Review of Literature

2. REVIEW OF LITERATURE

Identifying Web browsing strategies is a crucial step in Website design and evaluation, and requires approaches that provide information on both the extent of any particular type of user behaviour and the motivations for such behaviour (Clark *et al.*, 2006). This information can be retrieved from the server log files and forms the main focal point of the present work. The various works proposed with regard to the present is presented in this chapter.

2.1. WEB USAGE MINING

Web mining has been successfully applied to various areas like research trend identification (Tho *et al.*, 2003), robot detection and filtering to separate human and non-human behaviour (Kohavi, 2001), user profiling (Masand *et al.*, 2002), fraud and threat analysis (Lazarevic *et al.*, 2002) and identifying web communities (Gibson *et al.*, 1998). Several tools have been developed which analyze web log data to provide knowledge that will help web administrators in building effective websites. These tools help to understand the web usage process, web structure or web content from the web data. Chi *et al.* (1998) developed a Web Ecology and Evolution Visualization (WEEV) tool to understand the relationship between Web content, Web structure and Web Usage over a period of time. The site hierarchy is represented in a circular form called the "Disk Tree" and the evolution of the Web is viewed as a "Time Tube". This research work is focusing on extracting useful information from web data using web usage mining techniques. Hsien *et al.* (2005) described a method of web usage mining approach to discover patterns in the navigation of websites known as Unexpected Browsing Behaviours (UBBs) and called their technique as UBB mining. Web designers can review these UBBs to improve their website. Their results proved that the combination of predefined routing algorithm and UBB mining algorithm can discover interesting browsing patterns.

Web Usage Mining is the most sought-after tool in the Internet community where data from online web is converted to meaningful knowledge. The knowledge thus discovered can be used in web personalization, general system improvement, improve business intelligence, site modification and discover usage characteristics. The next section reviews some proposals made in web personalization, system and business intelligence through the use of web usage mining. Discovering usage characteristics from navigational patterns is the area of interest of the present research and is dealt in section 2.2.

2.1.1. Web personalization

Web personalization is the process of customizing a Web site to the needs of specific users, taking advantages of the knowledge acquired from the analysis of the user's navigational behavior (usage data) in correlation with other information collected in the Web context, namely, structure, content, and user profile data. In general, personalization techniques are divided into offline and online techniques. Offline personalization is based on simple user profiling and manual decision rule systems. Web usage mining is an online personalization data source. By evaluating site behavior and usage, a view about the website user is gained which yields to more effective personalization strategies. User profile is an important source of data for data personalization. Due to the explosive growth of the Web, the domain of Web personalization has gained great momentum both in the research and commercial areas (Cingil *et al.*, 2000).

Mobasher *et al.* (1999) propose an effective technique for capturing user profiles based on association rules discovery and usage based clustering combining with current status of an on-going activity to perform real time personalization. This was followed by the work of Toolan et al., (2002) proposed techniques based on web usage mining to deliver Personalized Site Maps that are specialized to the interest of each individual visitor.

Shahabi et al., (2003) described a complete framework for web-usage mining to satisfy the challenging requirements of web-personalization applications. They introduced a distributed user-tracking approach for accurate, scalable, and implicit collection of the usage data and proposed a feature-matrices (FM) model, to discover and interpret users' access patterns. A novel similarity measure based on FM was designed for accurate classification of partial navigation patterns in real time. This system worked well with both synthetic and real data for anonymous and efficient web personalization. It was at this period, web usage mining bloomed and a review of the popular techniques and tools available for web personalization was provided by Eirinaki et al., (2003).

Recently, Baraglia et al., (2007) proposed a dynamic personalization system which would personalize a site without the intervention of web users, using the information collected from log files and navigation pattern. In the same period, Ouamani *et al.* (2007) designed a web usage mining architecture for web personalization (PWUM) which was implemented using a multi-agent platform. This latter is composed of a set of autonomous agents interacting together in order to fulfill the main goal of the system. Agents are divided into modules that have well-defined tasks and that are further divided into two working groups: offline and online. The personalization agent uses the user model knowledge along with the previously discovered sequential patterns and applies a set of personalization rules in order to deliver the personalization tasks or functions like the memorization of personal information, user salutation, recommendation of links related to what users in the same group previously choose or links that the same user usually views, objects differentiation by presenting different features of each object.

2.1.2. General Site Improvement

Apart from developing web mining tools, work on general improvement of knowledge extraction from web data has also been proposed. Performance

and other service quality attributes are crucial to user satisfaction from services such as databases, networks, etc. Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission (Cohen, 1998), load balancing, or data distribution. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate (Fawcett et al., 1999). Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins, etc.

Almeida *et al.* (1996) propose models for predicting the locality, both temporal as well as spatial, amongst Web pages requested from a particular user or a group of users accessing from the same proxy server. The locality measure can be used for deciding pre-fetching and caching strategies for the proxy server. The increasing use of dynamic content has reduced the benefits of caching at both the client and server level. Schechter *et al.* (1998) has developed algorithms for creating path profiles from data contained in server logs. These profiles are used to regenerate dynamic Homepages based on the current user profile in order to reduce latency due to Page generation.

Cooley *et al.* (1999) in his investigation has successfully differentiated "web content mining" and "web usage mining". The contributions of Cooley and his research team-mates, Srivastava, Mobasher to web usage mining area is tremendous. During 1997, Cooley *et al.* presented a paper on the discovery and application of interesting patterns from web data using web mining techniques. This was followed by another work from the same team, (Srivastava *et al.*, 2000a) where they presented different methods for identifying and discovering usage patterns from web data.

This was followed by Seo *et al.* (2001) method of building intelligent systems for mining information and extraction rules from semi-structured Web pages by using domain knowledge. At the same period, Kohavi *et al.* (2001a) presented a method to mine log data across all customer touch points to extract web knowledge. Similarly, the same authors (Kohavi *et al.*, 2001b) also tested their system on e-commerce sites and identified the challenges and issues in it. Later in 2000, Srivastava *et al.* proposed hyperlink analysis technique and applied this technique to various applications and proved that their technique is superior to the existing methods. Madria *et al.* (1999) presented the issues involving web data mining and a comprehensive survey of various techniques is presented by Kosala *et al.* (2000).

2.1.3. Business Intelligence

Information on how customers use a Web site is critical information for marketers of e-businesses. Buchner *et al.* (1998) has presented a knowledge discovery process in order to discover marketing intelligence from Web data. They define a Web log data hypercube that will consolidate Web usage data along with marketing data for e-commerce applications. They identified four distinct steps in customer relationship life cycle that can be supported by their knowledge discovery techniques: customer attraction, customer retention, cross sales and customer departure.

There are several commercial products, such as Surf Aid, Accrue, Net-Genesis, Aria, Hotlist, and Web Trends that provide Web traffic analysis mainly for the purpose of gathering business intelligence. Accrue, Net Genesis, and Aria axe designed to analyze ecommerce events such as products bought and advertisement click-through rates in addition to straight forward usage statistics. Accrue provides apathy analysis visualization tool and IBM's Surf Aid provides LAP through a data cube and clustering of users in addition to page view statistics.

Padmanabhan *et al.* (1998) uses Web server logs to generate beliefs about the access patterns of Web pages at a given Web site. Algorithms for finding interesting rules based on the unexpectedness of the rule were also developed.

2.1.4. Site Modification

The attractiveness of a Web site, in terms of both content and structure, is crucial for many applications, e.g. a product catalog for e-commerce. While the results of any of the projects could lead to redesigning the structure and content of a site, the adaptive Web site project (SCML algorithm) (Perkowitz et al., 1998) focuses on automatically changing the structure of a site based on usage patterns discovered from server logs. Clustering of pages issued to determine which pages should be directly linked.

2.1.5. Usage Characteristics

While most projects works' on characterizing the usage, content, and structure of the Web, there is large amount of overlap between Web characterization research and Web Usage mining. Catledge *et al.* (1995) discusses the results of a study conducted at the Georgia Institute of Technology, in which the Web browser Mosaic was modified to log client side activity. The results collected provide detailed information about the user's interaction with the browser interface as well as the navigational strategy used to browse a particular site. The project also provides detailed statistics about occurrence of the various client side events such as the clicking back/forward buttons, saving a file, adding to bookmarks etc.

Huberman *et al.* (1998) proposed a model which can be used to predict the probability distribution of various pages a user might visit on a given site. This model works by assigning a value to all the pages on a site based on various attributes of that page. The formulas and threshold values used in the

model are derived from an extensive empirical study carried out on various browsing communities and their browsing patterns

Arlitt *et al.* (1997) discussed various performance metrics for Web servers along with details about the relationship between each of these metrics for different workloads. Manley (1997) developed a technique for generating a custom made benchmark for a given site based on its current workload. This benchmark, which he calls a self-configuring benchmark, can be used to perform scalability and load balancing studies on a Web server. Chi *et al.* (1998) describe a system called WEEV (Web Ecology and Evolution Visualization) which is a visualization tool to study the evolving relationship of web usage, content landsite topology with respect to time.

2.2. USER NAVIGATION PATTERN MINING

Pattern discovery from web data is the key component of web mining and it converges algorithms and techniques from several research areas. Catledge et al., (1995) say that the Web is a kind of open, highly dynamic and collaborative hypermedia system, a "dynamic information ecology" including two main types of user strategies: search and navigation. Cove et al., (1988) add a third strategy, "serendipitous browsing", when the users randomly walks through Web pages. Web designers must be aware of these strategies when planning a Web site, since there are different needs associated to each one. There's always the risk of users becoming "lost in cyberspace" (Nielsen, 1990), when these needs are insufficiently mitigated. Abundant information can be uncovered, if they are properly analyzed (Kimble et al., 2007). Recently, several Web Usage Mining (WUM) systems have been proposed to predicting user's behaviour, preferences and their navigation behaviors. Techniques that have been successfully exploited in pattern discovery fall under statistical pattern mining, association rules mining, clustering and classification. This section reviews statistical pattern mining and association rules, while Section

2.3 discusses classification of web data and Section 2.4 presents clustering techniques used in Web Usage Mining.

2.2.1. Statistical Pattern Mining

Statistical techniques are the most powerful tools in extracting knowledge about the visitors of a Web site. The analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file. By analyzing the statistical information contained in the periodic Web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitate the site modification task, and providing support for marketing decisions.

Statistical models for pattern matching and knowledge discovery is a technique that has been done by many researchers. Borges et al., (1998 and Levene et al., (1999) used statistical models to represent user navigation. In their works, WWW is considered to be a database of pages, described as a directed graph whose nodes are pages and the arcs are hyperlinks between pages. Through association of states to pages and probabilities to links, one can build Markov chain models to represent the navigation process, since this process has a strong regularity from a statistical point of view. A survey of various techniques on machine learning and statistical pattern mining for analyzing hypertext is provided by Chakrabarti *et al.* (1998).

Spiliopoulou *et al.* (1999) proposed the exploitation of mining technology to discover access patterns with "interesting" statistical properties and presented Web Utilization Miner (WUM) – a tool designed for the purpose. Mobasher *et al.* (1998) proposed a framework for web mining using various web mining task and implemented a prototype namely WEBMINER. Recently, AlMurtadha *et al.* (2010) used statistical techniques for mining web navigation profiles for recommendation system.

2.2.2. Association Rules

In the Web domain, the pages, which are most often referred together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions. In the term Web usage mining, the association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. The support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. When loading a page from a remote site, association rules can be used as a trigger for pre-fetching documents to reduce user perceived latency.

Mary et al., (2010) used association rules to improve the prediction accuracy and claim that the method proposed is better than Streaming Association Rule (SAR) model. They enhanced the existing SAR mining model with Apriori-like algorithm and Dynamic programming approach. An enhanced pruning rule method for eliminating the redundancy was also introduced in the preprocessing phase. This pruning of rules leads to better prediction accuracy.

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 pre-fetching documents in order to reduce user-perceived latency when loading a page from a remote site.

The problem of deriving Association Rules from data was first formulated in (Agrawal et al., 1993) and is called the "market-basket problem". The problem is that we are given a set of items and a large collection of transactions which are sets (baskets) of items. The task is to find relationships between the containments of various items within those baskets. Mannila *et al.* (1994) use page accesses from a Web server log as events for discovering frequent episodes. The major data mining technique used in their research work was association rules. Chen *et al.*, (1996) introduce the concept of using the maximal forward references in order to break down user sessions into transactions for the mining of traversal patterns.

Batista et al., (2001) perform mining process for online newspaper Web access logs by using Apriori algorithm. Apriori was the first scalable algorithm designed for association-rule mining algorithm. Apriori is an improvement over the AIS and SETM algorithms (Agrawal et al., 1994). The Apriori algorithm searches large item-sets during its initial database pass and uses its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent item-sets and those below are called small item-sets (Chen *et al.*, 1996).

Zhou *et al.* (2006) used association knowledge to discover knowledge from web logs and recommended their system for online applications such as web recommendation and personalization. Their experiments showed that the rules generated are comparable in quality.

Yu *et al.* (2001) proposed a novel incremental mining algorithm using FP-growth algorithm for identifying frequent patterns of web usage mining. Their results were comparable with the existing standards and had the potential to a web prototype of Web Log Analyzer in web usage mining.

Wang *et al.* (2004) proposed a method that can discover users' frequent access patterns underlying users' browsing Web behaviors using association rules. They proposed a technique which used a revised algorithm (FAP-Mining) based on the FP-tree algorithm to mine frequent access patterns. The algorithm is accurate and scalable for mining frequent access patterns with different lengths.

2.3. CLASSIFICATION AND CLUSTERING TECHNIQUES

Classification is the technique to map a data item into one of several predefined classes. In the Web domain, Web master or marketer use this technique to establish 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. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines, etc. **Clustering analysis** is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies. Clustering of users will help to discover the group of users, who have similar navigation pattern. This section presents a literature review regarding both these fields.

Liu et al., (2007) proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users' future requests. The approach is based on the combined mining of Web server logs and the contents of the retrieved web pages. They used character N-grams to represent the contents of web pages, and combined them with user navigation patterns by building user navigation profiles composed of a collection of N-grams

Later, Cadez *et al.* (2000) present a tool called WebCANVAS that displays clusters of users with similar navigation behavior. Prasetyo *et al.*

(2002) introduce "Naviz", a interactive web log visualization tool that is designed to display the user browsing pattern on the web site at a global level and then display each browsing path on the pattern displayed earlier in an incremental manner. The support of each traversal is represented by the thickness of the edge between the pages. Such a tool is very useful in analyzing user behavior and improving web sites.

Baraglia et al., (2002) proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. The main goal of SUGGEST is to find useful information from the user access data collected in web server logs. SUGGEST adopts a two levels architecture composed by an offline creation of historical knowledge and an online engine that understands user's behavior. After a pre-processing of the data recorded in the web server log files, SUGGEST creates clusters of related pages based on users past activity, and then classifies new users by comparing pages in their active sessions with pages inside the clusters created. A set of suggestions is then obtained for each request. The main disadvantages of this system are: Online component and offline component work separately, how to maintain and update the knowledge extracted in the offline phase and how the system can exactly understand the differences between index page and content page.

In the new architecture of SUGGEST they put together the previous two components into a single online module performing the same operation (Baraglia et al., 2004). As the requests arrive at this system it incrementally updates a graph representation of the Web site based on the active user sessions and classifies the active session using a graph partitioning algorithm. This architecture was designed to be usable on Web sites made up of pages statically generated, i.e. Web sites with a fixed number of pages. A list containing all the information describing a Web site pages was required as input by this architecture at its start-up time.

The last contribution of SUGGEST architecture proposed by Baraglia et al., (2002), introduces' a novel solution to implement WP (Web Personalization) as a single online module that performs user profiling, model updating, and recommendation building. It is designed to dynamically generate personalized contents of potential interest for users of large Web sites made up of pages dynamically generated. It is based on an incremental personalization procedure tightly coupled with the Web server. It is able to update incrementally and automatically the knowledge base obtained from historical usage data and to dynamically generate a list of page links (suggestions). The suggestions are used to personalize the HTML page requested on-the-fly. The adoption of a LRU-based (Least Recently Used) algorithm handling the knowledge base makes it possible for SUGGEST to manage large Web sites. But in this system quality of recommendations is not better than previous version of this system.

Another study towards Web Usage Mining proposes to cluster visitors of a website based on the page requests taking place on the sessions belonging to them. The aim of this study presented in Farzan (2004) is to discover the groups of pages that are visited together by many visitors. This information can be used by the Web master in redesigning the Web Site or updating it with extra links between these pages

Park *et al.* (2008) proposed a general sequence based clustering method in association with Markov models for user, web page clustering. A new, fuzzy ART-enhanced K-means algorithm is also developed and its superior performance is demonstrated.

Mobasher *et al.* (2000) and Nakagawa et al., (2003) presents a WebPersonalizer system which provides dynamic recommendations, as a list of hypertext links, to users. The analysis is based on anonymous usage data combined with the structure formed by the hyperlinks of the site. Data mining techniques (i.e. clustering, association rules and sequential pattern discovery)

are used in the preprocessing phase in order to obtain aggregate usage profiles. In this phase, Web server logs are converted in to clusters made up of sequences of visited pages, and cluster made up of set of pages with common usage characteristics. The online phase considers the active user session in order to find matches among the user's activities and the discovered usage profiles. Matching entries are then used to compute a set of recommendations which will be inserted into the last requested page as a list of hypertext links. WebPersonalizer is a good example of two-tier architecture for Personalization systems.

In Mobasher *et al.* (2000), the authors have developed a recommendation system, termed Yoda that is designed to support large-scale Web-based applications requiring highly accurate recommendations in real-time. With Yoda, they introduced a hybrid approach that combines collaborative filtering (CF) and content-based querying to achieve higher accuracy. Yoda is structured as a tunable model that is trained online and employed for real-time recommendations. The on-line process benefits from an optimized aggregation function with low complexity that allows real time weighted aggregation of the soft classification of active users to predefined recommendation sets.

Jespersen *et al.* (2002) proposed a hybrid approach for analyzing the visitor click sequences. A combination of hypertext probabilistic grammar and click fact table approach is used to mine web logs which could be also used for general sequence mining tasks.

Analog (Yan *et al.*, 1996) is one of the first WUM systems. It is structured according to an off-line and an online component. The off-line component builds session clusters by analyzing past users activity recorded in server log files. Then the online component builds active user sessions which are then classified according to the generated model.

2.4. GRAPH PARTITION CLUSTERING

A partitioning method was one of the earliest clustering method to be used in Web usage mining by Yan *et al.* (1996). They used an incremental algorithm that produces high quality clusters. Each user session is represented by an n-dimensional feature vector, where n is the number of Web pages in the session. The value of each feature is a weight, measuring the degree of interest of the user in the particular Web page. The calculation of this figure is based on a number of parameters, such as the number of times the page has been accessed and the amount of time the user spent on the page. Based on these vectors, clusters of similar sessions are produced and characterized by the Web pages with the highest associated weights. The characterized sessions are the patterns discovered by the algorithm. One problem with this approach is the calculation of the feature weights. The choice of the right parameter mix for the calculation of these weights is not straightforward and depends on the modeling abilities of a human expert.

A partitioning graph theoretic approach is presented by Perkowitz et al., (2000) have developed a system that helps in making Web sites adaptive, i.e., automatically improving their organization and presentation by mining usage logs. The core element of this system is a new clustering method, called cluster mining, which is implemented in the Page-Gather algorithm. Page-Gather receives user sessions as input, represented as sets of pages that have been visited. Using these data, the algorithm creates a graph, assigning pages to nodes. An edge is added between two nodes if the corresponding pages co-occur in more than a certain number of sessions. Clusters are defined either in terms of cliques, or connected components. Clusters defined as cliques prove to be more coherent, while connected component clusters are larger, but faster to compute and easier to find. A new index page is created from each cluster with hyperlinks to all the pages in the cluster. The main advantage of Page-Gather is that it creates overlapping clusters. Another partitioning clustering method is employed by Cadez *et al.* (2000) in the Web CANVAS tool, which visualizes

user navigation paths in each cluster. In this system, user sessions are represented using categories of general topics for Web pages. A number of predefined categories are used as a bias and URLs from the Web server log files are assigned to them, constructing the user sessions.

The Expectation-Maximization (EM) algorithm, based on mixtures of Markov chains is used for clustering user sessions. Each Markov chain represents the behavior of a particular subgroup. EM is a memory efficient and easy to implement algorithm, with a profound probabilistic background. The EM algorithm is also employed by Anderson *et al.* (2001) in two clustering scenarios, for the construction of predictive Web usage models. In the first scenario, user navigation paths are considered members of one or more clusters and the EM algorithm is used to calculate the model parameters for each cluster. The probability of visiting a certain page is estimated by calculating its conditional probability for each cluster. The resulting mixture model is named Naive Bayes mixture model since it is based on the assumption that pages in a navigation path are independent given the cluster. The second scenario uses a similar approach to Cadez *et al.* (2000). Markov chains that represent the navigation paths of users are clustered using the EM algorithm, in order to predict subsequent pages.

Baraglia et al., (2004, 2007) proposed a WUM system called SUGGEST, that provide useful information to make easier to the web user navigation and to optimize the web server performance. SUGGEST adopts a two levels architecture composed by an offline creation of historical knowledge and an online engine that understands user's behavior. As the requests arrive at this system it incrementally updates a graph representation of the Web site based on the active user sessions and classifies the active session using a graph partitioning algorithm.

Jalali *et al.* (2008a and 2008b) proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph

based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. This approach improved the quality of user navigation pattern discovery and can be used to predict user's next request in the huge Web sites.

Dixit et al., (2010) presented another user navigation pattern mining system based on the graph partitioning. An undirected graph based on connectivity between Referrer and URI pages was presented along with a preprocessing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph. From the results reported, the potential limitations of this architecture are:

a) The memory required to store Web server pages is quadratic in the number of pages. This might be a severe limitation in large sites made up of millions of pages

b) It does not permit administrators to manage Web sites made up of pages dynamically generated.

All of these works attempt to find architecture and algorithm to improve quality of clustering, but unfortunately the quality still does not meet satisfaction and hence research in this field is still ripe. In the present research an Ant-based clustering technique is used to solve the problem of graph partitioning clustering. The concept of Ant-base clustering and some of the reported literature is discussed in the next section.

2.5. ANT-BASED CLUSTERING

Ant-based clustering and sorting has been first introduced by Deneubourg *et al.* (1990), it explains' different types of naturally-occurring emergent phenomena. It is an instance of the broad category of ant algorithms, (i.e.,) algorithms with some behaviors' observed in real ants. In the case of ant-based clustering and sorting, two related types of natural ant behavior are modeled. When clustering, ants gather items to form heaps and when sorting,

ants discriminate between different kinds of items and spatially arrange them according to their properties (Handl *et al.*, 2004).

Applying the agent technology has improved the performance of web mining compared to traditional approach such as database approach. Han *et al.* (1999) proposed Web Agent for Document Categorization and Exploration on World Wide Web to automatically categorized a set of documents using Web Agent combine with a process for generating new queries used to search related documents and filters the result and extract the set of documents most closely related to the starting set, while Cooley *et al.* (1999) has developed Web Site Information Filter (webSIFT) to use the content and structure information from a Web site in order to identify potentially interesting results from mining usage data. These webSIFT is based on WEBMINER prototype.

With the growth of popularity for E-Commerce, Lee et al., (2001) proposed an intelligent multi-agent based environment known as intelligent Java Development Environment (iJADE) to provide an integrated and intelligent agent based platform in the e-commerce environment on Internet shopping. Mohammadian (2001) provides building blocks for integrating intelligent agent with current search engines. Intelligent agent can improve the performance of search and retrieval engines. The use of existing search and retrieval engines with the addition of an agent will allow a more comprehensive search with a performance that can be measured. Wahab *et al.* (2007), utilizing an ontology called COMRIS to provide shared understanding of a domain between Web mining agent and other agent accessing the gathered information.

Chau *et al.* (2003) developed Collaborative Spider is a multi-agent system to provide post-retrieval analysis and enable across-user collaboration in Web Search and mining. While Nasraoui et al., (2003) presented an Intelligent Web personalization based on web usage mining to discover useful knowledge about user access patterns followed by recommendation engine to response to these knowledge based on users' individual interest.

2.6. LOG FILE ANALYSIS FOR WEB USAGE MINING

Mining the user click-stream for user behavior, and using it to adapt the 'look-and-feel' of a site to a reader's needs was first proposed by Perkowitz et al., (1998). Analysis of Web server transaction logs provides comprehensive information about Web server traffic (Toolan *et al.*, 2003). Knowledge of server traffic can provide information about who is accessing a current Web site, and when and where they are visiting. This type of data can be beneficial in assessing what pages on the site receive the most frequent traffic and who is using them. This can help the design team to further identify target user groups for their current site or for redesigns (Blackett *et al.*, 2003).

Log analysis is typically conducted as an automated procedure with log analyzer software. During log analysis all Web server activity is recorded. This includes data such as the IP address and/or domain of the individual requesting a Web page from the server, the date and time of the request, the filename of the page accessed, and the number of bytes of data served (Peng *et al.*, 2005). This section reviews some work done in this area.

A number of articles have discussed Web server log analysis for libraries, since libraries began to develop Web presences (Li, 1991; Stabin et al., 1997; Nicholas *et al.*, 2000). These articles describe summary level metrics of Website usage, such as the total number of user sessions, broken down by variables such as date, time, or host domain of the requestor. As noted by many of these authors and by Goldberg (2003), these studies have been constrained by two main factors. First, data provided by the hypertext transfer protocol (HTTP) that governs user transactions on the Web is very limited. Second, usage logs are designed for use by system administrators, not for tracking users. While the information available in logs is limited, some user and resource usage data can be gleaned from them. For example, the Internet protocol (IP) or network address of users can provide some insight into who is

using a site, and the requested file can be used to make some conclusions about what content is being used.

The literature on data mining contains descriptions of sophisticated statistical analyses of Website usage, with applications including personalization and system improvement. These analyses apply a variety of techniques including ordinary least squares (OLS) and logistic regression, cluster analysis, decision trees, and neural networks. Web usage mining often analyzes sequences of page accesses to provide personalization and targeted marketing (Dunham, 2003; Han et al., 2001). Feng et al., (2000) provide an example of using Web usage mining techniques to develop a personalization system. Davis (2004) analyzes the information-seeking behavior of chemists based on Web log analysis.

A semantic session analysis model partitioning Web usage logs is presented by (Zhou *et al.*, 2006). The model enhances usage logs with semantic using Markov chain model based on ontology semantic measurement. The competitive method is applied to determine the end of the sessions. Compared with other algorithms, more successful sessions are additionally detected by semantic outlier analysis. Tanasa et al., (2004) in their work preprocessed web log files to reduce their size. They also used data summarization techniques to increase the quality of data obtained after classical preprocessing.

2.7. CHAPTER SUMMARY

The past nine 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. The various works proposed in this area with particular emphasize on web usage mining, clustering and classification was provided in this chapter. In the present work, the application of clustering to extract user navigation behaviour pattern is probed and the methods and techniques used are explained in the next chapter, Methodology.