CHAPTER 3

METHODOLOGY

Customer Relationship Management (CRM) may utilize data from within or outside a company to allow better understanding of its clients on a group basis or on a personal basis, by creating customer personal records. An enhanced understanding of the consumer's habits, interests and needs can raise up the trade. Thus, consistent information about the customers' preferences and requirements forms the basis for efficient CRM. As businesses become online (that is, develop into e-business), the struggle to keep the faithfulness of their old clients and to attract new customers is still more important, because a competitor's business site may be just one click away.

Voluminous of data existing in these online web have made it very important to use automated data mining and knowledge discovery processes to find out user navigation preferences. The different modes of website usage by a particular user may be discovered by means of Web usage mining techniques which can automatically retrieve frequent access patterns using the history of earlier user clickstreams used in web log files. These portfolios can be used towards personalizing the web site for the user and to support electronic marketing.

Web usage mining technology integrates techniques from two popular research fields, namely, data mining and the Internet. By analyzing the potential knowledge hidden in web logs, web usage mining can help webmasters to provide better design and business concerns to provide better navigation behavior. Many industries are focusing on customer orientation to retain frequent users for the development of customer relationship management. Study of interested web browsers, gives valuable information for web site designer to quickly react to

their individual needs. This chapter presents the research methodology used to design the next page prediction system.

3.1. WEB LOG DATA

Web log data is a file which has huge amount of information and from this data source, several data abstractions can be created. Examples include page views, server sessions, and click-streams. In all these abstracts, common terms and keywords are used as given in Table 3.1. This section also provides a detailed description of the weblog file format used in the present research work.

A log file is defined as a file that registers the operations of a web server. Log files yields information such as the files that are requested, the time of the file request, the user and the referring page. Every line of the log file defines a single "hit" on the log file in the server and contains a number of fields and the format of the log used for analyzes differ from server to server. Analysis of log file is considered advantageous for the following reasons:

- The web server already produces log files, so obtaining a raw data is not very difficult and does not require any modifications or additional programming effort,
- Company's own servers will maintain data in their own standard. This makes it trouble-free for a company to change programs later, use many different programs and examine chronological information with a new program,
- Creation and adding details to the log file does not require any additional Domain Name Server Lookups. Hence, there are no outside server calls that can slow down page load speeds, and leads to uncounted page views, and
- The web site's server records all transaction it makes and therefore is considered reliable.

The format of the log file is shown in Table 3.2. An hyphen ('-') in any of these fields indicates missing data.

TABLE 3.1

IMPORTANT TERMS IN WEB LOG DATA

S.No.	Terms	Description
1	User	Users accessing file from the web servers through a browser.
2	Page View	A page view is an abstract that consist of every file that is displayed on user's browser screen at one point of time. A page view may be associated with a single user action or can be related with several files such as scripts ,frames,and graphics, etc.,
3	Hit	Every successful file that is sent to the web browser is a hit
4	Click Stream	It is a sequential series of page view requests.
5	Server Session or visit	A Server Session or visit happens when a user or robot visits a website.
6	User Session	A user session is defined as a set of page requests made by a single user.
7	Web Log	These are files that stores into them details regarding all the visits made to a web site or a portal automatically and are maintained in the web server.

TABLE 3.2

WEB LOG FILE FORMAT

S. No.	Name of Field	Description	Example value
1	IP Address	IP address of the Client who request for a page on the web server	127.0.0.1
2	UserID and Password	Provides the username and their corresponding password used during the access of a content-secured transaction	Voder23 12ert35
3	Timestamp	The date, time and time zone when the server finished processing the request.	[10/Oct/2000:13:55:36 -0700]
4	Access Request	Request line from the client. It has three parts, the METHOD, URL STEM and PROTOCOL used during transmission.	GET http/www.yahoo.com/asctab31.zip HTTP/1.0
	Method	Can be GET (request made to get a program or document) or POST (during transmission indicates the server that data is following) or HEAD (used by link checking programs, not browsers and downloads just the information in the HEAD tag information)	GET POST HEAD
	URL	the address of the web content to be retrieved.	/download/windows/asctab31.zip
	Protocol	protocol used during	HTTP/1.0

S. No.	Name of Field	Description	Example value
		transmission along with the version number.	
5	Status	It gives the status of a web transaction	200
6	Bytes	The number of bytes transferred.	3784
7	User Agent	The User Agent is whatever software the visitor used to access this site. It is usually a browser, but it could equally be a web robot, a link checker, an FTP client or an offline browser.	Mozilla/4.7 [en]C-SYMPA (Win95; U)
8	Referral	Refers to the previous page visited by the same user	/movies/tamil

3.2. RESEARCH FRAMEWORK

Internet is a client/server design where a client sends a web requests for over the World Wide Web (WWW) to a web server. The web server responds by responding to the request. The transaction session involves the exchange of methods and protocols. Nevertheless, due to exponential rise of WWW, there are a large number of clients who interacts with the servers using a large number of networks connected with one another, leading to a considerable boost the WWW latency and load on the net. When a proxy server placed in between a browser and a server, it is an efficient tool that may be employed to reduce the WWW's latency. This means that, it can intercept any requests to the web server to ensure whether the request can be fulfilled by the client itself. If not, then it may be

forwarded to the web server. The presence of proxy servers provides two main advantages as given below.

- Reduce latency: Periodically, all the requested results from various clients are stored in a Proxy server. For instance, consider when two users A and B access the net through a proxy server. Assume user A requests for a particular web page (P1). Sometime later, user B also requests for the same page. Without forwarding the request to the web server, this page is returned by the proxy server its cache where the recently downloaded web pages are retained. Since proxy server and the user share the same network, the operations are much faster, hence reducing the perceived latency, and

- Filter Unwanted Requests: Unwanted requests are removed by the Proxy servers. For example, a college may restrict the students from accessing a specific set of web sites by using a proxy server.

To further reduce the WWW latency, the behavior of the user can be predicted and accordingly the predicted pages are prefetched and stored temporarily in the cache of the proxy server. The request of the user can be fulfilled quickly if the requested page is available in the cache. A general web page prediction model is shown in Figure 3.1.

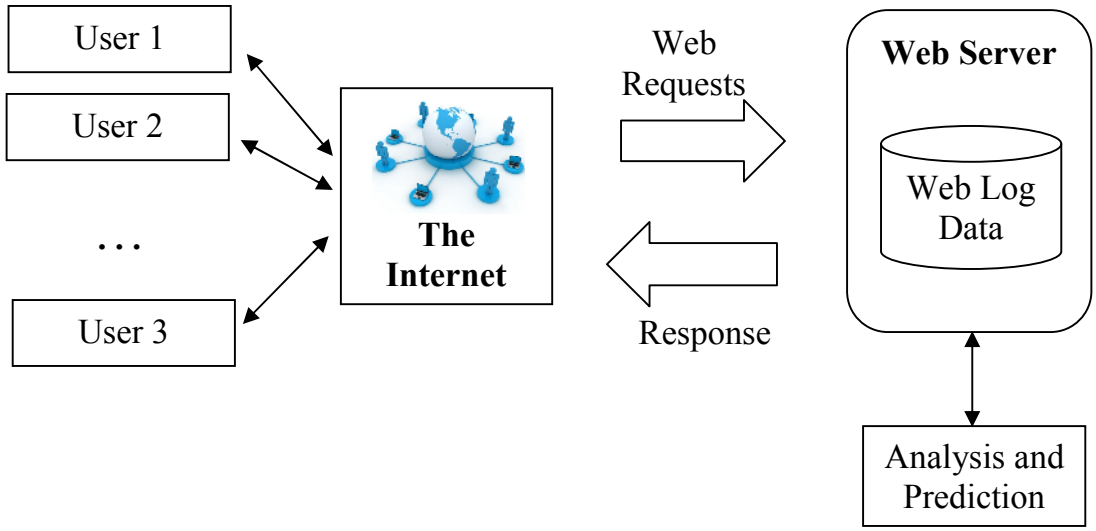

Figure 3.1 : General Architecture of Web System With Web Access Prediction

Prediction of user's consecutive steps in his/her communication with the Web poses a big challenge for researchers in the web engineering area and is the main focus of this research. As discussed in Chapter 1 (Introduction), prediction of user's future requests consists of various tasks and Figure 3.2 presents the flow of these tasks in the present research work. The proposed system is termed as next page prediction system. This work consists of three main steps, namely, preprocessing, potential user identification and prediction of future requests.

In this research work, each of the above mentioned steps is treated as a separate phase, which has to be applied in a sequential manner during the design and implementation of web page prediction system. The research methodology is planned in a manner that each step attempts to improve its respective task and works with the aim of improving its operation on prediction. During the flow of prediction, the output of one phase is used as input by the subsequent phase. The proposed research framework is presented in Figure 3.3 and the various techniques enhanced during the design of next page prediction system are introduced in the following subsections.

3.2.1. Phase I : Preprocessing Algorithms

Preprocessing of a web log file is nothing but simply reformatting the entries of a log file into a form that can be used directly by the subsequent steps of the log analyzer (Jalali *et al.,* 2008).

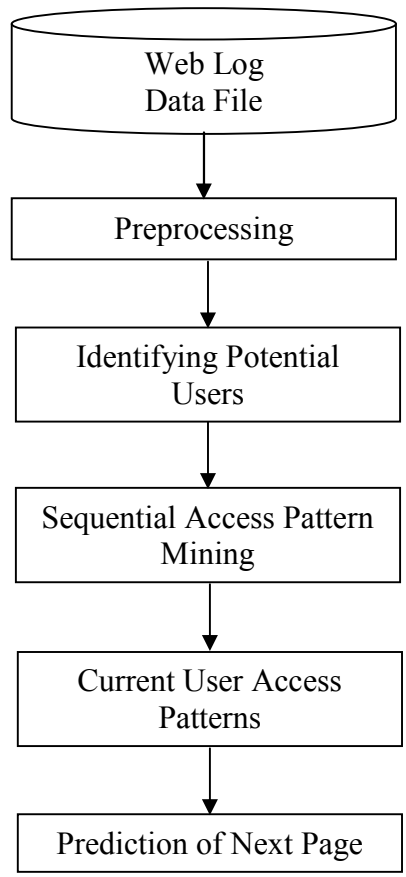

Figure 3.2 : Tasks in next page prediction system

Preprocessing of web data is essential in order to make the raw web log data more suitable for mining. The preprocessing is performed in four steps as given below:-

1. Cleaning,
2. User identification,
3. Session identification, and
4. Formatting

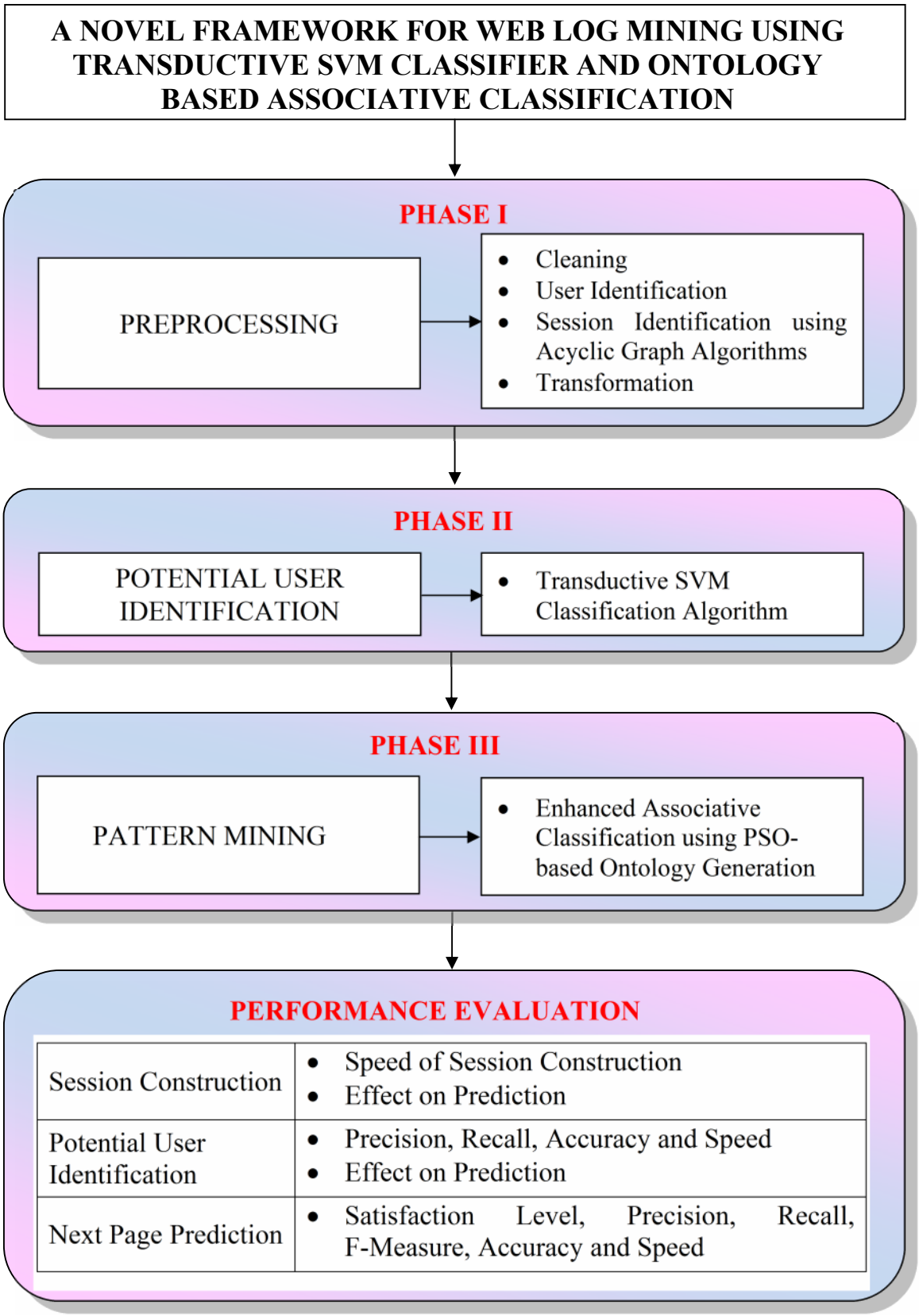

Figure 3.3 : Research Framework

All these steps are considered mandatory in next page prediction system, as these steps improve the process of extracting regularities of users' access behavior as patterns. The regularities of users' access behavior are called as patterns and the patterns are defined by combinations, orders or structures of the pages accessed in a session.

Cleaning is the process of removing unwanted or irrelevant data from the web log data file. The cleaning process used by the proposed web page prediction system uses algorithms which remove data that do not have any significant effect on the efficiency of subsequent steps of prediction.

The second important step of preprocessing is user identification, where individual users who access a web site are identified. Because of the capability of the internet and the advancement of the technology which is used for browsing the web, the users may not be identified in a insignificant way. In web page prediction system, this process of filtering and labeling unique users in the web log data is performed using an IP-address based method.

Session identification is used to divide the access records into several accessing sequences, in which the pages are requested. The main goal here is to segregate the page accesses of each user into individual sessions. Web sequential patterns are important for analyzing and understanding users' behaviour to improve the quality of service offered by the World Wide Web.

In this research, acyclic graph-based approach is analyzed for session identification of each user. Acyclic data graph is a data structure used for sequential pattern mining to derive patterns based on the directional pattern growth approach. This approach provides an easy way to represent groups that have underlying partial order. Graph-based approaches are analyzed, due to its recent success in various real world problems which the patterns have been modeled as traversals on graph and mining from these traversals have provided

effective results. Examples of such proposals include Chen (2010), Srikantaiah *et al.* (2013), Raguraman *et al.* (2013) and Iswarya and Sivakumar (2013).

The study analyzes four different acyclic graph-based approaches for user session identification as listed below. The methods are mainly analyzed in their ability to perform user's page transition in each session to a directed graph. The approaches analyzed are:-

- Directed Acyclic Graph-based Approach,
- Hierarchical Directed Acyclic Graph-based Approach,
- Partial Ancestral Graph-based Approach, and
- Mixed Ancestral Graph-based Approach.

Thus, the various steps of preprocessing converts the original raw web log file data, into a form that is pruned from unwanted data and formatted into a form that is readily accessible by the subsequent steps of next page prediction system. Detailed description of the methods used in this step are presented in Chapter 4, **Preprocessing Algorithms** and the performance advantages obtained by these steps to next page prediction system are discussed in Chapter 7, **Results and Discussion**.

3.2.2 Phase II : Potential User Identification

It is a well-known fact that the performance of an algorithm directly depends on the size of the dataset. In data mining, the curse of dimensionality is considered as a challenging and difficult problem and is an open area of research. In this research work, the second phase of next page prediction systemis used to reduce the size of the weblog dataset and is introduced also to identify the actual buyers of an e-commerce website.

Actual buyers or potential customers are in general identified using the time spent by n user in a webpage. This time, termed as browsing time, indicates their interest in the web content. The longer duration indicates the user has interest in the contents of the webpage and may become a potential customer. Studying the characteristics of these types of users will provide maximum advantage during the identification of users' browsing characteristics.

However, to extract these characteristics efficiently, just monitoring browsing time alone is not sufficient (Khilrani and Shandilya, 2011). According to Yan *et al.* (1996), one long access need not indicate that the user is actually a potential buyer and moreover, a long accessed transaction can completely obscure the importance of other relevant user accesses. Browsing time of the user in a page can be high due to various mode of actions like when the user is not navigating the web page throughout the time, is not attentive on the page during the navigation or navigate to some other application.

Thus, irrespective of the time spent by the user on the page, a user can either be potential (users with purchase interest) or non-potential (users without purchase interest) customer. As all e-commerce companies are more targeted towards actual buyers, this second phase of this study, groups the web log data into two groups, namely, potential and non-potential users. By pruning out the non-potential users from the preprocessed web log data, the size of web log data file can be reduced considerably. As the size of the web log data file has a direct impact on the analysis and prediction tasks, the performance of next page prediction systemcan be improved considerably.

For this purpose, this research work proposes the use of machine learning classification algorithm. As web log data are considered as unlabelled records, a semisupervised algorithm called Transductive Support Vector Machine Algorithm (TSVM) is used. For this purpose, a total of nine attributes, grouped into three

categories, are used. The three categories are temporal attributes, page attributes and communication attributes. Detailed description on these three groups of attributes along with the working of TSVM is presented in Chapter 5, Reducing Weblog Dataset Size. The performance of this classifier in identifying potential customers is presented and discussed in Chapter 7, Results and Discussion.

3.2.3 Phase III : Web Log Associative Classification

As stated in Srivastava *et al.* (2000), pattern discovery "draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition". Several methods and techniques have already been developed for this step. Some of the possible solutions are statistical analysis, clustering, classification, association rules, sequent patterns and dependency modeling. This research work uses associative classifiers for predicting future web page requests of a user.

According to Dolan (2002)], human emotion plays an important motivation factor during purchase and a customer buying behavior changes according to their mood and emotions. As the Internet is a medium where huge number of technology savvy people connect to the Internet for finding information pertinent to products and services before they commit to a purchase. Emotions have been found to influence a person's web surfing behaviors. This has been studied by Kalback (2007), Miao *et al.* (2007) and Choi and Ahn (2009). All these studies analyzed the effect of discovering and modeling consumers' emotions and surfing habits behaviors on many web applications such as personalized web search.

To relate emotions with user's navigation, this research work combines a recent concept called, semantic web usage mining (Berners-Lee *et al.*, 2001) with associative classification. The goal here is to associate each requested web page with one or more ontological entities to better understand the pattern of web navigation. A semantic web usage mining approach is one of the techniques which

is used to automatic creation of periodic web access pattern. Earlier, web usage mining method are focused on mining frequent access patterns which have occurred recurrently within the entire duration of all the user access sessions. So, this method analyzes the frequently used resources at a particular time period. Furthermore, ontology is generated to collect web access behaviors and emotional influence of the users for the specific resources.

However, analysis of this approach while predicting users' future requests revealed that the satisfaction level of the customers and prediction accuracy of the system still has room for improvement. This improvement is brought forward in this research, through the use of Particle Swarm Optimization (PSO). The PSO algorithm is used to identify optimal session intervals, which are then used to predict future requests using semantic web usage ontology. The proposed system, termed as OSIPSO, consists of various steps like generation of session intervals, application of PSO and of associative classification. Each of these steps has several sub-tasks, the details of which are discussed in Chapter 6, **Web Log Associative Classification**. The results of performance evaluation is tabulated and discussed in Chapter 7, **Results and Discussion**.

3.3. CHAPTER SUMMARY

This chapter presents the overall approach used in the research work. Details regarding the techniques used by Phase I of the study to preprocess the raw web log data is presented in the following chapter, Chapter 4, **Preprocessing Algorithms**.