. Uncommon sorts of programming exist which can be downloaded by the client to their program window. Despite the fact that the log document is available in the customer's program window the passages to the log record is done just by the Web server.

iii. Proxy Server Log

A Proxy server is said to be a transitional server that exists between the customer and the Web server. Therefore if the Web server gets a solicitation of the customer by means of the intermediary server then the sections to the log document will be the data of the intermediary server and not of the first client. These web intermediary servers keep up a different log document for a get-together the data of the client.

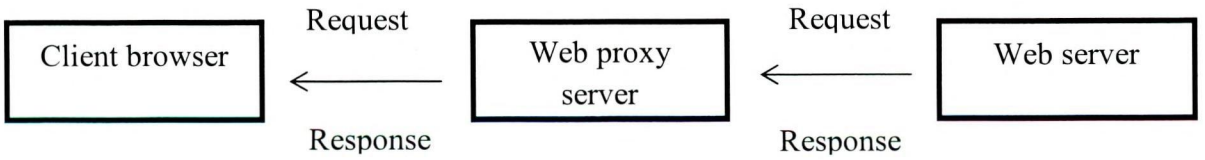

Figure 1.2 Log files in web proxy server

Challenges in the web usage mining

- Noisy data
- Curse of dimensionality
- Pattern navigation
- Session identification
- Security

Applications of web usage mining

- Personalization
- System improvement
- Business intelligence
- User characterization
- Identification of concept
- Website designing and development.

1.3 SESSION IDENTIFICATION

A session is characterized as a gathering of a request made by a single client for a solitary route reason. A client may have a single session or numerous sessions amid a timeframe. Just once these nuclear sessions have been distinguished can regular utilization designs among sessions be found by the web usage mining algorithm. The most regularly utilized session distinguishing proof strategy is called timeout, in which a client session is typically characterized as a grouping of solicitations from the same client such that no two back to back requests are isolated by an interim more than a predefined limit. This session recognizable proof technique experiences the issue that it is hard to set the time edge. Diverse clients may have distinctive route practices, and their time interval between sessions might be essentially diverse. Notwithstanding for the same client, intervals between sessions may fluctuate. A dynamic session recognizable proof technique that depends on the connection of solicitations would appear to be a great deal more proper.

The web log mining covers quite a while periods subsequently, clients may get to the website more than once. Session recognizable proof is keeping in mind the end goal to separate the entrance records into a few getting to arrangements, in which the pages are asked for in the meantime. Conventional session distinguishing proof calculation depends on a uniform and altered timeout. While the interim between two consecutive demands surpasses the timeout, a new session is decided. The underlying timeout is analyzed for every page as per the measurable result, consolidating with the significance level of page, system of session recognizable proof, time out is progressively balanced, client sessions are resolved judging by the element timeout.

1.4 IMPORTANCE OF NAVIGATION

Navigation on a website is accomplished by an accumulation of connections that frame the Website Navigation Menu or the Website Navigation Bar. This navigation menu or bar is typically the accumulation of links that the user saw see get graded vertically on the left or on a level plane close to the highest point of the page and some of the time on the footer of the website page.

In spite of the fact that it might appear to be insignificant, having sorted out and simple to-take after route for a website is extremely useful and imperative for the general client experience. Going to a website with scattered route resemble driving some place with dim bearing the user may get lost along the way on the grounds that the course is difficult to take after.

- **Expanded Visit Duration and Decreased Bounce Rate**

An effectively safe website expands the time a guest or client stays on your website. This permits guests more opportunity to investigate your website and find data about your organization. In the event that guests/clients visit your website and find that it is hard to explore, they might not have any desire to take an ideal opportunity to burrow through your website and may "skip." A bounce happens when a website guest just perspectives a solitary page on a website, as opposed to keeps, seeing different pages inside the same website.

- **Item Purchases**

It's simple to take the directions that are set up through key website route permit clients to easily experience the way toward survey and buying items.

- **Quick access**

On account of the rapid innovation that exists today, individuals like everything quick. Route bars with compact and clear classes permit individuals to quickly and effectively get to data about your organization.

- **General Design**

Since route bars are regularly put evenly at the highest point of a website or vertically on the left, it is essential to be predictable with these positions. Route bars put amidst a website page are non-standard and troublesome for guests to find. Institutionalized, sorted out and

uncluttered route outline expands the general refined submission of a website's configuration. Great navigation is a helpful technique to support website outline and page visits.

1.4.1 Benefits of Navigation Pattern

The examination of the pattern of navigation concentrates on the methods to contemplate the client conduct at the point when exploring a website. While the World Wide Web swings to be the biggest data asset accessible on the web, a consciousness of the navigation inclinations gets to be a fundamental step. It is not just during the time spent tweaking and adjusting the website's interface for people, also in enhancing the website's static structure of the hidden hypertext framework. Great learning in transit of guests navigate in a website could anticipate bewilderment and help the supplier to put the data appropriately.

The fundamental reason for navigation is to extricate already obscure connections between Web pages. This structure information digging gives use to a business to connect the data of its own webpage to empower route and bunch data into website maps. This permits its clients the capacity to get to the wanted data through watchword affiliation and substance mining.

1.5 PROBLEM JUSTIFICATION

Huge amount of data are present on the internet, the user can refer a large number of websites in order to obtain the desired information. In the web servers, log repositories play a key role as it keeps the record of user pattern for different users and thus it is great a source of knowledge. Web Usage Pattern is the process of getting the web user browsing patterns by analyzing their navigational behavior. A modern process of identifying a web user session is the average time spent by the user on pages. The problem lies in identifying the sessions which are useful for the user. The few existing algorithms that can be applied to find such contiguous approximate pattern mining have drawbacks like poor scalability, lack of guarantees in finding the pattern, and difficulty in adapting to other applications. Flame is a sequential pattern mining and therefore play an important role in finding the frequent pattern with a variety of definitions of pattern models.

1.6 PROBLEM STATEMENT

To device a strategy for improving the web page navigation prediction in web usage mining. Better session identification is a major challenge and the predictions of the user behavior presents a difficult task.

1.7 OBJECTIVE OF THESIS

The major objectives formulated in the thesis are as follows,

- To provide an improved session identification phase so that the pattern predictive accuracy will be increased.
- To predict user navigation patterns effectively using identified session.

1.8 ORGANIZATION OF THESIS

The rest of the thesis is organized as follows:

Chapter 2, Literature review presents the process of various methodologies in the area of web page navigation pattern prediction in the web usage mining.

Chapter 3, Methodology represents the method of implementation of DBSCAN, FP-growth, and Flame algorithm. This chapter represents how the proposed methods are implemented and the overall performance of the research is given.

Chapter 4, Results and Discussion, this chapter discusses the dataset used for the implementation, the experimental setup required to carry out the process. The result obtained with the experimental setup and the data set is analyzed.

Chapter 5, Summary and conclusion, this chapter summarizes the thesis work and presents the future direction.

1.9 CHAPTER SUMMARY

This chapter discussed the overview of session identification and navigation pattern prediction in the web usage mining. The web page navigation prediction and the objectives are discussed in this chapter.

The next chapter discusses the review of literature based on web page navigation pattern prediction in the web usage mining using various techniques.