INTRODUCTION AND DESIGN OF THE STUDY

1.1 Introduction

The World Wide Web (WWW) is playing an important role in many walks of life. Almost all businesses have to consider web strategies carefully as part of their business plan. Due to the ease of movement from one portal to another web users can be very mobile. If a particular website does not satisfy the needs of the user in a relatively short period of time, the user will quickly move on to another website. Therefore it is very important to understand the needs and characteristics of web users.

During the past few years the World Wide Web has become the biggest and most popular way of communication and information dissemination. It serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content, software and personal logs (blogs). Every day, the web grows by roughly a million electronic pages, adding to the hundreds of million pages already on-line. Because of its rapid and chaotic growth, the resulting network of information lacks of organization and structure. Users often feel disoriented and get lost in that information overload that continues to expand. On the other hand, the e-business sector is rapidly evolving and the need for web marketplaces that anticipate the needs of their customers is more than ever evident. Therefore, the ultimate need nowadays is that of predicting the user needs in order to improve the usability and user retention of a website. Moreover the study presents novel methods and techniques that useful to the present requirement of understanding the web users.
Nowadays Web users are facing the problems of information overload and drowning due to the significant and rapid growth in the amount of information and the number of users. As a result, how to provide Web users with more exactly needed information is becoming a critical issue in web-based information retrieval and Web applications. With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in finding the desired information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge.

As in classical data mining, the aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. There has been huge interest towards web mining. In web mining, this dataset is the huge web data. Web data contains different kinds of information, including, web documents data, web structure data, web log data, and user profiles data. There are two different approaches on the definition of web mining. One approach is process-based and the other is data-based. Data-based definition is more widely accepted today. In this perspective, web mining is the application of data mining techniques to extract knowledge from web data, where at least one of the structure or usage data is used in the mining process. There are no differences between web mining and data mining compared in general. All of web data can be mined mainly in three different dimensions, which are; web content mining, web structure mining, and web usage mining.
The study is mainly related to web usage mining, which is an important branch of web mining. Web usage mining can be defined as the application of data mining techniques to web log data in order to discover user access patterns. Web usage mining has various application areas such as web pre-fetching, link prediction, site reorganization and web personalization. Most important phases of web usage mining are the reconstruction of user sessions by using heuristics techniques and discovering useful patterns from these sessions by using pattern discovery techniques like a priori or similar ones.

The data processing phase for web usage mining can vary with respect to source and processing time. If the requests are processed after they are handled by the web server, the technique is called "reactive" while in "proactive" techniques the same (pre) processing occurs during the interactive browsing of the web site by the user. In proactive strategies, the raw data is collected when web server is processing client requests. Proactive strategies are more appropriate for dynamically created server pages. Also, in proactive strategies, association of an agent with a session is determined during the interaction of user with web site. However, in reactive strategies, the available raw data is mainly server logs containing information about client requests. Reactive strategies are mostly applied on static web pages. Because the content of dynamic web pages changes in time, it is difficult to predict the relationship between web pages and obtain meaningful navigation path patterns.

Web usage mining extracts useful user access pattern information, such as user profiles, from enormous amounts of Web log data for web site personalization. When users browse a website, they usually try to accomplish a certain task, such as finding information, buying products,
registering for classes, and attending classes online. The interaction between the users and the website can give the web engineers insight into the most common user tasks performed on the website. They can learn how most users navigate the website to finish their tasks and what changes can be made to the website structure in order to make the completion of the common tasks easier and faster.

Each user thus has his/her individual profile collected. The more user profiles a personalization system collects, the more data becomes available for precise recommendations. These user profiles are then saved either in flat files or are loaded into the databases. After the data is collected, preprocessing of the original data is conducted, which includes tasks such as, discretization, normalization and so on. Data preprocessing can help with rule evaluation, because well preprocessed data is essential for important rule generation. A rough set based rule evaluation system is demonstrated to show the effectiveness of the measures.

In this study, Association rule mining technique, along with a rough set based learning approach, is presented to enhance website usage. An active-user-based user identification algorithm is applied to the interaction log to group together records that belong to the same user.

To facilitate the knowledge understanding process various association rule generation algorithms are applied to generate rules from the web log data and several rule evaluation measures are applied to facilitate the knowledge understanding process. In order to automatically discover meaningful and important knowledge from huge amounts of web log data rough set reduct generation algorithms are applied for feature selection which identifies the significant features, eliminates the relevant dispensable features to the learning task and builds a good learning model Automatic rule
evaluation measures are applied to extract and rank important rules which form the association patterns. The resulting association patterns which predict user behavior are used for the construction of web personalization.

Then, personalizations are generated by certain rule generation algorithms, such as association rule algorithms, clustering algorithms, classifications and so on. The amount of personalization can be huge when first generated; therefore, post-processing for the generated results is performed in this stage.

There are several rule evaluation techniques available towards rule post processing of rules. The rule interestingness measure is used to find important rules and to rank the rules. Furthermore, how the evaluation measures can be used in a web personalization system is investigated.

The rules generated based on users’ profiles therefore serve as the available knowledge base for personalization systems. In the real world situation, when the personalization system observes a new customer whose profile is an exact match or similar match to the profiles in the databases, the recommendations from the personalization database are generated and provided to this new customer.

Personalization towards individuals recently became an important focus for business applications, such as personalized home pages and a personalized shopping cart. In an online shopping application, individuals’ online purchasing patterns and online browsing experiences may be personalized as well. Such personalization is helpful to predict customers’ interests and to recommend relevant advertisements of interested products to facilitate customers’ online shopping experiences. However an online web user normally browses hundreds of web pages before making a purchase.
online, and different users visit different websites. Personalization based on other people’s past histories may not be very interesting to another user. A user-centric personalization system based on an individual user’s search histories is needed for precise personalization.

1.2 Online Product Purchasing System

The following Figure 1.1 illustrates the prototype of a potential user-centric personalization system, combining data mining and machine learning algorithms on predicting online product purchases. User-centric data is collected and stored in the databases. Features related to user-centric click stream data are selected and the data is preprocessed for the prediction engine. The search terms the users input into the search engines, and the search terms they use on the leading shopping online stores are considered as strong indications of the purchasing interests, and the terms are categorized first to classify potential users into different product purchasing categories. Classification algorithms such as decision tree, logistic regression and Naive Baye’s, association rules algorithms such as a priori, and other prediction algorithms are applied in the following steps to further predict whether a user is an online buyer or non-buyer according to the observed browsing behaviours across multiple websites.
In Figure 1.1, one possible approach of modeling users’ browsing behaviours to study their browsing histories across multiple websites is shown. Personalized products and advertisements through predictions generated by such models can be of great benefit from the business point of view.

1.3 Personalization Systems

Click stream data collected across all the different websites a user visits reflect the user’s behaviours, interests, and preferences more completely than data collected from the perspective of a single website. For example, the intentions of users who not only searched on Google but also visited the HP shopping website and the Dell website can be modeled and predicted than if the users know only one of those pieces of information. The
complete data set is termed user-centric data, which contains site-centric data as a subset. Current research on click stream data analysis is centered around site-centric data. The site-centric personalization systems collect customers’ browsing histories based on the click stream data from the individual web site perspective, and personalizations are generated according to these click stream data to recommend items to the internet users who browse this specific web site. Predictions can be generated for a new customer based on the profile matching of the existing customers (such as name, location, gender, occupation, IP address, operating system and browser information), browsing histories (such as the web pages the customers visited during a certain period of time, application tasks and their sequences the customers performed during a certain period of time), and the preferences of the browsed items (such as some customers expressing great interests on specific items or tasks, whereas some customers show no interest on the same items or tasks).

1.3.1 Site-Centric Personalization

Figure 1.2 shows a model for a site-centric data personalization system. Data collected from different users, including the browsing histories, personal preferences and demographic data, are sent for creating the personalization engine. When a new customer comes, based on the browsing histories and the demographic information, the engine recommends personalized interesting items (such as web pages) to this new user.
Figure 1.2 Personalization for Site-Centric Data

1.3.2 User-Centric Personalization

The personalization systems introduced above are also called “site-centric personalization systems”. Such systems make predictions for “site-centric data”, which is data collected on one single server. Site-centric personalization systems collect customers’ browsing history from the click stream data on the web site side, and personalizations are made based on these click stream data from the site to recommend items to the customers who browse this specific web site. Much current research on click stream data personalization focuses on site-centric personalization. It is important to study the difference between user-centric data and site-centric data, to determine the potential value of the user-centric approach. Due to the limitations of site-centric data it is difficult to fully capture customers’ online shopping behaviours for precise personalization modeling and predictions. The difficulties are explained in the following example.
Let an online shopping and retail website such as shopping.com or Amazon.com as an example, be a site-centric website. Consider the online click stream data collected by this site as site-centric data. Knowledge such as customers’ demographic information, the web pages the customers visited, the time the customers spent on each of the particular web pages, the incoming and outgoing URLs for each of the customers and so on are collected on the server side. Information for the previous web pages each customer visited is collected, and recommendations based mostly on the buyers’ (customers who are observed to make a purchase at this site) behaviours are suggested. For those customers who visited but do not make a purchase at such sites, although later they may make a purchase elsewhere (e.g., HP shopping websites), this site captures the browsing histories for people who visited, but such browsing information is not considered to be important for making online product purchase predictions. The available information is thus not fully captured and utilized. On the other hand, demographic information for customers’ background are also taken into consideration for making recommendations. Therefore, customers’ privacies are not well protected.

User-centric studies are proposed for personalization based on customers’ individual behaviours while greatly preserving customers’ privacy issues. User-centric data is collected to capture each of the individual customers’ browsing behaviours. Data containing customers’ browsing histories, purchasing histories, and so on are then processed for personalization generation. In user-centric data personalization, the limitations of not effectively capturing complete information collected from only certain sites no longer exist.
Users’ web search histories across multiple websites are all used towards the construction of the engine. User-centric personalization has the advantage of protecting users’ privacy as well. By considering more complete information collected from the users, without using their demographic information, the personalized model can fully capture the behaviours while greatly taking care of the privacy issues without using the individual’s demographic information (such as users’ login names, zip code, age, occupations and so on).

The personalization techniques for site-centric data are quite mature, which are techniques originating from traditional web log mining, machine learning, data mining and so on. Given the differences between site-centric data and user-centric data, it is important to study whether these site-centric personalization techniques can be applied to user-centric data, and whether new issues (in terms of data collection, data preprocessing, user behavior modeling and so on) and new challenges should be taken into consideration for user-centric data personalization.

User-centric data is collected for each of the individual user. The data contains users’ browsing histories on all the web sites they visit and their own preferences of interested web sites. Figure 1.3 depicts a sample user-centric personalization system.
1.4 Differences between User-centric and Site-centric Data

1.4.1 Data

Site-centric data is collected from a particular web site due to the limitations of accessing other web sites. User-centric data is collected based on each of the individual users. The data contains click streams from multiple web sites that users browsed.

1.4.2 Session contents

Given a session containing all the users’ browsing history within a limited time sequence (i.e., 30 minutes), the session data for site centric data contains all the web pages a user visited on one website; the session data for user-centric data contains web pages a user visited on multiple websites.

1.5 Rough Set Theory

Rough Set Theory (RST) proposed by Zdzislaw Pawlak is a mathematical approach to intelligent data analysis and data mining. It is an emerging soft computing tool with wide range of applications in many domains especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems,
inductive reasoning and pattern recognition. Rough set is an approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations which are classification of the domain of interest into disjoined categories. The lower approximation is the description of the domain objects which are positively classified as belonging to the subset of interest whereas the upper approximation is a description of objects which possibly belong to the subset. The rough set approach provides efficient algorithms for finding out hidden patterns in data, minimal sets of data (data reduction), evaluating significance of data and generating sets of decision rules from data. In RST all computation are performed directly on data sets. It requires no additional parameters to operate such as thresholds or a grade of membership as in fuzzy set theory, other than the supplied data. It works by making use of the granularity structure of the data. One advantage of Rough Set theory is that, it provides a well understood formal model, which is very helpful in generating several kinds of information such as relevant features or association rules using minimal model assumptions.

1.5.1 Rough Sets based Knowledge Discovery Systems

A brief survey of current rough sets based knowledge discovery systems is undertaken. The individual functions of each system based on general characteristics, such as the data sets, the preprocessing tasks, the related rough sets tasks, the rule generations and so on are discussed.

1.5.1.1. ROSETTA : The software supports the complete data mining process, from data preprocessing, including handling incomplete data, data discretization, generating reduct sets which contain essential attributes for the given data set, to classification, rule generation, and cross validation evaluation. Some discretization and reducts generation packages are from the RSES library.
1.5.1.2. **RSES 2.2** : The system supports data preprocessing, handling incomplete data, data decomposition, reducts generation, classification, and cross validations.

1.5.1.3. **ROSE 2** : ROSE stands for Rough Sets Data Explorer. This software is designed to process data with large boundary regions. The software supports data preprocessing, data discretization, handling missing values, core and reducts generation, classifications and rule generation, as well as evaluations. This software provides not only the classical rough set model, but also the variable precision model.

1.5.1.4. **LERS**: LERS stands for Learning from Examples based on Rough Sets. It is not publicly available. The system was designed especially to process missing values of attributes and inconsistency in the data set. Certain rules and possible rules are both extracted based on the lower and upper approximations. In addition to the rough sets based systems mentioned above, there are other available knowledge discovery systems based on the methodologies of rough sets such as DBROUGH and GROBIAN.

These systems demonstrate the use of rough sets theory for knowledge discovery by several researchers. ROSETTA is used for reduct generation. The software provides Genetic reducer, Johnson reducer, Holte1R reducer, Manual reducer, Dynamic reducer, RSES Exhaustive reducer and so on. Genetic reducer is an approximation algorithm based on a genetic algorithm for multiple reducts generation. The Johnson reducer generates only a single reduct with minimum number of attributes.
1.6 Association Rules

The association rule algorithm was first introduced and is commonly referred to as the a priori association rule algorithm. It can be used to discover rules from transaction datasets. The algorithm first generates frequent itemsets, which are sets of items that have transaction support more than the minimum support; then based on these itemsets, the association rules are generated which satisfy the minimum confidence. Many contributions on how to efficiently generate frequent itemsets and generate rules have been reported.

Association rule algorithms can be used to find associations among items from transactions. For example, in market basket analysis, by analyzing transaction records from the market, one could use association rule algorithms to discover different shopping behaviours such as, when customers buy bread, they will probably buy milk. This type of behavior can be used in the market analysis to increase the amount of milk sold in the market.

An association rule is a rule of the form $\alpha \rightarrow \beta$, where $\alpha$ and $\beta$ represent itemsets which do not share common items and, $\alpha$ - antecedent, and $\beta$ – consequent. The association rule $\alpha \rightarrow \beta$ holds in the transaction set $D$ with confidence $c$ if $c\%$ of transactions in $D$ that contain $\alpha$ also contain $\beta$. Confidence can be represented as $c = |\alpha \cup \beta| / |\alpha|$. The rule $\alpha \rightarrow \beta$ has support $s$ in the transaction set $D$ if $s\%$ of transactions in $D$ contain $\alpha \cup \beta$. Support can be represented as $s = |\alpha \cup \beta| / |D|$ .Confidence gives a ratio of the number of transactions that the antecedent and the consequent appear together to the number of transactions the antecedent appears. Support measures how often the antecedent and the consequent appear together in the transaction set.
1.7 Statement of the Problem

In order to automatically discover meaningful and important rules from huge amount of web log data for web personalization, the study focuses on obtaining web access patterns of the users from the user’s profiles based on intelligent rough set rule mining approach. A profile can consist of a set of URLs (Uniform Resource Locator) that are relevant to the sessions. Once these profiles are discovered, they can be exploited as part of an automated personalization on the website, by treating them as summarized user models against which all future user sessions are compared. The discovered knowledge or any unexpected rules are likely to be imprecise or incomplete, which requires a framework with soft computing techniques like rough sets. The study provides efficient algorithm for finding hidden patterns in web log data in the form of rules that are automatically generated from the given data. The investigation of this thesis presents rule evaluation techniques which help not only to reduce the number of rules but also to extract higher quality rules. Empirical studies on both artificial data sets and real world data sets demonstrate how such techniques can contribute to practical systems.

1.8 Objectives of the Study

The study focuses on using rough set based association rule mining to investigate a database of URL information regarding access to electronic sources. The specific objective is to determine the different rules that describe associations between sets of items in the web server user access logs. The aim of this study is to perform association rules analysis between web pages and the users to extract sequential association rules among large sets of URL links.
The specific objectives of the study are as follows.

- To apply rough sets theory to knowledge discovery in web log data.
- To preprocess the data set taken from Internet Information Server.
- To apply association rule algorithms to extract item-item relationships among large web log data set.
- To evaluate the association rules in order to rank the important rules
- To extract user access pattern information from web log data for web site personalization.
- To assess the performance of interestingness measures which evaluate the generated rules.
- To make suggestion and recommendations for the effective use of web sites.

1.9 Scope of the Study

A challenging problem in rule generation is that an extensive number of rules are extracted by data mining algorithms over large data sets, and it is infeasible for human beings to select important, useful, and interesting rules manually. How to develop measures to automatically extract and evaluate interesting, relevant, and novel rules becomes a requirement of the area. Many existing methods such as rule interestingness measures and rule quality measures are reported in.
Comparisons of different measures are reported for general purpose applications. A measure is said to be a subjective measure if it is defined based on a domain expert’s opinions towards the particular application. A measure is an objective measure if it measures the data itself without any predefined opinions. Subjective measures that use real human evaluators are the optimal measure to evaluate rules, although they are sometimes infeasible and expensive, because they may require humans to look at a large number of rules. Previous studies on rule evaluations focus mostly on objective measures, which do not contain any knowledge from the domain of the data. Therefore such objective measures may not sufficiently evaluate whether a rule is indeed interesting for a certain domain.

In recent years, with the rapid growth of the amount of information on the web, most research in web mining has concern on web personalization. Web usage mining techniques have been widely used for discovery of interesting usage patterns from access log files. Soft computing methods neural networks, fuzzy logic, genetic algorithms, and rough sets, etc., have been intensively used in web usage mining studies.
1.10 Methodology

This thesis applies rough set theory to the extraction of association rules in Web log data and evaluates the rules to rank the important rules and uses the rule set for web personalization.

![Design of the Web Personalization](image)

**Figure 1.4 Design of the Web Personalization**

First, the study performs data preprocessing on web log data, a step in which salient features are selected using rough set. From the preprocessed Web log data, Web access sequence lists are constructed. Second, decision tables are constructed according to the length of the list in order to apply rough set theory. Then the rules are generated on the decision tables by
association rule mining algorithms. After the rules are generated, rule evaluations are performed. Redundant or not important rules are removed by rule interesting measures, and useful rules are presented as knowledge as the output of the system. Third, the user access behaviors are predicted by the use of rule set for web personalization.

The rules generated based on users’ profiles therefore serve as the available knowledge base for personalization systems. In the real world situation, when the personalization system observes a new user whose profile is an exact match or similar match to the profiles in the databases, the recommendations from the personalization database are generated and provided to this new user.

The study analyses several rule evaluation measures for the purpose of facilitating the knowledge understanding process. The aim of the study is to design automatic rule evaluation measures that can bring both domain related knowledge and the objective measures together into the rule evaluations. Such measures are proposed to help to extract and rank important rules from a large number of rules generated by a learning algorithm. The rules generated based on users’ profiles therefore serve as the available knowledge base for personalization systems. In the real world situation, when the personalization system observes a new user whose profile is an exact match or similar match to the profiles in the databases, the recommendations from the personalization database are generated and provided to this new user.
1.11 Chapterisation Scheme

The present study has been organized into six chapters. The layout of the chapters is delineated below.

Chapter I

Encompasses the research methodology of the study. This chapter contains Introduction, scope of the study, statement of the problem, objectives of the study, research methodology and chapterisation scheme.

Chapter II

Presents a brief review of related literature.

Chapter III


Chapter IV

Outlines the important concepts of rough sets theory and current rough set based knowledge discovery system. A novel rule interestingness measures to facilitate recommending rules generated by association rule algorithms is also discussed.

Chapter V

Describes the performance of preprocessing the web log data, reduct generation, generation and evaluation of association rules to rank the important rules to construct a web personalization system.

Chapter VI

Embodies the summary of recommendations for the construction of web personalization. Towards the end of this research work future areas of research on this related topic and comprehensive references have been added.
REFERENCES


7. Jiye Li and Nick Cercone. A rough set based model to rank the importance of association rules. In Dominik Slezak, Jingtao Yao, James F. Peters, Wojciech Ziarko, and Xiaohua Hu, editors, RSFDGrC (2), volume 3642 of Lecture Notes in Computer


12. Jiye Li and Nick Cercone. Empirical analysis on the geriatric care data set using rough sets theory. Technical report, School of Computer Science, University of Waterloo,


