CHAPTER – II
REVIEW OF LITERATURE

This chapter provides a relevant literature review that contains a wide range of topics, including knowledge discovery, rough set theory, association rule mining, web mining, and web personalization to give the readers the necessary background knowledge and the conceptual framework of the development of the proposed work.

The online business retail spending has grown from $20.3 billion to $197 billion by 2011. According to the EIAA Mediascope Europe 2008 study\(^1\), Europeans are deepening their experience of the internet by not only increasingly using it for leisure, but actively enhance and manage their daily lifestyle. 179 million Europeans (60\%) are online each week. Over half (55\%) of European users are online every single day. Three quarters (75\%) of internet users are online during their evenings compared to 67\% in 2007. 51\% of Europeans use the internet at the weekend, an increase of 13\% since 2007. According to the figures 54\% have booked more holidays than in previous years. Because of the growing confidence, consumers made a huge number of purchases online in 2008 or 9.2 purchases per person versus 7.7 in 2007. By having these figures, it can be concluded that market will successfully withstood only by companies taking significant attention into their web data and making serious analysis, others will be choked off\(^2\).

Therefore, with no doubt, web log data is the largest source of information concerning human interaction with the www and continue to grow rapidly. This enables knowledge discovery from web logs to recognise practical users, improve marketing strategies and increase web site’s retention etc. Currently, most web sites are designed without taking into account how web logs can be used to tune and evaluate the usefulness of the
web site. The success of the web site cannot be measured only by hits and page views. Usually, the need to know how users behave comes later. Ideally, web site design should enclose techniques, which relate web pages and access activity into standard database format and make data preparation process easier. Unfortunately, web site designers and web log analysers do not usually cooperate. This causes problems such as identification, unique user’s construction, discrete users sessions and collection of essential web pages for analysis. The result of this is that many web log mining tools have been developed and widely exploited to solve these problems.

However, neither commercial nor free tools solve adequately these problems. Several steps called knowledge discovery must be passed through in order to observe patterns from data. These steps are (a) data preprocessing which includes such stages as data cleaning, feature selection, transformation, (b) data analysis and (c) finally results visualisation and examination.

Data mining is a very important part of modern business. Mining large-scale transactional databases is considered to be a very important research subject for its obvious commercial potential. It is also a major challenge due to its complexity. New techniques have been developed over the last decade to solve classification, prediction and recommendation, problems in these kinds of databases. To find and evaluate valuable and meaningful patterns requires huge quantities of data. Each business can produce amounts of data just by recording all customers’ events into the database and it is called data warehousing. Although large quantities of data are typically generated by dotcom web servers by monitoring requested packages going to web site visitors. This information is logged into special purpose – web log files.
A lot of studies have been conducted in Web usage mining. Some focus on the mining of association rules and navigation patterns in the user access paths\textsuperscript{20, 21}. A session is viewed as a transaction in association rule mining and algorithms for association rule mining are employed to find frequent paths that are followed by many users. Others build data cubes from Web server logs for OLAP and data mining \textsuperscript{21, 23}. The statistics along pages, IP domains, geographical location of users, and access time are calculated from sessions. Some others cluster users based on their access patterns\textsuperscript{24}. There is also research on data preparation and query language for Web usage mining, and Web personalization based on Web usage mining.

Nina et al.,\textsuperscript{25} suggests a complete idea for the pattern discovery of Web usage mining. Web site creators must have clear knowledge of user's profile and site intentions and also emphasized information of the approach users will browse Web site. The creators can examine the visitor's behavior by means of Web analysis and identify patterns of the visitor's activities. This Web analysis includes the transformation and interpretation of the Web log records to identify the hidden data or predictive pattern by the data mining and knowledge discovery process. This result provides a great view coupled with the Web warehousing.

Wu et al.,\textsuperscript{26} have given a Web Usage Mining technique based on the sequences of clicking patterns in a grid computing environment. Examining user's browsing pattern is a significant process of web usage mining. It can assist the web supervisors or creators enhance the web structure or increase the performance of the web servers. Mining on the Sequences of such Clicking Patterns (MSCP) can be regarded as a data mining task. MSCP is generally an expensive procedure because of its significant quantity of time for computation and storage for archiving a large quantity of information.
Running MSCP becomes ineffective or even not practical on a computer with restricted resources. The author discovered the usage of MSCP in a distributed grid computing surroundings and expresses its effectiveness by empirical cases.

To analyze the usage of the Web, Web mining, especially Web usage mining, has been proposed by many researchers. Web usage mining is the mining of Web usage data. In most Web usage mining studies, Web server logs are used as the primary data source, although client and proxy level logs may be used. A Web server log collects a large amount of information about user activities on the Web site by keeping information about the requests of pages on the server.

Personalized Web page recommendation is strictly restricted by the nature of web logs, the intrinsic complexity of the problem and the higher efficiency needs. When handled by existing Web usage mining methods, because of the existence of a large number of meaningful clusters and profiles for visitors of a usually highly rated Website, the model-based or distance-based techniques are likely to create very strong and simple assumptions or, on the other hand, to turn out to be highly complex and slow. The author designed a heuristic majority intelligence technique, which effortlessly adjusts to changing navigational patterns; with the low cost explicitly individuate them ahead of navigation.

Zdzisław Pawlak introduced Rough set theory to deal with uncertainty and vagueness. Rough set theory became popular among scientists around the world due to its fundamental importance in the field of artificial intelligence and cognitive sciences. The building block of rough set theory is an assumption that every set of the universe of discourse is associated with some information in the form of data and knowledge.
Hu et al.,\textsuperscript{33} presented the formal definitions of rough set theory. A. Kusiak described the basic concepts of rough set theory, and other aspects of Data mining. The other aspects of data mining are Equivalence classes, Atoms, Approximation accuracy, Boundary approximation, Classification accuracy, Classification quality, Sensitivity, Specificity, Positive predicted value, Negative predicted value, Rule length, Rule strength, Exact rule, Approximate rule, Rule support, Rule coverage, Rule acceptance and Discrimination level.

Zhong et al.\textsuperscript{29} applied Rough Sets with Heuristics(RSH) and Rough Sets with Boolean Reasoning(RSBR) for attribute selection and discretization of real-valued attributes. Komorowski et al.\textsuperscript{30} studied an application of rough sets to modeling prognostic power of cardiac tests. Carlin et al.\textsuperscript{31} presented an application of rough sets to diagnosing suspected acute appendicitis.

A well known rule pruning approach has been developed for improving rule-based models and studied in rough-set literature. Implementations are available both in the ROSETTA\textsuperscript{34} and the RSES\textsuperscript{35} systems. RSES system has both filtering and generalization methods available. A comprehensive study on rule filtering is given in\textsuperscript{36}, and applications of this method are described. A method performing rule generalization based on clustering of rules is described in\textsuperscript{37}. Dynamic reducts\textsuperscript{38}, or any other kind of approximate reducts were considered as a method for rule pruning as it resulted in general rules and prevented the incorporation of noise-derived information in the induced model.
Mining association rules in large databases was first approached by Agrawal, Imielinski and Swami\textsuperscript{39}. A database of transactions is the training data from which rules will be generated. The aim of association rule discovery is to find rules with strong association between items form the training data.

To mine association rules from a large database, minimum constraints are defined\textsuperscript{40}. The number of rules to be generated can get very large, e.g. for 1000 rules there are $2^{1000}$ possibilities. The rules generated will be based upon a minimum support constraint. The data space can be very large, therefore the minimum support constraint is needed to prune the data space to generate rules.

The support of a rule is the frequency with which the rule occurs in the training data. For example 25 transactions in out of 100 transactions, assuming the training data includes 100 transactions, contain Pepsi. Then the support of Pepsi is 0.25. The itemsets within the rules generated are considered to be frequent itemsets; Pepsi is frequent in the previous example. Further constraints such as Confidence and Lift can be applied to the rules that are generated. Applying other constraints is defined as interest.

Some constraints that can be applied as a measure of interestingness are:

1) Confidence (A $\Rightarrow$ B) = support (A $\Rightarrow$ B) / support (A)

2) Lift (A $\Rightarrow$ B) = support (A $\Rightarrow$ B) / support (A) x support (B)

3) Leverage (A $\Rightarrow$ B) = support (A $\Rightarrow$ B) – support (A) x support (B)

Confidence is a common measure of interest for generating association rules in Association rule discovery.
Discovering association rules can be thought of as a three part process:

1) Search the data space to find all the items or itemsets whose support is greater than the user specified minimum support. These itemsets are the frequent itemsets.

2) Generate interesting rules based on the frequent itemsets. A rule is said to be interesting if its confidence is above a user’s specified minimum confidence.

3) Remove (prune) all rules which are not interesting from the rule space.

Further constraints can be applied to generate rules specific to a user’s needs. However, this can only be done when the rules are already available from the first two parts of the rule generation process.

In association rule discovery, the frequent itemset generation process is the part which needs improvement in efficiency. When the training data has thousands of transactions finding frequent itemsets can take up a lot of time. Therefore, most research related to association rule discovery has been conducted to improve the frequent itemsets generation process. Specifically most of the attention has been given to improve part one of the three part process. There are several algorithms which have been developed to mine association rules quickly.

The Apriori algorithm has proved to be an efficient algorithm in mining association rules. It has been used effectively and has become the standard algorithm used for association rule discovery.

Apriori follows a two step process to generate rules. The first step is to find all frequent itemsets. From these frequent itemsets association rules are generated which is the second step. The first step will limit the number of
itemsets which are considered for the antecedent and consequent of the rules and information about the itemset frequency is maintained. An Itemset is frequent if it satisfies the minimum support constraint.

Some variants of the Apriori approach have showed that very few passes through the training data may be necessary to generate association rules \(^{41}\).

There have been several algorithms proposed to solve the task of generating frequent itemsets efficiently \(^{42, 43}\). Webb \(^{44}\) argues that for some applications a direct search may be more effective than the two part process of the Apriori algorithm. The algorithm presented maintains the data in memory from which the association rules can be generated through constraints.

The need for interestingness measures originates from limitations of the support and confidence approach. Even though there is a reasonable concept behind the support and confidence approach there are still some cons in using this approach for rule extraction. Brin et al. \(^{45}\) and Aggarwal et al. \(^{46}\) address the weaknesses of the support-confidence framework.

One of the analyzing approaches was proposed\(^{47}\) as a subjective approach that assists the user in finding interesting rules by analyzing the discovered association rules using the user’s existing knowledge about the domain.

Liu et al. \(^{48}\) presented a ranking method for mined patterns according to the user’s existing knowledge and general impressions. The main disadvantage is that the user is required to express his/her knowledge in the specifications, which might not be an easy and standard task.
In other research Tuzhilin and Silberschatz\textsuperscript{49} discussed subjective measures depending on two concepts’ actionability and unexpectedness and the relation between them. Actionability states that the pattern is interesting if the user can act on it to his advantage. Unexpectedness focuses on the surprising factor for the pattern. Also, it relates to beliefs which can be defined as logical statements. There are two types of beliefs: hard, which the user is not willing to change; and soft, which the user can change if suggested by newly discovered patterns.

Piatetsky-Shapiro\textsuperscript{50} considered as first proposal using statistical independence of rules as an interestingness measure. More methods have since been proposed using different statistical approaches. Brin, \textit{et al.}\textsuperscript{51} proposed lift and $\chi^2$ (chi-squared) as correlation measures and developed an efficient mining method. Hilderman and Hamilton and Tan, \textit{et al.} have comparative studies of different interestingness measures and address the concept of null-transactions. Because the probability of an item appearing in a particular transaction is usually very low, it is desirable that a correlation measure should not be influenced by these transactions which they call “null-transactions”, i.e., the transactions that do not contain any of the items in the rule being examined. In another study related to the correlation, Omiecinski and Lee, \textit{et al.} found that all\textunderscore confidence, coherence, and cosine are null-invariant and are thus good measures for mining correlation rules in transaction databases.

After all these studies, Tan \textit{et al.}\textsuperscript{52} discussed the properties of twenty-one objective interestingness measures and analyzes the impacts of Support based pruning and contingency table standardization. This study ended with the conclusion that there was no measure that was consistently better than others in all application domains. However, some of these measures were correlated with each other.
Three measures for capturing relatedness between item pairs were proposed by Shekar and Natarajan\textsuperscript{53}. These measures used the concept of function embedding to appropriately weigh the relatedness contributions due to Mutual Interaction, complementary and substitutability between items. At the end they proposed interestingness coefficient by combining the three relatedness measures. All the three measures were calculated based on the probability without taking into account the transaction itself (large or small).

Berberidis, \textit{et al.}\textsuperscript{54} and George, \textit{et al.}\textsuperscript{55} introduced a data mining paradigm, involving the discovery of contiguous frequent itemsets and presented level-wise algorithm for finding these item sets, which led them to generate a two level global support (gsup) the first-level support and local support (lsup), the second-level support. For the evaluation, they introduced new metric (Mutual Exclusion Metric) to evaluate the degree of the mutual exclusion between two items.

Another measure based on information theory designed a rule interestingness measure named Directed Information Ratio. This measure filters out these rules whose antecedent and consequent are negatively correlated and these that have more counter examples than examples.

A detailed survey,\textsuperscript{56} reviews the interestingness measures for rules and summaries, classifies them from several perspectives and compares their properties. They present thirty-eight probability based objective interestingness measures for association rules Table 2-1. Another paper by Lenca, \textit{et al.} studied twenty interestingness measures by using 10 data sets. This study is compared to an analysis of formal properties of the measures which make a best choice of user’s needs.
There are some research papers where they tried to improve the measures quality by using some existing information. One of these papers Hilderman, et al. proposed a concept of share-confidence and support which involves the quantity and price of the items in the confident and support computation.

Moreover, in specific application domains some attributes can have very different degrees of interestingness for the user, depending on which attributes occur in the rule antecedent. Thus, in some applications, different attributes might have very different “costs” to be accessed. The typical example is medical diagnosis. For example, some health-related attributes can only be determined by performing a very costly examination. Suppose that the antecedent (“if part”) of a discovered rule $r_1$ involves the result of an exam $e_1$ which costs $200, while the antecedent of a discovered rule $r_2$ involves instead the result of another exam $e_2$ which costs $20. All other things including prediction accuracy being equal, one would rather use rule $r_2$ for diagnosis. So the cost of the attribute becomes part of interestingness decision. There are some data mining algorithms which take into account attribute costs like what had been described in Ming (1993) and Turney (1995).

Han et al. (2000) proposed the FP-growth algorithm for mining the complete set of frequent patterns. This algorithm is based on a new frequent pattern tree (FP-tree) structure, which is a prefix tree for storing necessary information about frequent patterns. Only frequent 1-item(s) are stored in each node of the tree. The FP-tree is applied to restrict generation of a large number of candidate sets. This concept eliminates the multi-scan inefficiency the Apriori algorithm. FP-growth is adapted to the pattern growth approach to avoid scanning the database for every level of frequency and handles a large number of candidate sets. The algorithm begins with frequent 1-items
which are kept in the FP-tree to perform recursive mining. The search technique is a partitioning-based divide-and-conquer method to increase the running time efficiency.

Web servers around the world generate thousands of giga-bytes of such data every day. According to the Internet Software Consortium, the World Wide Web (WWW) since 1994 has grown from two million servers to more than 110 million in 2001 (Internet_Systems_Consortium). The number of home users has increased from 3 million to more than 89 million for the same period in the US, estimated by another Internet research company - Jupiter MM. It also states that almost 33 million Europeans in December 2001 used the Internet to make their Christmas shopping. Forrester analysts in (Schmitt et al. 1999)\textsuperscript{59} reported that 84\% of interviewed companies received demand for site data to skyrocket by 2001.

There are several researches done in the area of web structure mining. Social Network Analysis is one of the most famous ones [Kautz 97, Chakrabarti 00]\textsuperscript{60,61}. With social network analysis, it is possible to discover specific types of pages such as hubs and authorities. The separation is made with respect to the number of incoming and outgoing links. One of the works, which is related to social networks analysis, aims to model network of AI researches. Another research in this area is modeling Web topology such as HITS [Kleinberg 98]\textsuperscript{62} and improved version of HITS by using content information with link structure. The methods addressed in these works are mainly used to calculate quality and relevance relation of two web pages. Another application of web structure mining is discussed in the context of web Warehouse. Measurement of frequency of local links that is on the same web server helps measure completeness of the web site. Also measuring replication of documents across web warehouse contributes to find out the mirrored sites.
Web structure mining has another dimension, which is the discovering the structure of Web document. This dimension of web structure mining is used to extract the schema of Web pages. This information is beneficial for navigation purpose and provides comparison and integration of Web page schemes. This type of application of web structure mining provides information to database systems for accessing to web pages by providing a reference schema [Madria 99].

S. Madria et al. [Madria 1999] gave details about how to discover interesting facts describing the connectivity in the Web subset, based on the given collection of connected web documents. The structure information obtained from the Web structure mining has the followings:

Zhu et al., recently developed a user-side web personalization system “Web-IC” to predict information content (IC) pages that a web user will be interested to visit. The motivation of this system is to help web users locate these IC pages everywhere on the web for the users themselves based on their own behaviours. The words contained in the web pages a user visits, as well as the actions such as back pages browsing or follow-up pages the user makes on such pages are taken into consideration as users’ interests for behavior modeling. It is shown that classifiers built from such features as extracted from user-side browsing properties can effectively predict the interested web pages for the users.

Other researchers studying user-centric personalization include Lieberman, who developed the Letizia web search agent for web page recommendation, Billsus and Pazzani, who query users to get feedback for recommending news web pages and Ardissono et al. who customize the presentation of a website advertising a product to a user, based on a
monitoring of the user’s interests. These researchers are more focused on user modeling and machine learning, whereas we are most interested in developing effective data mining techniques.

The personalization of services offered by a Web site is an important step in the direction of alleviating information overload, making the Web a friendlier environment for its individual user and hence creating trustworthy relationships between the Web site and the visitor-customer. In (Mobasher et al., 1999 a) Web Personalization is simply defined as the task of making Web-based information systems adaptive to the needs and interests of individual users. Typically, a personalized Web site recognizes its users, collects information about their preferences and adapts its services, in order to match the users’ needs. Web personalization improves the Web experience of a visitor by presenting the information that the visitor wants to see in the appropriate manner and at the appropriate time. In e-business, Web personalization additionally provides mechanisms to learn more about customer needs, identify future trends and eventually increase customer loyalty to the provided service.

Hitherto, Web personalization has been related mainly with recommender systems and customized Web sites for newspapers, electronic shops, etc. (Schafer et al., 2001), (Pretschner and Gauch, 1999). However, Web personalization comprises a variety of functions ranging from simple user recognition to more advanced functionality, such as performing certain tasks on behalf of the user. This functionality is offered by the Web personalization systems according to a personalization policy that specifies the manner in which personalization will be delivered to the final user.
In the majority of the existing commercial personalization systems, the personalization process involves substantial manual work and most of the time, significant effort on the part of the user. Despite these problems, the number of personalized Web pages is increasing. A survey by Data monitor (2001) predicts that global investment in personalization technologies will reach $2.1 billion in 2006, where half of the investment will be made by firms in the financial services and retail sectors. Personalization technologies are also popular with telecommunications and entertainment firms. Other surveys reflect the adoption of this new technology by users. According to a poll conducted by Cyber Dialogue (2001) 56% of the participants said that ‘they are more likely to shop at a site that offers personalization’ and 63% were ‘more likely to register for a site that offers personalization.’

One way to expand the personalization of the Web is to automate the adaptation of Web-based services to their users. Machine learning methods have a successful record of applications to similar tasks, i.e., automating the construction and adaptation of information systems (Langley, 1999; Pohl, 1996 and Webb et al., 2001). Furthermore, the incorporation of machine learning in larger process models, such as that of Knowledge Discovery in Data (KDD or Data Mining), can provide a complete solution to the adaptation task. Knowledge Discovery in Data has been used to analyze data collected on the Web and extract useful knowledge. This effort was named Web Mining (Etzioni, 1996) and one branch of it is concerned with the analysis of usage data, i.e., records of how a Web service is used.

Early work in Web usage mining (Srivastava et al., 2000) did not consider extensively its use for personalization. Its primary focus was on the discovery of decision-support knowledge, expressed in terms of descriptive data models to be evaluated and exploited by human experts. However, Web usage mining can also be a useful tool for Web personalization.
All that is required for the application of Web usage mining to Web personalization is a shift of focus from the traditional, decision-support knowledge discovery, i.e., the static modeling of usage data, to the discovery of operational knowledge for personalization, i.e., the dynamic modeling of users. This type of knowledge can be directly delivered back to the users in order to improve their experience in the site, without the intervention of any human expert. Thus, it is now widely recognized that usage mining is a valuable source of ideas and solutions for Web personalization.

A Web personalization system can offer a variety of functions starting from simple user salutation, to more complicated functionality such as personalized content delivery. Kobsa et al. (2001) recommends a classification of the Web personalization functions, which is extended here to a generic classification scheme.

Clustering has been used for grouping users with common browsing behavior, as well as grouping Web pages with similar content (Srivastava et al., 2000). One important constraint imposed by Web usage mining to the choice of clustering method is the fact that clusters should be allowed to overlap (Paliouras et al., 2000b). This is important for Web personalization since a user or a Web page may belong to more than one group.

A partitioning method, was one of the earliest clustering methods to be used in Web usage mining by Yan et al. (1996). In this work, the Leader algorithm (Hartigan, 1975), is used to cluster user sessions. Leader is 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 a 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.

Further more, the use of the Leader algorithm is problematic, as the construction of the clusters depends on the presentation order of the input vectors. For instance, if three training vectors (a, b, c), are submitted in that order to the algorithm, a different set of clusters may result than if the vectors are submitted in a different order, e.g. (b, a ,c ).

A partitioning graph theoretic approach is presented by Perkowitz and Etzioni (1998, 2000), who 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, as signing 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. Furthermore, in contrast to the other clustering
methods, the clusters generated by this method group together characteristic features of the users directly. Thus, each cluster is a behavioral pattern, associating pages in a Web site. However, being a graph based algorithm, it is rather computationally expensive, especially in the case where cliques are computed.

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, (Dempster et al., 1977) 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. However, there are cases where it has a very slow linear convergence and may therefore become computationally expensive, although in the results in Cadez et al. (2000), it is shown empirically that the algorithm scales linearly in all aspects of the problem.

The EM algorithm is also employed by Anderson et al. (2001a) 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 Baye’s 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.

An extension of partitioning clustering methods is fuzzy clustering that allows ambiguity in the data, by ‘distributing’ each object from the data set over the various clusters. Such a fuzzy clustering method is proposed in (Joshi and Joshi, 2000) for grouping user sessions, where each session includes URLs that represent a certain traversal path. The Web site topology is used as a bias in computing the similarity between sessions.

The site is modeled using a tree, where each node corresponds to a URL in the site, while each edge represents a hierarchical relation between URLs. The calculation of the similarity between sessions is based on the relative position in the site tree of the URLs included in the sessions. Clustering is implemented using two newly devised algorithms: Fuzzy c-medoids and Fuzzy c-trimmed-medoids, which are variants of the Fuzzy c clustering method (Bezdek, 1981). Fuzzy clustering is also employed by Nasraoui et al. (1999), who use the Relational Fuzzy Clustering Maximal Density Estimator (RFC-MDE) algorithm to cluster user sessions identified in the Web server logs. The employment of fuzzy clustering methods allows the creation of overlapping clusters, since they introduce a degree of item-membership in each cluster. However, this item-membership is specified by a membership function, the design of which is a non-trivial issue.

A hierarchical clustering approach is employed by Fu et al. (1999) who use the BIRCH algorithm (Zhang et al., 1996) for clustering user sessions. Data from the Web server log are converted into sessions represented by vectors, where each vector contains the ID of each Web page
accessed, together with the time spent on that page. In order to improve the efficiency of the clustering algorithm and to discover more general patterns, sessions are generalized using the page hierarchy of the Web site as a bias. Each Web page in a session is replaced by a more general Web page according to the page hierarchy, using the attribute-oriented induction method (Han et al., 1992). The resulting generalized sessions are used as input to the BIRCH algorithm.

BIRCH is a very efficient and incremental algorithm for processing large volumes of data. It can produce qualitative clusters by scanning the data only once and improve them with additional scans. However, similar to the Leader algorithm, BIRCH also depends on the presentation order of the input data. Furthermore, due to the specification of the algorithm, it does not always create ‘natural’ clusters since each cluster, is allowed a maximum size of members (Halkidi et al., 2001).

A variety of model-based clustering methods have been used in (Paliouras et al., 2000b). A probabilistic method, Autoclass (Hanson et al., 1991), a neural network, Self-Organizing Maps (Kohonen, 1997), a conceptual clustering method, COBWEB (Fisher, 1987), and its non-hierarchical variant, ITERATE (Biswas et al., 1998), are exploited in order to construct user community models, i.e., models for groups of users with similar usage patterns. Community models are derived as characterizations of the clusters and correspond to the interests of user communities.
Summary

In this chapter, the background of web log data retrieval, filtering, machine-learning and web mining has been discussed. Also reviewed is related research work regarding rough set, association rule mining and evaluation of the rule set. The research background sets up the research scope and takes research problems into consideration. During the discussion and analysis of corresponding issues, it discloses the direction and related methods of this thesis research.
Reference


22. Buchner and M. Mulvenna, Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining, SIGMOD Record, 27, 1998.


50. Tuzhilin, A., & Silberschatz, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowledge and Data Eng.*, (pp. 8, 6.).


