University of Modern Sciences

A Proposed Approach to Generate Dynamic User Profiles Based on Periodic Web Usage Mining

By

Emad Mahmoud Mohammed Al Azazi

Supervisor

Prof. Ahmed Sultan Al-Hegami

Submitted to Graduate Studies of University of Modern Sciences to fulfillment of the requirements for the degree of Master of Department information technology

2022

ACKNOWLEDGEMENTS

In the name of ALLAH, the Beneficent, the Compassionate, and the one who gives me strength and patience to complete my duties despite many challenges situations during my research study.

I would like to take this opportunity to thank everyone who has helped me throughout my journey. First and foremost, I wish to acknowledge the support, guidance, and valuable observations accorded by my supervisor, Prof. Ahmad Sultan Al-hegami who gave me a lot of good intellectual help and valuable guidance to face the challenges l have faced throughout my Master study. Professor Ahmed Sultan has always take time to listen to my ideas, and he has patiently answered my questions, with fruitful discussions, patience and continued encouragement supply at every stage of this work. Furthermore, he always provides the gold recommendations and suggestions to my inquiries tranquilly and accurately.

Besides my advisor, I would like to express my sincere gratitude for My Father for his encouragement And to provide moral and material support for Me during My study.

Many thanks to all my friends in In graduate studies for the many fruitful discussions we have shared together. I would specially like to thank my close friends : Abu Alfath Aldhabey, Abdu Alrahman Alshareef, Bassam AlKebsi, Mohamed Saree, Najeeb Albazeli, Ahmed Aladran

Finally yet importantly, I am very grateful to my Family for unconditional love and unlimited support in my live. I would specially lik to thank my father, my mother, and my Wife. I appreciate all the opportunities they made available to me. I would like to convey my sincere gratitude to my great brothers and sisters.

ABSTRACT

Nowadays, personalization of web field is growing exponentially from e-mail, e-trading, internet forums to social networking based websites directly or indirectly to utilize web personalization and recommendation system for providing customized services to their users. Successful businesses is growing through adopting most suitable approaches that analyze the customer’s behaviors to provide them with the services according to their behaviors. So web personalization systems have been established to enhance user’s experience that stored in the "user profile". Performance of Web personalization systems depends on building the user profile efficiently to capture the user’s interests and reflect them to provide him with most useful services. Actually, various valuable browsing patterns happen repeatedly during a specific period of time. During other time periods this can depict the user surfing habits.

Regular web personalization utilization is using the regular web usage mining activities that do not depend on the user's present browsing information. This can be largely ambiguous and eventually which result in constraining the personalization process and recommending irrelevant services and items over time. Periodic web personalization approaches simply can be an intermediate web page prior to the current browsing to reach the requested page.

Many methods for building the user profile are available that neglect the dynamic nature of the user interests. In this thesis, a solution this problem through novel approach to extract interesting patterns that occur during the user's interaction with different resources through browsing period, its "browsing patterns". This approach should be able to track the changes in user interests that called periodic user profile, and based on these patterns, the appropriate services are provide to the user.

In this thesis, an approach is proposed for Periodic browsing patterns extracted at each browsing period and evaluated by a novelty measure, whether these patterns are novel or not, based on the new patterns, the model is updated. Thus, the user profile built dynamically and incrementally.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

KDD	Knowledge Discovery in Database
DK	Domain Knowledge
KK	Known Knowledge
PDK	Previously Discovered Knowledge
NM	Novelty Measure
SW	Semantic Web
RF	Relevance Factor
LCS	Longest Common Subsequence
URL	Uniform Resource Locator
PRNG	A pseudorandom graph
OLAP	Online analytical processing

CHAPTER 1

1.1 Introduction

Community intelligence business need an expert decision-making[73]. Most of today's websites designed with a one-size-fits-all philosophy: a single view of (the content and navigational structure) for each visitor in all conditions. However, one size of the structure does not fit all: visitors are browsing from different devices, and they look for different contents, so they have different styles of navigation where the value of web interaction maximized. Web personalization systems to personalize the experience for each user. It is of a two-step process: the first is modeling the users and then adapting the content to meet user best needs. Periodic web usage and a novel approach provide an expert decision to determine the resources that the user is interested in during a given period. Constructing the personalization page in the real time is very expensive [3]. To overcome this problem, a proposed approach "periodic web usage mining for web personalization" in order to minimize processing the overload at runtime and to provide the user with the appropriate service according to his behavior.

The Proposed incremental algorithm and the novel approach integrate interests criteria during the process of building the model. One of the main features of this approach is capturing the user background knowledge, which is monotonically augmented. One of the main drawbacks with the classical algorithms do not consider the time in which the data arrived[60,61]. In practice, data is acquired in small batches over the time. In such scenario a combination of old and new data is used to build a new model from scratch that results in losing the previously discovered knowledge (PDK). Therefore, researchers have been strongly motivated to propose techniques that update the model as new arrival data, rather than running the algorithms from scratch, resulting in incremental models. Incremental algorithms build and refine the model as new data arrival at different time intervals, in comparison to the traditional algorithms where they perform model building in batch manner [60,61,62,63,64]. The incremental algorithms that reflect the changing data trends and the user beliefs are attractive in order to make the over all web usage mining process more effective and efficient.

1.2 Problem Statement

A number of studies provide different ways to model user profiles[74]. Most of such ways do not reflect the user interests in the current browsing, however they refer to other time periods. The periodic nature of browsing patterns is repeatedly ignored while, the browsing patterns happen during a specific period of time that can depict the user surfing habits at the time and don't during other time periods.

Regular web personalization utilization is using the regular web usage mining activities and do not depend on the user's present browsing information. This can be largely ambiguous and eventually result in constraining the personalization process and recommending irrelevant services and items over time. To overcome this problem through this novel approach to build the user profile dynamically that should be able to track the changes in the user interests including any interest-shift or drift and adapt accordingly.

1.3 Research Objective

Building user profile, which represents the user's interests for Periodic accessibility can help developing effective models for constructing user profiles. This can capture the user's interests dynamically. Our goal in this thesis is to establish an approach for the users dynamically who share common interests for periodic accessibility and also for having better performance in analyzing user behaviors which helping finding knowledge about the web usage. Building a user profile that represent the user's interests, the recommendations to individual users are provided in the best quality. Consequently, the appropriate service for the user is provided according to his behavior.

1.4 Motivations

Most available personalization systems do not capture and track the change in the user behaviors and they don't indicate at the user behavior in a holistic way.

In practice, the data is acquired in small batches over the time. Therefore, researchers have been strongly motivated to propose techniques that update the model as new data arrived, rather than running the algorithms from scratch.

We propose to personalize the experience for each user to maximize the value of the web interaction. User profile changes dynamically according to the client's interests, because each client gets the services needed. A proposed approach dynamically learns the user's areas of interest, at different period. One of the main features of this approach is to capture the user knowledge incrementally and dealing with time changing data and user beliefs. It is attractive to update the discovered knowledge each time new data arrive.

1.5 Research Organization

The thesis is to be broadly organized as follows:

Chapter 1 introduction and the major problems that has been addressed through this thesis, and what are the objectives of the proposed method and the motives for it .

In chapter 2, We review the literature related to web usage mining and dynamic user profile.

The chapter3, Describes the proposed incremental methods and interestingness measures to building user profile

Chapter4 presents results of a computational implementation and experimentation

Finally, chapter 5 concludes the thesis.

CHAPTER 2

LITERATURE REVIEW AND RELATED WORK

2.1 Introduction

World Wide Web is a major source of information and it creates new challenges of information retrieval as the amount of information on the web increasing exponentially so came the need to web mining[3]. Where the application of data mining techniques to extract and analyze useful information from Web data. Based on kind of data to be mined web mining can be classified into three different categories: Web content mining, web structure mining and web usage mining [39]. Web usage mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from web log files on the web server are major source of data for web usage mining. In web usage mining, data mining techniques are applied to preprocessed web log data in order to find interesting and useful patterns Based on these patterns and behavior, there are systems that adapt web content or provide appropriate services to users as including Yahoo!, Amazon, eBay, Netflix, NewsWeeder, IBM and many more[75]. A recommendation system learns from a customer's behavior and recommends a product in which users may be interested. It helps to build up a long lasting relationship with loyal users of website. Various vendors offer web personalization tools that can be employed in existing systems to achieve personalized web system[12]. Personalization is a relatively new and challenging field for web content delivery[4]. In order to meet expectations of visitors, customers and loyal users, web world is struggling to offer excellent customized services during their interaction with the system. The impact of personalization and recommendation system can be experienced by the rapid popularity that this area has gained in the last few years. Customers preferably choose to visit those websites, which understand their needs, provide them rapid value added customized services and easy access to required information in simple understandable format. Web personalization and recommendation system plays a major role in meeting this goal [14].

A number of techniques have been explored in the web usage logs for web personalization to discover various web accessibility patterns[25]. Even though, the periodic nature of browsing patterns is repeatedly ignored. Actually, various valuable browsing patterns happen repeatedly during a specific time period, this can depict a user surfing habits, but not during other time periods. These browsing patterns are called periodic browsing patterns. Many different techniques have been considered to discover patterns from web browsing logs for web personalization. It consist of rule-based filtering approaches, content-filtering database to approach, collaborative filtering approach, and "hybrid approaches [15]. The personalization of a web page dynamically changes the content of a web page according to the preferences of the client or client, so that each client gets the information that suits their needs. Periodic web personalization is the process by which a Web server and its associated applications vigorously modify content for a specific user depend on information about their activity on a website above a specific period of time [16]. User profile is collected through two ways[46], explicit profile, and implicit profile, Explicit profile. The attributes either come from online questionnaires or registration forms: all this is an external feedback that is collected by the user that provides rating or preferences. User personal information like age, religion, economic ,environment etc. It can be used to generate user profile. Implicit profile can be derived from browsing patterns, cookies, and other sources. In this case there is no external feedback collected but it is based on the characteristics of an individual such as his interest.

User profile can represented by two main types of the user profiles[76] , Content-based profile, and Collaborative profile. In Content-based profile this approach based on similar queries as a vector of terms. Collaborative profile in this approach is based on patterns of the similar users. It is assumed that individuals with similar patterns “think similarly”. So, the collaborative profile can be expressed as a list of similar users[17]. The proposed approach in this thesis is related to the second one of these techniques

2.2 Web Personalization

Web personalization and recommendation system utilizes data, which is collected explicitly or implicitly during the interaction of user with website. Such a collection of data is known as web data that can be divided into four major categories: Content data, Structure data, Usage data and User profile data [25].Content data is intended for end-user in simple text format, images or structured information retrieved from databases. Structure data represent how the contents are organized internally. It may include data entities used in web pages, such as HTML or XML tags, and hyperlinks to interconnect web pages. Usage data represents usage of website, it consists of visitors’ IP Address, time and date of access, complete paths of files or directories accessed, and other attributes that can be included in a Web access log [25]. User profile data comprises of personal information of each website user, such as name, age, sex, country, qualification etc. and information about users interests and preferences. It is explicitly provided by user during the interaction with website while filling registration forms or questionnaires. Once the web data is procured, task of personalization can be initiated. The process of web personalization includes identifying the visitors of website, retrieving their profile data, selecting contents that suit to their profile, and then displaying those contents in most pleasing and easily understandable format.

2.3 Web Mining

Web Mining is use of Data Mining techniques to automatically discover and extract information from web data [46]. Web mining is the general name of the data mining technique used in an attempt to make content analysis from the online web sites. Web mining has the Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user click streams stored in web log files . The method of discovering useful knowledge from collected web data is known as web mining [27], which can take any of the following three forms as shown in Figure 2.1 [11, 28].

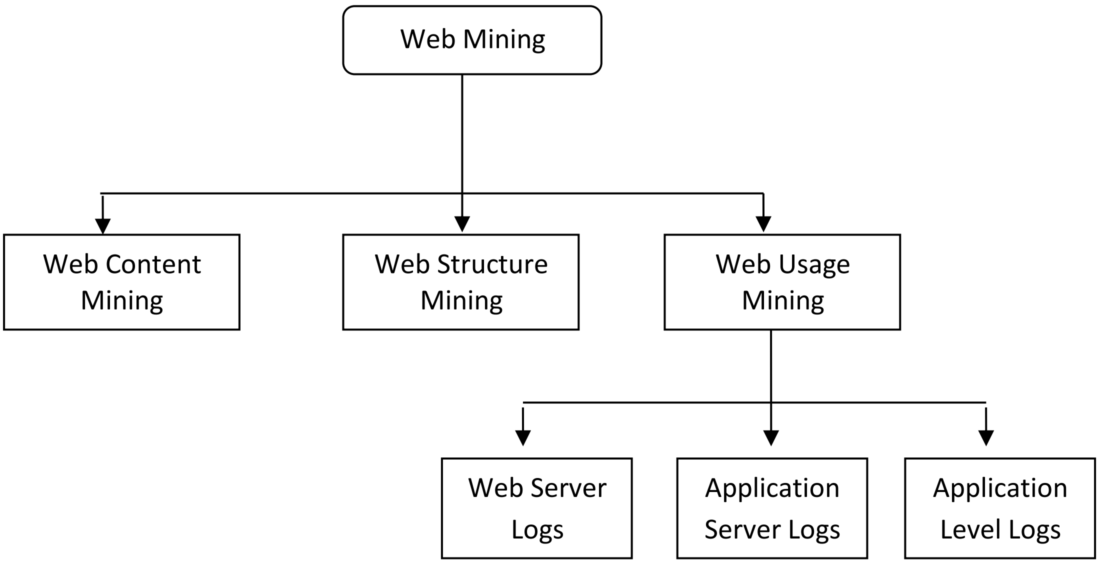


Figure 2.1 :Taxonomy of Web Mining

Content data is intended for end-user in simple text format, retrieved from databases. Structure data represent how the contents are organized internally. It may include data entities used in web pages, such as HTML or XML tags, and hyperlinks to interconnect web pages. Usage data represents usage of website, it consists of visitors' IP Address, time and date of access, complete paths of files or directories accessed. User profile data comprises of personal information of each website user, and information about users interests and preferences.

A proposed approach is more interested in personalization system through web usage mining, which plays a major role in predicting user interests and behavior. These predictions act as foundation of personalization and recommendation systems.

To analyze the web data, various approaches have been proposed by researchers across the globe in past few years [25] classified these approaches into four major classes: Content-based filtering systems, social or collaborative filtering systems, manual rule-based filtering systems, and web usage mining based systems. Content-based filtering systems model the behavior of individual user based on his past interests, personal preferences, and browsing behavior. Once user modeling is completed, system starts recommending items (to users) that match individual user's profile.

The goal of collaborative filtering systems is to achieve personalization without analyzing the web contents. Such systems invite users to explicitly rate the available items, or reveal their personal preferences which are recorded by the system. The system then performs categorization of users, based upon the information provided by them. Such recommendation engines work on assumption that the users with similar behavior have correlated interests. Thus depending upon the category in which a user falls, system can acquire the information that a particular user may be interested in. Manual rule-based filtering system requires manual intervention of website designer and user's co-operation in order to achieve personalization. Under this approach, a set of questionnaires derived from a decision tree is presented to users[39]. Based on the answers given by user, a set of rules are defined manually, and a static user model is created. Depending upon the underlying rules and user model, contents of web pages are tailored according to user's needs. In the proposed approach concentrate on web usage mining based systems that focuses on discovering interesting patterns from usage data. It is now widely recognized that usage mining is a valuable source of ideas and solutions for web personalization [10].

The web usage data represents details of user-website interaction, which is appropriate to create user model that represents user's behavior, interests and personal preferences. The constructed user model can be used by personalization system in fully automated way without any human interference, for carrying out the personalization task. A general architecture for web usage mining shown in Figure 2.2

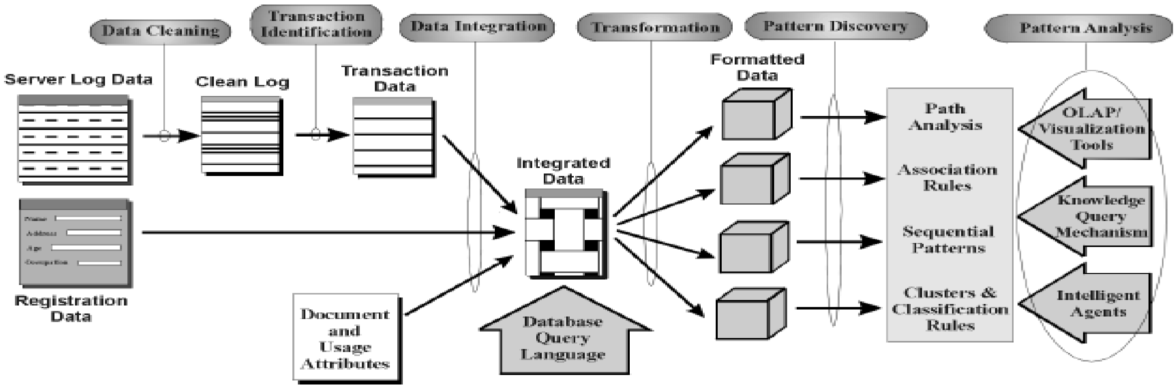


Figure 2.2 :General Architecture for Web Usage Mining[77]

2.4 Web Usage Mining

2.4.1 Periodic Web Usage Mining

The web personalization is a two step process: the first one is modeling the users and then adapting his content to meet the best needs of each user. Periodic web usage constructing the personalization page in the Real time and provide expert decision to determine the resources that the user is interested in during a given period.

2.4.2 Web Usage Mining In Web Personalization

When data mining techniques are applied on web usage data in order to extract useful knowledge regarding user behavior. It is an approach for collecting and preprocessing web usage data, and then constructing models that represent the behavior and interests of users. Such models can automatically be used by personalization system for predicting user's personal interests and thus enhance his surfing experience with the website. The web usage mining involves[78]: collection of data from various sources, pre-processing of collected data, discovering useful knowledge, and finally post processing the knowledge.

- **Collection of Data From Various Sources**

The log file : what is it and how do we store information?

The log file definition : A log file is defined as “a file that lists actions that have occurred” [40]. These files are generated by servers – a computer or a device on a network that manages network resources and contain a list of all requests made to the server by the network’s users. A web log file [41] records activity information when a web user submits a request to a web Server. The main source of raw data is the web access log, which we shall refer to as log file, these logfiles can be located in places [42,43].

Proxy Server- A proxy server is a intermediary compute that acts as a computer hub through which user requests are processed. Web Client- A Web client is a computer application, such as a web browser, that runs on a user local computer or workstation and connects to a server as necessary. A web server log file is a log file that automatically creates and maintains the activities performed in it. This file is used to record each and every hit to a web site [38]. Table 2.1 represents the sample web server log entries

IP address	**Date and time of request**	Request	Status	Bytes	Referer	User_agent
128.101.34.92	**[08Mar2018 00:04:44 +1200]**	"GET/harum/HTTP/1.0	200	3014	http:/www.msss.edu	Mozilla/4.7[en];sunos5.8sun4

Table 2.1 represents the sample web server log entries

Where entries in the server log mean: Sources for collecting web usage data through web server.

IP address: IP address of the remote host.

Rfc931: the remote login name of the user.

Authuser: the username as which the user has authenticated himself.

Date: date and time of the request.

Request:the request line exactly as it came from the client.

Status: the HTTP response code returned to the client.

Bytes: The number of bytes transferred.

Referer: The URL the client was on before requesting your url.

User_agent: The software the client claims to be using.

2.4.3 Data Pre-Processing

The data collected in web log file is incomplete and not suitable for mining directly. The data collected during first stage is usually diverse and voluminous. Therefore it is necessary to preprocess it by filtering unnecessary and irrelevant data, predicting and filling in missing values, removing noise, transforming it into more useful format for pattern discovery, and resolving the inconsistencies. In Web usage mining, this stage includes identifying the users and their sessions. In addition, it is also necessary to provide a good trade-off between insufficient preprocessing and excessive preprocessing [24]. Data pre-processing consists of data filtering, user identification, and user session identification

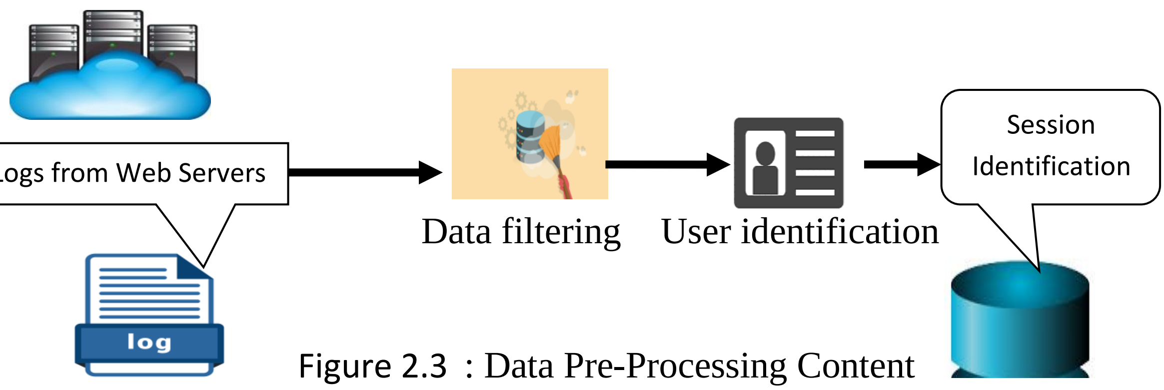


Figure 2.3 : Data Pre-Processing Content

As shown in the Figure 2.4 for data pre-processing consists : Data filtering (or data cleaning) is the first step in data preprocessing, where the major task is to clean and filter the raw web data. During this step, the available data is examined, and irrelevant or redundant items are removed from it. Data generated by client-side agents does not require cleaning as it is intentionally collected by the software agents. Since data preprocessing task is domain-dependent hence depending upon the domain of website, it is necessary to differentiate between relevant and irrelevant data, otherwise it may cause loss of valuable information. The records created by spiders and crawlers are also deleted during filtering since it is not considered as usage data [51].

User identification is very important because it constitutes a basic processing unit for discovery of interesting, prominent access patterns and so it largely affects the quality of pattern discovery result. The construction of the user activities in sessions for a particular web site contains a correct mapping of activities to distinct web users. A long sequence of visits by the users has to be broken into user sessions as Table 2.2 , Currently a user informs the system explicitly about who he/she is, by entering a pair of username and password. But a lot of research is being conducted for automating the process of user identification.

User session is a delimited set of web pages visited by a particular user in single visit to the website. Identification of user sessions has also received significant attention as it reveal the navigational behavior and surfing habits of a user, which forms the foundation of personalization system. A user may have a single or multiple sessions during a particular time period. Once a user is identified, the click stream of each user is segmented into logical clusters and this process is known as sessionization or session reconstruction as shown sequence of visits by the users broken into user sessions in the next Table 2.2 .

Time	IP Address	URL	Agent
13:44:29	192.168.1.3	A	MSIE 6.0 Windows NT 5.1
13:48:22	192.168.1.3	B	MSIE 6.0 Windows NT 5.1
13:49:09	192.168.1.3	C	MSIE 6.0 Windows NT 6.0
13:53:13	192.168.1.3	D	MSIE 6.0 Windows NT 5.1
14:54:23	192.168.1.6	A	MSIE 6.0 Windows NT 6.0
14:55:18	192.168.1.6	C	MSIE 6.0 Windows NT 5.1
15:14:07	192.168.1.7	B	MSIE 6.0 Windows NT 5.1
15:28:12	192.168.1.7	C	MSIE 6.0 Windows NT 6.0
15:29:39	192.168.1.7	D	MSIE 6.0 Windows NT 5.1
15:30:24	192.168.1.7	A	MSIE 6.0 Windows NT 6.0

16:24:20	192.168.1.4	C	MSIE 6.0 Windows NT 5.1
16:30:26	192.168.1.6	B	MSIE 6.0 Windows NT 5.1

The result is a list where all entries generated by the same user are clustered together and stored

Session 1 for user 1

13:44:29	192.168.1.3	A	MSIE 6.0 Windows NT 5.1
13:48:22	192.168.1.3	B	MSIE 6.0 Windows NT 5.1
13:49:09	192.168.1.3	C	MSIE 6.0 Windows NT 6.0
13:53:13	192.168.1.3	D	MSIE 6.0 Windows NT 5.1

Session 1 for user 2

15:14:07	192.168.1.7	B	MSIE 6.0 Windows NT 5.1
15:28:12	192.168.1.7	C	MSIE 6.0 Windows NT 6.0
15:29:39	192.168.1.7	D	MSIE 6.0 Windows NT 5.1
15:30:24	192.168.1.7	A	MSIE 6.0 Windows NT 6.0

Session 1 for user 3

16:24:20	192.168.1.4	C	MSIE 6.0 Windows NT 5.1
16:30:26	192.168.1.6	B	MSIE 6.0 Windows NT 5.1

2.4.5 Pattern Discovery

To discover potentially useful and interesting information, several methods and data mining algorithms are applied such as path analysis, association rules, sequential patterns, clustering, classification etc. After identifying user sessions, the various techniques of web usage pattern discovery are applied in order to detect interesting and useful patterns. There are several kinds of access pattern mining that can be performed depending on the needs of the analyst. Some of pattern discovery techniques are discussed below. This stage employs machine learning and statistical methods on pre-processed web data in order to extract the patterns of website usage. There are four major machine learning approaches: clustering, classification, association discovery and sequential pattern discovery [10].

Clustering is a technique to group those users who exhibit similar browsing patterns, or web pages which exhibit similar contents [20]. This approach is employed in majority of pattern discovering methods. Web usage mining allows the overlapping of clusters [11]. It is very important in web personalization because a user or a web page may not necessarily belong to a single group. According to [19] clustering methods can be classified as: Partitioning methods, Hierarchical methods, and Model-based methods. First approach is used to break up a given data set into n clusters (groups). Second approach is used to decompose a given data set into a hierarchical structure of clusters, and third approach is used to find the best match between a given data set and a mathematical model. So far, a number of clustering approaches have been proposed by various researchers, Classification of pre-processed data involves assigning a web page or a user to one or more predefined classes. It also helps to develop a profile for items belonging to a particular group according to their common attributes. It is a supervised learning problem [40] where a set of labeled data acts as target vector for training a classifier that can be employed for labeling future data.

In classification analysis, data items are classified according to predefined categories. If it is needed to develop a profile of user belonging to a particular class or category then features are extracted that well describe the properties of given class or category. There are many algorithms such as decision trees, neural networks, Bayesian classifier, and probability theory for classification [4]. After classification, business activities can start according to the characteristics of this type of clients, providing targeted personalized information services. In e-commerce, after classifying the data with the same characteristics, e-commerce enterprise can provide personalized information services according to the characteristics of such customers.

Association discovery the process of mining association rules consists of two main parts. First, we have to identify all the itemsets contained in the data that are adequate for mining association rules. These combinations have to show at least a certain frequency to be worth mining and are thus called frequent itemsets . The second step will generate rules out of the discovered frequent itemsets. In the first step mining frequent patterns from a given dataset is not a trivial task. All sets of items that occur at least as frequently as a user-specified minimum support have to be identified at this step. An important issue is the computation time because when it comes to large databases there might be a lot of possible Itemsets all of which need to be evaluated. Different algorithms attempt to allow efficient discovery of frequent patterns. After having generated all patterns that meet the minimum support requirements, rules can be generated out of them. For doing so, a minimum confidence has to be defined. The task is to generate all possible rules in the frequent itemsets and then compare their confidence value with the minimum confidence (which is again defined by the user). All rules that meet this requirement are regarded as interesting. Frequent sets that do not include any interesting rules do not have to be considered anymore. All the discovered rules can in the end be presented to the user with their support and confidence values. The goal of sequential pattern discovery is to identify those event sequences, which frequently occur in the dataset. It is useful in identifying the navigational patterns of users. Under this approach, two types of method are most popular: deterministic methods and stochastic methods. Deterministic methods employ recording of navigational behavior users, whereas stochastic methods analyze the sequence of web pages visited by users for predicting the web pages that the users may be interested to visit in future[29].

- **Path Analysis**

A graph can be formed to perform path analysis. A graph represents some relation defined on web pages. The physical layout of the web site can be represented by graph in that web pages are nodes and link between pages are directed edges. Using it frequent paths traversed by users, entry and exit points can be determined easily [36]. For example what paths do users traverse before they go to particular URL? What percentage of clients left the site after five or less page references?

- **Sequential Patterns**

The technique of sequential pattern discovery can be applied to web server logs. It attempts to find intersession patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions [39]. In e-commerce, customer access data is recorded in the web server log with a unit of a time period. Using sequential pattern discovery, useful user trends can be discovered, website navigation can be improved and adopt web site contents to individual client requirements or to provide clients with automatic recommendations that best suit customer profiles. The user's visit patterns can also be predicted using sequential pattern discovery in web server log and helps in targeting adverting aimed at groups of users based on these patterns. The sequential patterns can be discovered as the following form :x% of client who bought items using URLA, also placed an order online within 10 days using URLB.

- **Association Rules**

The correlations between web pages that are most often referenced together in a single user session can be discovered by association rules. In e-commerce, association rules can be used to find the relevant pages browsed by customers. It can provide the information: What are the set of pages frequently accessed together by web users? What page will be fetched next? What are paths frequently accessed by web users?.

Implement association rules to on-line shopper can generally find out his/her spending habits on some related products [35,50].

For example, if a transaction of an on-line shopper consists of a set of items, while each item has a separate URL. Then the shopper's buying pattern will be recorded in the log file, and the knowledge mined from it, can be the form like this: x% of clients who accessed the web page with URL A.html, also accessed B.html, y% of clients who accessed S.html, placed an online order in P.html. The association rule can be used to restructure the web site by adding links that interconnect pages which are often viewed together

- Fuzzy Association Rules

Most real-life data are neither only binary nor only numerical but a combination of both. Quantitative attributes such as age, take values from a partially ordered, numerical scale which is often a subset of the real numbers. The general method adopted is to convert numerical attributes into binary attributes using ranges (for example, any numeric value for attribute Age would fit in ranges like “up to25”, “25-60”, “60 and above”). This reduces the paradigm to traditional association rule mining with binary values. A better way to solve this problem is to have attribute values represented in the interval [0, 1], instead of just 0 and 1, and to have transactions with a given attribute represented to a certain extent (in the range [0, 1]) [31]. Thus, the proposed approach uses fuzzy methods, by which quantitative values for numerical attributes are converted to fuzzy binary values. Doing so ensures that there is no loss of information whatever may be the value of any numerical attribute. Moreover, the inherent uncertainty that is present in numerical data is also appropriately taken care of. The corresponding mining process yields fuzzy association rules. For example, we have a crisp association rule like “People between the ages of [25, 41] earn an income [$80K, $150K] . With this rule a 24-year old person earning $100K is not accounted for. But a fuzzy association rule like, “Middle-aged people earn high incomes”, is more flexible, and reflects this person’s salary in a more appropriate manner.

Based on classical association rule mining, a new approach has been developed expanding it by using fuzzy sets. The new fuzzy association rule mining approach emerged out of the necessity to mine quantitative data frequently present in databases efficiently. When dividing an attribute in the data into sets covering certain ranges of values, we are confronted with the sharp boundary problem as shown in Figure (2.4)

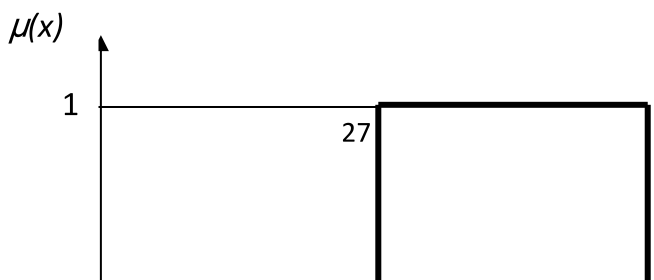


Figure 2.4 : Sharp Boundary Problem

Elements near the boundaries of a crisp set will either be ignored or overemphasized [30]. For example, one can consider a set representing persons of middle age, ranging from 30 to 50 years old . In this example, a person aged 29 years would be a 0% representative and a 31 year old would be 100%. In reality, the difference between those ages is not that great. Implementing fuzziness can overcome this problem. The same problem can occur if one is dealing with categorical data. Sometimes, it is not ultimately possible to assign an item to a category. As an example, one can say that a tomato is a vegetable but also, in a way, a fruit. Crisp sets would only allow assigning the item to one single category, fuzzy sets allow different grades of membership to more than one set. Three different approaches to fuzzy association rules : The quantitative approach, fuzzy taxonomic structures and the approximate item set approach

2.5 Incremental Algorithms

In many applications, incremental updates (representing newer data) may arrive at any time. Some algorithms cannot incorporate incremental updates into existing grouping structures and, instead, have to re compute a new group from scratch. Incremental algorithms may also be sensitive to the input data order. That is, given a set of data objects, this algorithms may return dramatically different group depending on the order in which the objects are presented. Incremental algorithm that is insensitive to the input order are needed[32,37]. The proposed methodology view an incremental algorithm that integrates interestingness criteria during the process of building the model. One of the main drawbacks with the classical algorithms is that they do not consider the time in which the data arrived In such scenario a combination of old and new data is used to build a new model from scratch. This results in losing of the previously discovered knowledge (PDK).

Although these approaches require the user to provide constraints, the user background knowledge is not implicitly/explicitly stated. This lack of incorporating user Domain Knowledge (DK) and Previously Discovered Knowledge (PDK) into the tree induction process results in a decision tree which may be optimal in size and accuracy but may generate branches that are similar to the earlier discovered tree and hence does not reflect the user interest.

Commonly used techniques to discover interesting patterns in most KDD end are partially effective unless combined with subjective measures of interestingness. Subjective measures quantify interestingness based on the user understandability of the domain [44, 45]. Capturing the user subjectivity in dynamic environment requires a great deal of knowledge about databases, the application domain and the user's interests at a particular time . Therefore, it is difficult for the user to analyze the discovered patterns and to identify those patterns that are interesting from his/her point of view.

2.6 Interestingness Measures of Discovered Patterns

Data mining research has shown that we can measure a rule's interestingness using both objective and subjective measures [64, 65, 66]. Objective measures involve analyzing the rule's structure, predictive performance, and statistical significance. Such measures include accuracy, support and confidence. However, objective measures are insufficient for determining a discovered rule's interestingness. Subjective measures are needed [64, 65, 66]. Subjective interestingness has three main measures:

•Unexpectedness: rules are interesting if they are unknown to the user or contradict the user's existing knowledge (or expectations)

•Actionability: rules are interesting if users can do something with them to their advantage.

•Novelty: rules are novel if they add knowledge to the user prior knowledge.

Although novelty, actionability and unexpectedness of the discovered knowledge are the basis of the subjective measures, their theoretical treatment still remains a challenging task. Actionability is the key concept in most applications. Actionable rules let users do their jobs better by taking some specific actions in response to the discovered knowledge. Actionability, however, is an elusive concept because it is not feasible to know the space of all rules and the actions to be attached to them. Actionability is therefore is implicitly captured by novelty and unexpectedness [72] .

2.7 Challenges In Web Personalization

Recent advances in web personalization and recommendation system has brought the researchers together from all over the world, for resolving the challenges that are yet to be met[79]. This section presents the major issues that demand rigorous research in this area. The most important issue to be considered during user profiling is privacy violation. Many users are reluctant to provide their personal information either implicitly or explicitly (such as those obtained from registration forms). Aware users also hesitate to visit those web sites, which make use of cookies or agents. In such an environment, it becomes difficult to achieve personalization. In data pre-processing phase, the web log data may need to be cleaned from entries of pages that returned an error or graphics file accesses. For some domains, such information may be important whereas for other domains, same data should be eliminated from a log file. Thus overcoming the domain dependence problem is important so that a standard technique can be employed for all the domains. Another problem to be met has to do with web caching. Accesses to cached web pages are not recorded in the server log and hence such information is missed. Identification of a user from server log entries is also very important but this task becomes difficult in presence of a proxy server between web server and the host. A visitor having no previous interaction with the website poses a problem to the personalization system as there is no data available for personalizing the interactions with such a user. A similar problem occurs for a newly added item. Due to the absence of rating history, system cannot recommend a new item to users until its rating history has been collected. Many personalization systems use static profile of users. It should be remembered that user's interests are not static (fixed) and it can keep changing with respect to time, which supports research in dynamic profiling of users. Those personalization systems, which highly depend upon item ratings provided by users, are vulnerable to receive false ratings from users who have their vested interest in making an item popular among visitors[47].

Although research is already being conducted to resolve these issues since past few decades and as a result, researchers have come up with a range of solutions. But the growing size of data and rapidly increasing demand for personalization in various contexts is continuously posing new challenges to investigators.

2.8 Related Works

The foremost, module of the web personalization construction is log mining and processing [18]. Log analysis and web usage mining is a practice that enables it to apply useful statistics and data mining practices such as clustering, association rule discovery, classification, and sequential pattern retrieval to process information accumulated in the web server logs to disclose valuable patterns and it able to analyze later [21]. These patterns depend on the method used and the input facts and can be correlations between users page clusters, usage patterns, user groups and web pages. It can later store these patterns in a storage to perform query systems or OLAP work with visualization techniques.

In [22], the author proposed a framework for personalizing web mining based on web content along with web usage data and a site structure for more precisely predicting future requests from users. It suggested an algorithm known as modified IncSpan for efficient mining of chronological patterns in growing databases. This algorithm can detect database-based sequential patterns in sequential patterns, occur in the insert and append databases, and clogged sequential patterns come from result sequential patterns.

In [23], has provided a solution supported by the LCS algorithm to analyze and process user search patterns for subsequent web page predictions. Their architecture improves classification accuracy and provides efficient online forecasting. Some evaluation techniques applied to evaluate the quality of found predictions.

In [24], they study the effectiveness of the proposed algorithm and reduce database search time, They proposed a dual algorithm for Web-enabled mining based on sequential numbers that are appropriate for mining interval patterns. They used an algorithm that changed the user's interval pattern to binary. Then they used the inconsistent search strategy to generate frequent itemset candidates twice. They also calculate and maintain for the sequence number aspect to scan the user's interval pattern once, unlike the existing dual search mining algorithm. Their experiments indicate that the competence of the algorithm is faster and additional proficient than suggesting similar algorithms.

In [26], the authors proposed an efficient and new architecture for Web search personalization by Web-enabled mining without user's explicit feedback. The author performs web personalization utilizing the sequential browsing pattern mining algorithm and the Apriori algorithm. The Apriori algorithm provides frequent patterns and explosive candidate sequences but requires a lot of space for processing. The author used the FP-growth tree and the Markov model to reduce the limitations of this algorithm. It utilized prNG Graph to find user browsing patterns. Used FP-growth trees instead of prNG graphs to improve performance.

Many researches in the field of web mining and web personalization that performed in the past aim to retrieve patterns by means of data mining practices to identify the user's browsing activity to make expert decisions about reorganizing or modifying the site. In many cases, according to the browsing behavior of users recommended engine helps users to navigate the site, there are some sophisticated systems offer additional benefit, to consolidate the concept of adapting the site structure according to browsing behavior

In research works was performed to find periodic patterns in a time series database. Most of these tasks, however, focus on regular pattern mining instead of pertaining these patterns to realistic applications. They only converse a few applications for periodic accessibility patterns, identify periodic accessibility patterns., no approach has been for the periodic accessibility based web personalization.

A number of studies provide different ways to model user profiles, most of them do not deal with the user's changing interests over time, while others only attempted to model low-level dynamic user profiles[83]. The distinction made in these works between short and long-term interests, and interests at current runtime is often blurred and most such approaches treat users as having fixed size interests and the periodic nature of browsing patterns is repeatedly ignored. Actually, various valuable browsing patterns happen repeatedly during a specific time period, this can depict a user surfing habits, but not during other time periods. These browsing patterns are called "periodic browsing patterns".

The proposed approach dynamically learns the user's areas of interest different period like Early Morning, Morning, Noon, Early Afternoon, Late Afternoon, Evening, Night and Late Night. The propose an incremental approach that integrates interestingness criteria during the process of building the model. . One of the main features of this approach is to capture the user background knowledge, which is monotonically augmented. The proposed approach deals with time changing data and user beliefs.

chapter3

The Proposed Approach

3.1 Introduction

Web mining is a way of extracting patterns automatically in click streams and related data collected through a user's access to one or more websites during the web browsing. The main purpose of Web mining is to examine the behavioral patterns and user profiles that interact with a Web site. The detected patterns are identified as a set of pages, objects, or resources regularly accessed by groups of users with shared interests. We use the techniques to extracts the periodic access patterns frequently occurring in a particular period. The proposed approach abled to recognize resources used frequently at a specific time. In this chapter, we present the proposed approach in order to generating dynamic user profiles for web personalization system.

3.2 Configuration of Periodic User Profiles

A user profile can be either static, when the information it contains is never or rarely altered (e.g. demographic information), or dynamic when the user profile's data change frequently. Such information is obtained either explicitly, using on line registration forms and questionnaires resulting in static user profiles, or implicitly, by recording the navigational behavior and the preferences of each user [48,49], resulting in dynamic user profiles.

The system should be able to differentiate between different users or groups of users. This process is called user profiling and its objective is the creation of an information base that contains the preferences, characteristics and activities of the users.

In this thesis, a user profile is built dynamically for web personalization systems through analyze the usage patterns from the web usage log. The web usage log based on the browsing interval and provide appropriate resources based according to the user's behavior as shown in Figure 3.1.

There are different steps to manage incoming data streams for representing the dynamic behaviors of the users. Dynamic user profiles save the changes in a user's interests or behavior and make recommendations to users based on their behavior. These recommendations change as user behavior changes.

In Figure3.1, the data is stored in log file to perform pre-processing such as : data cleaning, user identification, and interval identification. Then, the patterns are extracted in order to create periodic web usage rules, which are subjected to novelty measure. The novelty measure is used in order to extract only interesting periodic user profile. Subsequently, these patterns are used to update incremental Model Mi To be used in other time period . The following subsections describe the proposed approach:

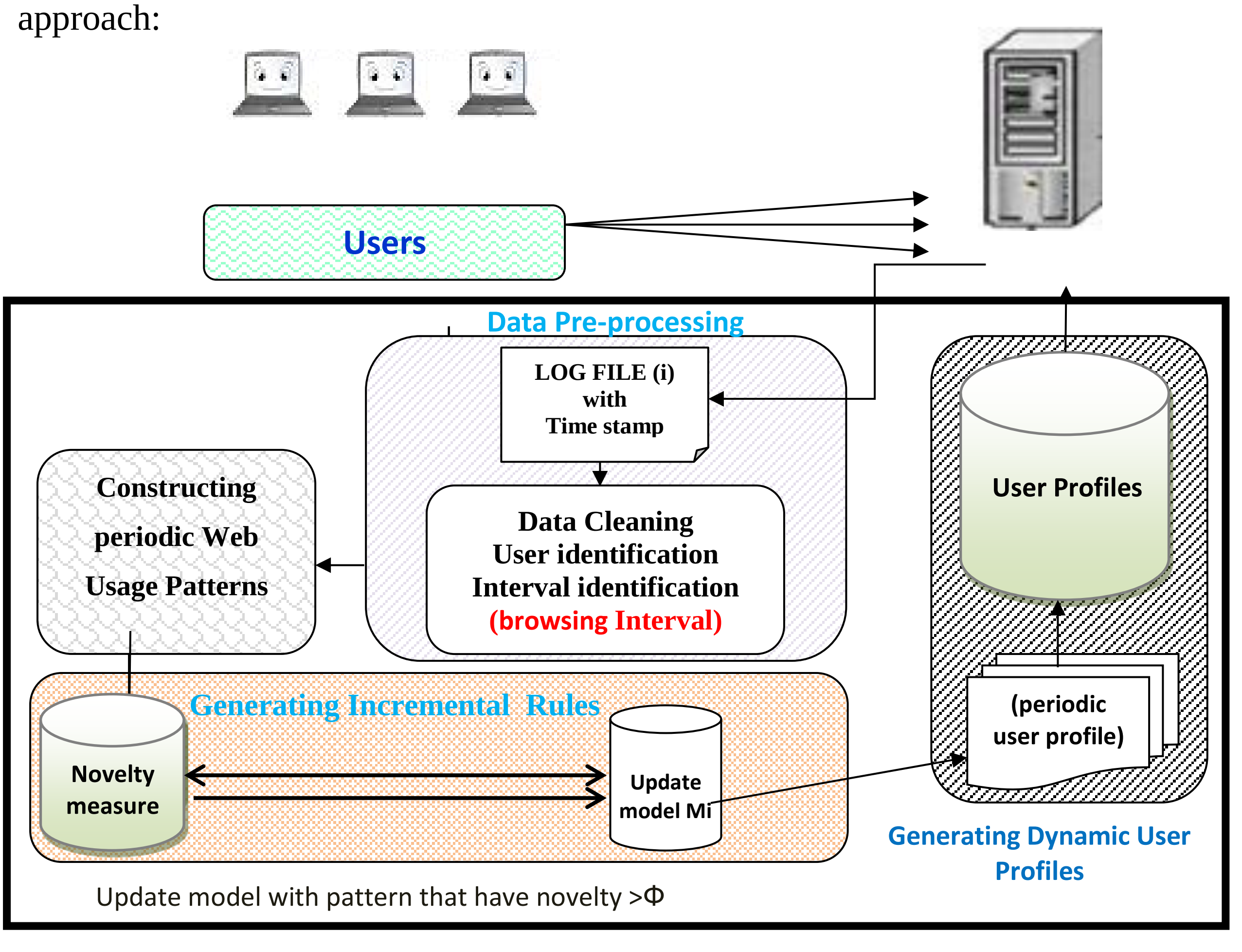


Figure 3.1 :Structure for configuration of Periodic User Profiles

The configure of "User Activity" is configured for the individual web usage pattern model. When a period condition is particular to an individual user or a web server, an individual periodic user profile depends on an individual web usage pattern.

The web server afterward provisions the configuration and presents the resource to the user for a specified period of time, the proposed approach consists of three Phases, Data Preparation for browsing activity, constructing periodic individual web usage pattern, and constructing incremental rules to generating periodic user profile .

3.3 Data Preparation

Data preparation is responsible for processing the web usage log and identifying all user access sessions individually , and split the entire time into set of browsing intervals . There is semantic information in the URL of the web usage log that consist of the user's accessed content . This study assumes that each requested URL in the Web usage log contains semantic information that is interpreted from one or more predefined category , such as News, Sports, or Entertainment. This task can be done either manually or semi-automatically [69] . This stage contains: Browsing Interval, constructing web usage context and preprocessing of data

3.3.1 Browsing Interval

Consider E be a set of unique access events, which represent web resources accessed by the user, i.e. URL_s of web pages. A user access session $S=e_1 e_2, e_3 \ldots e_n$ ($e_i \in E$ for $1 < i < n$) is a sequence of access events.

Each $e_i = (ts_i, te_i, URl_i)$, where t_{si} is the start time of event e_i , t_{ei} is the end time of event e_i , and URl_i is the URL accessed by the user in the event e_i . Note that it is not necessary that $URl_i \neq URl_j$ for $i \neq j$ in S . The URl_s in web usage logs contain little semantic information about the web contents accessed by users. To overcome this problem, each URL should be mapped into a predefined category such as: News, Sports and any other events . The category information can be obtained by using a web page category classification technique [54]. As such, each user access session can be classified into a sequence of categories. Suppose that the set of event attributes M_c consists of all the valid predefined categories, the function Category is defined by the $(e_i) = M_c$ $(m_c \in M_c)$ in order to define converts each e_i into Event Attribute m_c in M_c. The total duration for each category accessed estimates the level of user interest in that category for each access session by the user. Therefore the duration ($t_{e\ i}$- $t_{s\ i}$) of an access event indicate the level of user interest in that web content., After classifying the categories and computing the duration, each access session S is defined into S*=(t_s , t_e , D) where Ts is the start time of the session (i.e. t_{si} in S), T_e is the end time of the session (i.e. t_e in S) and D the total duration [1,2,3] :

$$D =\{d (S,m_i)|d(S,M_i)= \sum_{e_i \in S \text{ Category (ei)} = m_i} (te_i - ts_i), m_i \in M_c , 1 < i <|M_c|\}.$$

The browsing interval(S) = {(URL_1, t_1), (URL_2, t_2), ... , (URL_n, t_n)}" is the repeated requested URL_i with a timestamp for each $t_i(1 \leq i \leq n)$". The interval d_i of $URL_i$is basically as calculated di = $(t_{i+1} - t_i)$. For URL_n that does not have t_{n+1}, it's time interval will be computed by standard recent interval.

i.e. $d_n = (d_1+d_2+ . . . + d_{n-1}) / (n - 1) = (t_n - t_1) / (n - 1)$. To recognize the d_n the correlated interval having n >1 should be maintain, as it contains one or more requested URLs. Every URL_i in the user access session is related the setup of the access event attributes $M_{ci} \in M_C$ is for denoting the semantics of the content in a specific URL . so it's supposed that each URL_i in the duration of session d (S) is correlated with the access event feature collected as $M_{ci} \in M_C$, that represents URL_i contenets . In this consideration, each browsing interval can be defined as a sequence of group access event features as M_{ci}, and denoted as S={(M_{c1}, t_1 , d_1), (M_{c2}, t_2 , d_2,)), . . . , (M_c, t_n , d_n)}. The estimation of the interest level depends on the total duration for each access event attribute M_k during the user access session M_C. According to the above mentioned we can conclude the flowing: the level of user interest for specific access session can be estimated through the summation of the periods of the sub sequent URLs requests during specific session by the user. As a result the user interest can be estimated for entire browsing interval by converting the session time to a value from 0 to 24 as a successive time interval that is start time as "t_s (S) ϵ [0, 24]" and the end time "t_e (S) ϵ [0, 24]". We called the period of the browsing interval p(S) it is indicated as [55],

$$p(S) = \begin{cases} [t_s(S), t_e(S)], & \text{if } t_s(S) \le t_e(S) \\ [0, t_e(S)] \cup [t_s(S), 24], & \text{otherwise} \end{cases} \quad (3.1)$$

3.3.2 Constructing Web Usage Context

Temporal attributes for Web activities are referred to by eight real-life time concepts, namely Early Morning, Morning, Noon, Early Afternoon, Late Afternoon, Evening and Night. For events attributes referred to by News, Sports, others. So the user access behavior can be represented by a set of temporal and events attributes.

In addition, we also use fuzzy logic [53] to represent both temporal and events attributes, and incorporated into Formal Concept Analysis [52] for constructing Web Usage Context. While fuzzy temporal based Web Usage Context can be represented by a cross table with rows indicating user session IDs and columns indicating the temporal and events attributes , which can be assigned by experts or computed automatically. A membership value μR (g, m) ∈ [0, 1] in row g and column m indicates a fuzzy relation between the object g and attribute m. The relation between a user access session, and a temporal attribute or an event attribute is represented by a membership value within [0, 1] , Table 3.1: Web Usage Context

SID	T1 (A)	T2 (E)	T3 (N)	C1 (S)	C2 (N)	C3 (T)
S1	0.6	0.5	-	0.8	-	0.6
S2	-	0.4	0.7	-	-	0.9

S3	-	-	0.9	-	0.7	0.5

Table 3.1 an example shows the Web Usage Context, which consists of three user access sessions, three temporal attributes “T1 (A - Late Afternoon)”, “T2 (E - Evening)” and “T3 (N - Night)”, and three events attributes “C1 (S - Sports)”, “C2 (N - News)” and “C3 (T - tec)” [67].

3.3.3 Data pre-processing

Pre-processing consists of cleaning data, user identification , and browsing interval discovery . The main idea of the pre-processing process is to eliminate unsuccessful requests, redundant data and to identify all the personal browsing interval for each individual user [41]. Further more, the proposed approach needs to process the Web usage log sequentially for all intervals individually browsed by the individual user. So here we are using the traditional way of pre-processing Web server logs, namely data cleaning, user identification and interval identification .

- Data Cleaning

In data cleaning we depend on resource properties, entries are considered useful in Web usage logs when one of the resource properties is mentioned in the requested URLs, unless unwanted records in the Semantic Annotation field will be removed .

- User identification:

User identification in data preparation is important step in web personalization system , once the web log files become cleaned next step is user identification , So consider each combination of IP and user-agent unique as separate user. It analyzes individual browsing activity, the unique user should be recognizing the unique "UserIP" value. So that, the website could be personalized and the methods should have to be capable in differentiate the dissimilar individual or users groups.

- Interval identification

It defines the time stamp gap for each two repeated requests in an interval by the same user , assuming that the latest surfed interval has started for constant 30 minutes session limit for the timeout. The session is defined as a sequence of requests made by a single user within a certain navigation period the user can have a single or multiple sessions during a period of time, For that session identification objective is to separate the accesses pages into individual sessions for each user . Reconstruction of precise user sessions from server access logs is a difficult task because the access log protocol (HTTP protocol) is **status less and connectionless.** There are two simple methods for session identification[80]. One is based on the total session time ranging from 25.5 minutes to 24 hours[81]. and other is based on a single page stay time . The set of pages visited by a specific user at a specific time is the "page viewing time". Depending on the length of stay in a page calculated by the difference between time stamps[82].

3.4 Creating Periodic Web Usage Pattern

The relevant functions in this study are described as follows[71]:

$$z(g_i, m_c) = \frac{d(g_i, m_c)}{T_{ei} - T_{si}} \quad , \; z(m_c) = \frac{\sum_{g_k \in G} d(g_k, m_c)}{\sum_{g_k \in G} d(Te_k, Ts_k)} \quad (3.2)$$

"$Z(M_c)$" is the ratio of the entire interval of browsing to the access event "m_c" in all browsing intervals of the user, which represents the user's overall interest in the access event "m_c". The "$z(g, m_c)$" is described as the ratio of how long to browsing access event "m_c" in browsing interval "g", which represents the user's confined interest in access event "m_c".

The method of fuzzy interval web usage : it is a cross table with rows and columns, where rows present browsing intervals while columns are period and access event features as shown in the table 3.1 . The elements of $\mu_R(g, m) \epsilon$ [0, 1] where $g \epsilon G$, $m \epsilon M_p \cup M_c$, is represented by a membership value $\mu_R(g, m)$ in [0, 1], of row "g" and column "m" have the fuzzy relationship among browsing interval "g" and feature " m". It can organize the individual web usage patterns supported by the fuzzy periodic web usage context. $Z(m_c)$ is defined as the proportion of the total duration of accessing a web category m_c in all user access sessions, which indicates the user's global interest of the web category m_c.

$z(g_i, m_c)$ is defined as the proportion of the duration of accessing a web category $m_c$within a user access session g_i, which indicates the user's local interest of the web category m_c . Then we use novelty measure of interestingness of discovered patterns are used as a constraint to discover temporal association rules that are novel and hence interesting in an incremental manner. The temporal rules that reflects the changing data and the user beliefs is attractive in order to make the overall KDD process more effective and efficient , the novelty measure used in temporal and incremental association rule mining in order to find not only accurate knowledge but also comprehensible and novel knowledge [67].

3.4.1 Association Rules

Web usage mining is the method of extracting interesting patterns from Web usage log file.Web usage mining is subfield of data mining uses various data mining techniques to produce association rules . There are many more techniques used to find association between Web pages. But instead of only page sequence if we consider page view time while accessing the web page then the Web log sequence can be seen as quantitative data. Fuzzy logic, which may be viewed as an extension of traditional logical systems, provides an effective conceptual framework for dealing with the problem of knowledge representation in an environment of uncertainty and imprecision [6, 7, 8, 9] Here first, we applied fuzzy concept to Web usage log data to find fuzzy labels and then applied A priori algorithm to find interesting association rules. We can recognize the individual web usage pattern through recognizing the activity of web browsing by the user and configure the individual we usage pattern during the individual browsing interval. Individual web usage pattern is the level of user interest during a time in browsing specific URL .

Which can be estimate by $d_i = (t_{i+1} - t_i)$. That means a particular user is interested in a particular resource for specific period of Time. To Constructing Web Usage Patterns here we use association rules approach to extracting the user interesting patterns for quantitative data instead of association rules between only page sequences

3.4.2 Fuzzy Association Rule

There are different types of association rules that is a previous type association of binary rules and represented by Boolean value. Actually, the data base does not include binary attributes only but it includes quantitative and categorical attributes. Among these types we cannot use the classical method for mining to discover rules in such types of data. Instead of using real numbers between zero and 1 we need a technique that deals with the quantitative attribute replaced by Boolean attributes [52].

The web usage data includes quantitative values, to deal with it we use the fuzzy data mining algorithm to extract the association from the quantitative log file to generate the fuzzy association rules. at the beginning we designed a membership function which is used to convert the quantitative values into fuzzy terms. These values should be divided into intervals and map each attribute into new corresponding attribute. We use the fuzzy web usage pattern in the browsing interval of user represent the features for both periodic and access events referred to $K=(G, M_p, M_c, I)$ where G denotes set of browsing intervals for a user, M_p is a set of periodic attributes, M_c is a set of access event attributes, I is fuzzy set in the domain of $(G \times (M_p \cup M_c))$, a fuzzy set of the domain to denote the associations to browsing intervals $\mu(g,m)$, to fuzzy where $g \in G$ and features $m \in (m_p \cup m_c)$ [54].

Each fuzzy association $\mu(g,m) \in I$ is symbolized by elements of $\mu(g, m) \in [0, 1]$, where,

$$\mu(g, m) = \begin{cases} \mu_p(g, m_p) & \text{if } m_p \in M_p \\ \mu_c(g, m_c) & \text{if } m_c \in M_c \end{cases} \qquad (3.3)$$

the user browsing behavior can be represented by a set of temporal and events attributes $\mu(g, m)$= a set of temporal attributes $\mu_p(g, m_p)$ and a set of events attributes $\mu_c(g, m_c)$, each browsing interval $g \in G$ can be computed using the period of (g) symbolized by p(g), Can also be represented as a fuzzy set on the domain for a periodic attribute $m_p \in M_p$ the membership value.

$\mu_p(g, m_p)= \max_{t \in p(g)} \{\mu_p(t,m_p)\}$,where $\mu_p(t, m_p)$is described relation between a user access session (t), and a temporal attribute m_p, This relation is illustrated in Fig 3.2

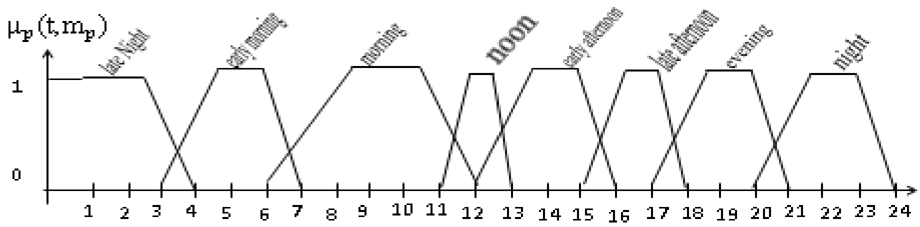


Figure 3.2: Periodic attributes

Fuzzy set on the domain for a events attribute , in the beginning, classifying the categories that we have mentioned earlier, then computing the duration of an access event (ei) can be used to indicate the level of interest the user has in that web content. and the total duration for each category accessed can be used for estimating the level of user interest in that category for each user access session. events attribute is represented by a membership value within [0, 1] Also for a resource attribute, the membership value $\mu_c(g, m_c)$, The entire interval for every $m_c \in M_C$ is used to estimate the user's point of interest in the resource for the period of the browsing interval S, which can be calculated as follows $d(g, m_c)$ [55].

3.4.3 Incremental Mining of Association Rules to Generate Dynamic User Profile

The proposed approach uses fuzzy Apriori algorithm for dynamic mining of association rules requires the following steps: frequent itemset are generated using support measure: This is which required to recognizing the web browsing activity for the user also configuring the individual web usage pattern in its individual browsing interval. Then compute a confidence for rules through the time T, subsequently, generate strong association rule ,

We compute the novelty measure (NM) by as in the following equation [60,62,63,64]

$$\mathrm{NM} = \frac{\{|S_1| + |S_2| - 2 * K\} + \sum_{i=1}^{k} \delta(c_1^i, c_2^i)}{|S_1| + |S_2|} \qquad (3.4)$$

Where,

S_1 and S_1 be two conjunct sets with cardinalities $|S_1|$ and $|S_2|$ respectively.

K= the pairs of compatible conjuncts between S_1 and S_2

(c_1^i, c_2^i) is the i^{th} pair of compatible conjuncts [60,62,63,64].

The computation of (NM) of the periodic rules (P) with respect to the existing model M is performed to discover only Interesting rules and update the model incrementally. The advantage of pushing such measure inside the algorithm is that the search space is reduced and the algorithm discovers relatively smaller sized model. Construction of periodic user profile in the personalization system is the construction of the periodic personalization resources through understanding user interests within browsing intervals, the discovered knowledge reflects the user's requirement of interestingness.

The user interest may change over time and for that reason, it is suggested to generate dynamic user profile for tracking these changes in the user behaviors and to capture these characteristics. The Figure 3.3 shows the algorithm that extract the periodic interesting patterns .

Figure 3.3: fuzzy Apriori algorithm for extracting periodic interesting patterns

Input: Pre-processed dataset
Output: Set of periodic interesting patterns
Variables :

C_K: Candidate itemset of size k
L_K: frequent itemset of size k
Φ : Novelty measure threshold

Process :

```
    Get initial membership functions, support value and confidence value
    Divide dataset into partitions
    Set n to number of partitions to be processed
           repeat Transfer quantitative values into fuzzy terms with fuzzy values
          Calculate the counts of fuzzy terms
     repeat
          for each fuzzy term
               Generate the candidate set by counting of each fuzzy term incrementally
          end for
          for each fuzzy term
                if count ≥ min support
                    generate large frequent
                end if
          end for
   join the large Itemsets
   until large Itemsets  = = NULL
   until n=0
    repeat
    for each large Itemsets
          if confidence ≥ min confidence
                    Construct association rule
          end if
    end for
    merge all association rules
until n=0
Extract periodic interesting patterns
return ∐k LK;
```

3.5 A Detailed example

For

date	timestamp	session ID	file name
2022-01-01	05:39:56	1	Diseases.asp
2022-01-01	05:40:26	1	Dance.html
2022-01-01	05:42:46	1	Basketball.htm

Example: Given the following logfile dataset is shown in Table (3.2).

2022-01-01	05:44:17	1	Diseases.asp
2022-01-01	05:40:52	1	newspaper.html
2022-01-01	05:41:54	2	Diseases.asp
2022-01-01	05:42:25	2	Basketball.htm
2022-01-01	05:44:07	2	stockmarket.html
2022-01-01	05:41:30	2	Diseases.asp
2022-01-01	05:43:02	3	Basketball.htm
2022-01-01	05:44:31	3	newspaper.html
2022-01-01	05:43:46	3	Dance.html
2022-01-01	05:44:06	4	Technology
2022-01-01	05:45:47	4	newspaper.html
2022-01-01	05:47:45	4	Dance.html
2022-01-01	05:47:56	4	Basketball.htm
2022-01-01	05:48:38	4	Diseases.asp
2022-01-01	05:46:46	4	newspaper.html
2022-01-01	05:47:53	5	Dance.html
2022-01-01	05:48:19	5	Basketball.htm
2022-01-01	05:49:33	5	Diseases.asp
2022-01-01	05:48:50	5	Dance.html
2022-01-01	05:50:13	6	Technology
2022-01-01	05:51:14	6	newspaper.html
2022-01-01	05:53:16	6	Basketball.htm
2022-01-01	13:44:31	7	newspaper.html
2022-01-01	13:43:46	7	Dance.html
2022-01-01	13:44:06	7	Technology
2022-01-01	13:45:47	8	newspaper.html
2022-01-01	13:47:45	8	Dance.html
2022-01-01	13:47:56	8	Basketball.htm
2022-01-01	13:48:38	9	Diseases.asp
2022-01-01	13:46:46	9	newspaper.html
2022-01-01	13:47:53	10	Dance.html
2022-01-01	13:48:19	10	Basketball.htm
2022-01-01	13:49:33	11	Diseases.asp
2022-01-01	13:44:31	11	newspaper.html

Table(3.2): Log file

Data Preparation

This stage responsible for processing the web usage log and identifying all user access sessions individually as shown in table(3.3), and split the entire time into set of browsing intervals. There is semantic information in the URL of the web usage log that consist of the user's accessed content in order to convert each URL into Event Attribute mc in Mc. The following steps are performed are data preparation stage:

1- identifying all user access sessions individually:

Session ID	date	timestamp	file name
session1	2022-01-01	05:39:56	Diseases.asp
	2022-01-01	05:40:26	Dance.html
	2022-01-01	05:42:46	Basketball.htm
	2022-01-01	05:44:17	Diseases.asp
	2022-01-01	05:40:52	newspaper.htm
Session2	2022-01-01	05:41:54	Diseases.asp
	2022-01-01	05:42:25	Basketball.htm
	2022-01-01	05:44:07	stockmarket.ht
	2022-01-01	05:41:30	Diseases.asp
Session3	2022-01-01	05:43:02	Basketball.htm
	2022-01-01	05:44:31	newspaper.htm
	2022-01-01	05:43:46	Dance.html
Session4	2022-01-01	05:44:06	Technology
	2022-01-01	05:45:47	newspaper.htm
	2022-01-01	05:47:45	Dance.html
	2022-01-01	05:47:56	Basketball.htm
	2022-01-01	05:48:38	Diseases.asp
	2022-01-01	05:46:46	newspaper.htm
Session5	2022-01-01	05:47:53	Dance.html
	2022-01-01	05:48:19	Basketball.htm
	2022-01-01	05:49:33	Diseases.asp
	2022-01-01	05:48:50	Dance.html
Session6	2022-01-01	05:50:13	Technology
	2022-01-01	05:51:14	newspaper.htm
	2022-01-01	05:53:16	Basketball.htm
Session7	2022-01-01	13:44:31	newspaper.htm
	2022-01-01	13:43:46	Dance.html
	2022-01-01	13:44:06	Technology
Session8	2022-01-01	13:45:47	newspaper.htm
	2022-01-01	13:47:45	Dance.html
	2022-01-01	13:47:56	Basketball.htm
Session9	2022-01-01	13:48:38	Diseases.asp
	2022-01-01	13:46:46	newspaper.htm
Session10	2022-01-01	13:47:53	Dance.html
	2022-01-01	13:48:19	Basketball.htm
Session11	2022-01-01	13:49:33	Diseases.asp
	2022-01-01	13:44:31	newspaper.htm

Table(3.3): identifying all user access sessions individually

2-

Time	Session ID	date	timestamp	file name
Time t1 Early	session1	2022-01-01	05:39:56	Diseases.asp
		2022-01-01	05:40:26	Dance.html
		2022-01-01	05:42:46	Basketball.ht

Splitting the entire time into set of browsing intervals as shown in table (3.4):

Morning		2022-01-01	05:44:17	Diseases.asp
		2022-01-01	05:40:52	newspaper.ht
	Session2	2022-01-01	05:41:54	Diseases.asp
		2022-01-01	05:42:25	Basketball.ht
		2022-01-01	05:44:07	stockmarket.h
		2022-01-01	05:41:30	Diseases.asp
	Session3	2022-01-01	05:43:02	Basketball.ht
		2022-01-01	05:44:31	newspaper.ht
		2022-01-01	05:43:46	Dance.html
	Session4	2022-01-01	05:44:06	Technology
		2022-01-01	05:45:47	newspaper.ht
		2022-01-01	05:47:45	Dance.html
		2022-01-01	05:47:56	Basketball.ht
		2022-01-01	05:48:38	Diseases.asp
		2022-01-01	05:46:46	newspaper.ht
	Session5	2022-01-01	05:47:53	Dance.html
		2022-01-01	05:48:19	Basketball.ht
		2022-01-01	05:49:33	Diseases.asp
		2022-01-01	05:48:50	Dance.html
	Session6	2022-01-01	05:50:13	Technology
		2022-01-01	05:51:14	newspaper.ht
		2022-01-01	05:53:16	Basketball.ht
Time t2 Early Afternoon	Session1	2022-01-01	13:44:31	newspaper.ht
		2022-01-01	13:43:46	Dance.html
		2022-01-01	13:44:06	Technology
	Session2	2022-01-01	13:45:47	newspaper.ht
		2022-01-01	13:47:45	Dance.html
		2022-01-01	13:47:56	Basketball.ht
	Session3	2022-01-01	13:48:38	Diseases.asp
		2022-01-01	13:46:46	newspaper.ht
	Session4	2022-01-01	13:47:53	Dance.html
		2022-01-01	13:48:19	Basketball.ht
	Session5	2022-01-01	13:49:33	Diseases.asp
		2022-01-01	13:44:31	newspaper.ht

Table(3.4): Splitting the entire time into set of browsing intervals

3- Converts each URL into Event Attribute m_c in M_c, where mc and Mc are used to represent web pages. Let Mc1, Mc2, Mc3, Mc4 and Mc5 represent Diseases.asp, Dance.html, Basketball.htm, newspaper.html and stockmarket.html, as shown in

Time	Session ID	date	timestamp	file name	categories
Time t1 Early Morning	session1	2022-01-01	05:39:56	Diseases.asp	Health
		2022-01-01	05:40:26	Dance.html	Music
		2022-01-01	05:42:46	Basketball.ht	Sport
		2022-01-01	05:44:17	Diseases.asp	Health
		2022-01-01	05:40:52	newspaper.ht	News
	Session2	2022-01-01	05:41:54	Diseases.asp	Health
		2022-01-01	05:42:25	Basketball.ht	Sport
		2022-01-01	05:44:07	stockmarket.h	economy
		2022-01-01	05:41:30	Diseases.asp	Health
	Session3	2022-01-01	05:43:02	Basketball.ht	Sport
		2022-01-01	05:44:31	newspaper.ht	News
		2022-01-01	05:43:46	Dance.html	Music
	Session4	2022-01-01	05:44:06	Technology	Technology
		2022-01-01	05:45:47	newspaper.ht	News
		2022-01-01	05:47:45	Dance.html	Music
		2022-01-01	05:47:56	Basketball.ht	Sport
		2022-01-01	05:48:38	Diseases.asp	Health
		2022-01-01	05:46:46	newspaper.ht	News
	Session5	2022-01-01	05:47:53	Dance.html	Music
		2022-01-01	05:48:19	Basketball.ht	Sport
		2022-01-01	05:49:33	Diseases.asp	Health
		2022-01-01	05:48:50	Dance.html	Music
	Session6	2022-01-01	05:50:13	Technology	Technology
		2022-01-01	05:51:14	newspaper.ht	News
		2022-01-01	05:53:16	Basketball.ht	Sport
Time t2 Early Afternoon	Session1	2022-01-01	13:44:31	newspaper.ht	Health
		2022-01-01	13:43:46	Dance.html	Music
		2022-01-01	13:44:06	Technology	Sport
	Session2	2022-01-01	13:45:47	newspaper.ht	Health
		2022-01-01	13:47:45	Dance.html	News
		2022-01-01	13:47:56	Basketball.ht	Health
	Session3	2022-01-01	13:48:38	Diseases.asp	Sport
		2022-01-01	13:46:46	newspaper.ht	economy
	Session4	2022-01-01	13:47:53	Dance.html	Health
		2022-01-01	13:48:19	Basketball.ht	Sport
	Session5	2022-01-01	13:49:33	Diseases.asp	News
		2022-01-01	13:44:31	newspaper.ht	Music

table(3.5) .

Table(3.5): semantic information

4- Browsing Interval

In this step, the browsing interval(S) = {(URL_1, t_1), (URL_2, t_2), . . . ,(URL_n, t_n)}" is obtained. It represents the URLi with its timestamp.

In our example for the browsing interval:

S = {(Mc1, 05:39:56), (MC2, 05:40:26),(MC3 , 05:42:18),(Mc4, 05:43:56). . , (URLn, tn), where each browsing interval can be defined as a sequence of group access event features as Mci, and denoted as S={(Mc1, t1 , d1), (Mc2, t2 , d2,)), . . . , (Mc, tn , dn)}. The estimation of the interest level depends on the total duration for each access event attribute Mk during the user access session MC as shown in Table (3.6).

Session ID	Web Recourse
1	Mc1, 30
1	Mc2, 42
1	Mc3, 98
1	Mc4, 91
2	Mc3, 62
2	Mc1, 31
2	Mc3, 102
3	Mc5, 92
3	Mc3, 89
4	Mc1, 20
4	Mc4, 101
4	Mc2, 118
4	Mc1, 11
4	Mc4, 42
5	Mc3, 64
5	Mc1, 29
5	Mc4, 74
6	Mc3, 80
6	Mc4, 61
6	Mc2, 122
6	Mc1, 17

Table(3.6):The durations of all pages browsed by each client

S = {(Mc1, 05:39:56, 30), (MC2, 05:40:26, 42),(MC3 , 05:42:18, 98),(Mc4, 05:43:56, 91). . , (URLn, tn)

each browsing interval can be considered as a sequence of a group resource features as "M_{ci}", and denoted S = {(Mc1, 01:39:56 , 30), (MC2, 01:40:26 , 42) ,(MC3 , 01:41:08 ,98),(Mc4, 01:42 :46 ,91) . . . , (URLn, tn , dn)

The interval di of URLi calculated basically as shown in table (3.7) by di=(ti+1−ti)

d1 = 0.30-0.00=0.30 (Mc1 , 30)

d2 = 01:12- 0.30 =0. 42 (Mc1 , 42)

TID	Browsing Sequence
1	(Mc1, 30) (Mc2, 42) (Mc3, 98) (Mc4, 91)
2	(Mc3, 62) (Mc1, 31) (Mc3, 102)
3	(Mc5, 92) (Mc3, 89)
4	(Mc1, 20) (Mc4, 101) (Mc2, 118) (Mc1, 11)(Mc4, 42)
5	(Mc3, 64) (Mc1, 29) (Mc4, 74)
6	(Mc3, 80)(Mc4, 61) (Mc2, 122)(Mc1, 17)

Table (3.7): The durations of all pages browsed by each session

5- Constructing Web Usage Context

This step is one of the stages of data preparation. At this stage we convert the quantitative values into weights, using the fuzzy technique, in order to extract the user's interest in each interval.

To get the user's interest in a particular resource during a certain period. The data set in Table (3.7) is converted into fuzzy set in order to represent the level of user interest as shown in Table (3.8)

SID	Fuzzy set
1	$\left(\frac{0.8}{\text{mc 1. Short}}+\frac{0.2}{\text{mc 1. Middle}}\right),\left(\frac{0.6}{\text{mc 2. Short}}+\frac{0.4}{\text{mc 2. Middle}}\right),\left(\frac{0.6}{\text{mc 3. Middle}}+\frac{0.4}{\text{mc 1. Long}}\right),\left(\frac{0.8}{\text{mc 4. Middle}}+\frac{0.2}{\text{mc 4. Long}}\right)$
2	$\left(\frac{0.2}{\text{mc 3.Short}}+\frac{0.8}{\text{mc 3.Middle}}\right),\left(\frac{0.8}{\text{mc 1.Short}}+\frac{0.2}{\text{mc 1.Middle}}\right),\left(\frac{0.6}{\text{mc 3.Middle}}+\frac{0.4}{\text{mc 3.Long}}\right)$
3	$\left(\frac{0.8}{\text{mc 5.Middle}}+\frac{0.2}{\text{mc 5.Long}}\right),\left(\frac{0.6}{\text{mc 3.Middle}}+\frac{0.4}{\text{mc 3.Long}}\right)$
4	$\left(\frac{1.0}{\text{mc 1.Short}}\right),\left(\frac{0.6}{\text{mc 4.Middle}}+\frac{0.4}{\text{mc 4.long}}\right),\left(\frac{0.2}{\text{mc 2.Middle}}+\frac{0.8}{\text{mc 2.long}}\right),\left(\frac{1.0}{\text{mc 1.Short}}\right),\left(\frac{0.6}{\text{mc 4.Short}}+\frac{0.4}{\text{mc 4.Middle}}\right)$
5	$\left(\frac{1.0}{\text{mc 3.Middle}}\right),\left(\frac{0.8}{\text{mc 1.Short}}+\frac{0.2}{\text{mc 1.Middle}}\right),\left(\frac{1.0}{\text{mc 4.Middle}}\right)$
6	$\left(\frac{1.0}{\text{mc 3.Middle}}\right),\left(\frac{0.2}{\text{mc 4.Short}}+\frac{0.8}{\text{mc 4.Middle}}\right),\left(\frac{0.2}{\text{mc 2.Middle}}+\frac{0.8}{\text{mc 2.Long}}\right),\left(\frac{1.0}{\text{mc 1.Short}}\right)$

Table (3.8)fuzzy set

In fact, an individual user in a browsing session can browse more than one source, and each source has a period of time, and this represents the user's behavior during this interval. This can be done through Table (3.9), this table contains rows and columns. The intersection of the row with the column gives us different weights for each source reached. An this stage, constructing Web Usage Context by a cross table with columns indicating user session IDs and rows indicating the temporal and events attributes

SessionID	Time t1 early Morning			Time t2 early after noon		
	S 1	**S2**	**S3**	**S 1**	**S2**	**S3**
Mc1.Short	0.8	0.8	0.0	1.0	0.8	1.0
Mc1. Middle	0.2	0.2	0.0	0.0	0.2	0.0
Mc1.Long	0.0	0.0	0.0	0.0	0.0	0.0
Mc2.Short	0.6	0.0	0.0	0.0	0.0	0.0
Mc2. Middle	0.4	0.0	0.0	0.2	0.0	0.2
Mc2.Long	0.0	0.0	0.0	0.8	0.0	0.8
Mc3. Short	0.0	0.2	0.0	0.0	0.0	0.0
Mc3. Middle	0.6	0.8	0.6	0.0	1.0	1.0
Mc3.Long	0.4	0.4	0.4	0.0	0.0	0.0
Mc4. Short	0.0	0.0	0.0	0.6	0.0	0.2
Mc4. Middle	0.8	0.0	0.0	0.6	1.0	0.8
Mc4.Long	0.2	0.0	0.0	0.4	0.0	0.0
Mc5. Short	0.0	0.0	570.0	0.0	0.0	0.0

Mc5. Middle	0.0	0.0	0.8	0.0	0.0	0.0
Mc5.Long	0.0	0.0	0.2	0.0	0.0	0.0

Table(3.9): constructing Web Usage Context

6- Creating Periodic Web Usage Pattern

In this step, all the weights for each resource are collected in all browsing [illegible] regions for recourse as shown in Table (3.10). The higher value represents the user's interest in the current browsing period. The "z(g, mc)" is described as the ratio of how long to browsing access event "mc" in browsing interval "g", which represents the user's interest in access event "mc".

COUNT z(g,Mc)	fuzzy regions
4.4	Mc 1.Short
1.6	Mc2. Long
4.2	Mc 3.Middle
3.2	Mc 4.Middle
0.8	Mc 5.Middle

Table (3.10): largest count value among all possible regions for recourse

Importance of web page is divided into five fuzzy regions as follows: Very Unimportant, Unimportant, Ordinary, Important, and Very Important. The linguistic terms for the importance of the web pages are transformed into fuzzy sets by the membership functions. Then the minimum support value of the linguistic is converted which has been calculated, to a fuzzy set of minimum supports, in order to we calculate fuzzy weighted set of minimum supports. Then calculated the average height of the fuzzy weighted support, through dividing by the number of three durations for each web resource, e.g. long, short, middle.

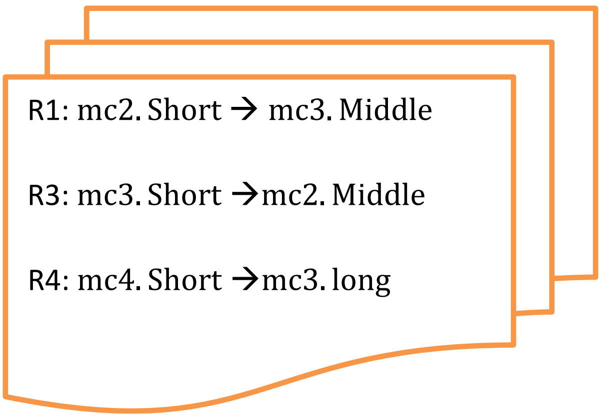
As a result, we will have the set of fuzzy weighted large 1-sequences as shown in Table (3.11)

1-SEQUENCE	COUNT
mc1. Short	4.4
mc4. Middle	3.2
mc3. Middle	4.2
mc2. Long	1.6

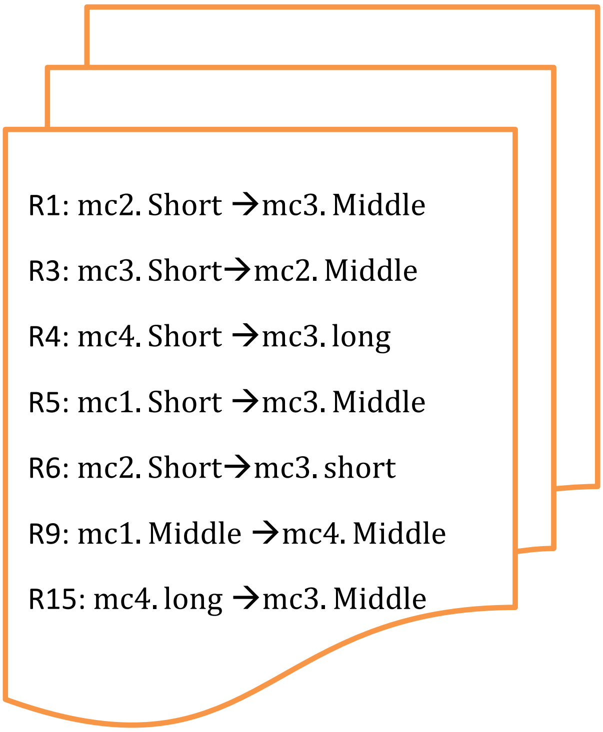
Table (3.11):The set of fuzzy weighted large 1-sequences

7- Incremental Mining of Association Rules to Generate Dynamic User Profile

In this phase, association patterns are extracted from the frequent itemsets. These patterns are evaluated using confidence measure and prune the patterns that do not satisfy this criterion resulting in a set of strong association patterns which are subjected to the novelty criterion with the aim of deciding either these patterns are interesting or not. This phase takes into consideration the existing model Mi representing the known association rules and consequently resulting in discovering of Mi+1. For each frequent itemsets, only novel rules are extracted and used to update the model Mi+1. We compute novelty degree rule with the novelty measure (NM).



Model in Time t1



Model in Time t2

Chapter 4

Implementation and Experimentation

4.1 Introduction

This chapter describes the experiments to evaluate performance of the proposed approach. In this section, experimental results are presented, in particular those related to the proposed approach performance. The proposed approach and other algorithms are written in Java and implemented on Hadoop. The experiments are conducted using Anonymous, Web Data Set available at http://kdd.ics.uci.edu. The datasets are considered as evolving with time.

4.2 Evaluation Measures

Real-time web personalization is a costly approach. Periodic personalization satisfies users more effectively. Periodic browsing patterns allow it to easily infer resources that the users might be most interested in over a period of time.

We have defined eight actual Intervals as, Early Morning, Morning, Noon, Early Afternoon, Late Afternoon, Evening, Night and Late Night. To indicate the temporal nature of the web activity, and effect of novelty threshold Φ on the number of rules. The dataset contains a 48 -hour period of Hypertext Transfer Protocol (HTTP requests to a Web server).

In these experiments, we built dynamic user profiles by creating individual web usage pattern to generate periodic web usage pattern with novelty measure .

4.2.1 Experiment One

We created the data by sampling and processing the log file. The data records, the use of by 38000 anonymous, randomly selected users. For each user, the data lists all the web sites, that user visited in a one week timeframe. Users are identified only by a sequential number, for example, User #14988, user #14989, etc. The file contains personally identifiable information. And we assumed that they have arrived at times T1, T2

Intervals		The extracted rules with Novelty threshold Φ=0.5
Time T1	Early Morning	2798
	Morning	2645
	Noon	2573
	Early Afternoon	2077
	Late Afternoon	1562
	Evening	1329
	Night	1082
Time T2	Early Morning	852
	Morning	599
	Noon	363
	Early Afternoon	308
	Late Afternoon	287
	Evening	131
	Night	84

respectively.

Table 4.1: The discovered rules on Anonymous dataset using Apriori algorithm

With Novelty measure on T1; T2; in the proposed Approach

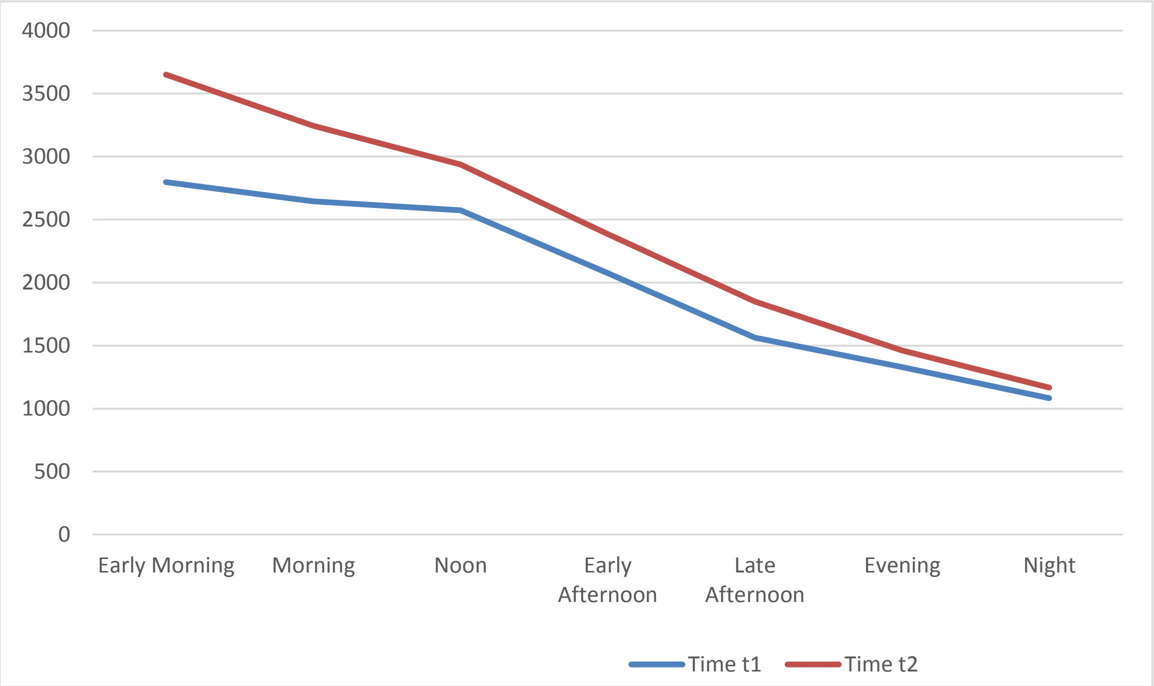


Figure 4.1: Discovered Rules and Novel Rules for Data Set at Time T1 and Time T2 for different Browsing Intervals and Novel Threshold=0.5

The performance is measured in term of the number of discovered rules with various thresholds of minimum support and confidence and fixed novelty threshold $\Phi = 0.50$. The dataset used is (the log file) and it is considered evolving with time and partitioned into two parts representing times T1 ; T2 ; respectively. Note that as

Table 4.1, and Figure 4.1 the reduction of the number of discovered rules using (log file) dataset at T1,and T2 times.

4.2.2 Experiment two

The second experiment was performed using public dataset to study the effect of novelty measure on the number of rules. The dataset contains a 48 -hour period of Hypertext Transfer Protocol (HTTP requests to a Web server) . The dataset has HTTP request from 05:39:56 - 1th March 2018 to 23:53:07 2th March 2018.

The dataset has total 748 requests: 514 GET requests, 22 POST requests, 17 HEAD requests and 6 invalid requests. Shown the result as Table(4.1) the proposed approach is implemented in C language and tested using public dataset available at

http://kdd.ics.uci.edu.

Intervals		Discovered Rules without Novelty	Discovered Rules with Novelty $\emptyset = 0.50$
T1	Morning	300	90
	Noon	267	45
	Afternoon	376	19
T2	Morning	233	40
	Noon	159	36
	Afternoon	188	31

Table 4.2: the resualt of experiment Two

The objective of the second experiment is to show the effectiveness of our a proposed approach in reducing the number of discovered rules against other appr oac h with out (N M). As sho wn in

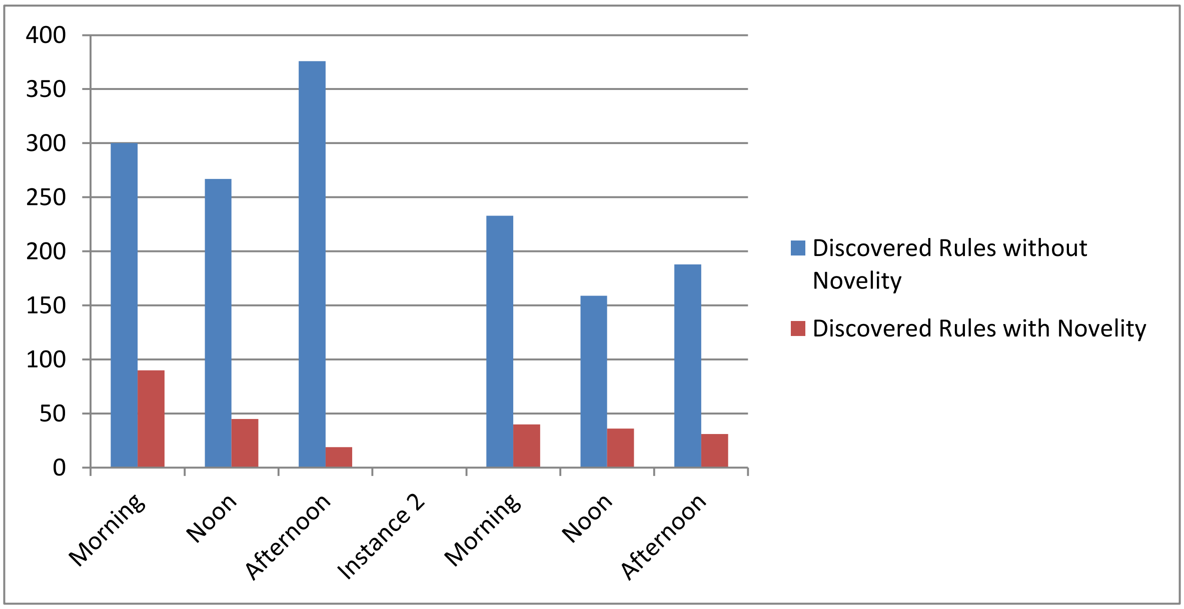


table 4.2 and fig 4.2

Figure 4.2: Effect of Novelty Threshold Φ=0.5 on the Number of Rules

4.2.3 Experiment Three

The objective of the third experiment is to show the effectiveness of the proposed approach in reducing the number of discovered rules . It is expected that the number of discovered rules keeps on decreasing over the time. We work with two browsing interval datasets and considered these datasets as evolving with time, and partitioned them into two browsing interval increments: mined at times T1 ,T2 respectively. For each dataset used, the minimum support that the number of interesting rules decreases in a proposed approach in contrast to the number of rules discovered by Apriori algorithm at T1 , and T2. Intuitively, the interesting rules discovered by our approach at time T1 is no more interesting at time T2 and the interesting rules discovered at time T2 is no more interesting at time T3. Consequently, as the value of novelty threshold Φ increases, the number of discovered interesting rules decreases at each time as per our expectations. The results are demonstrated in Table (4.3).

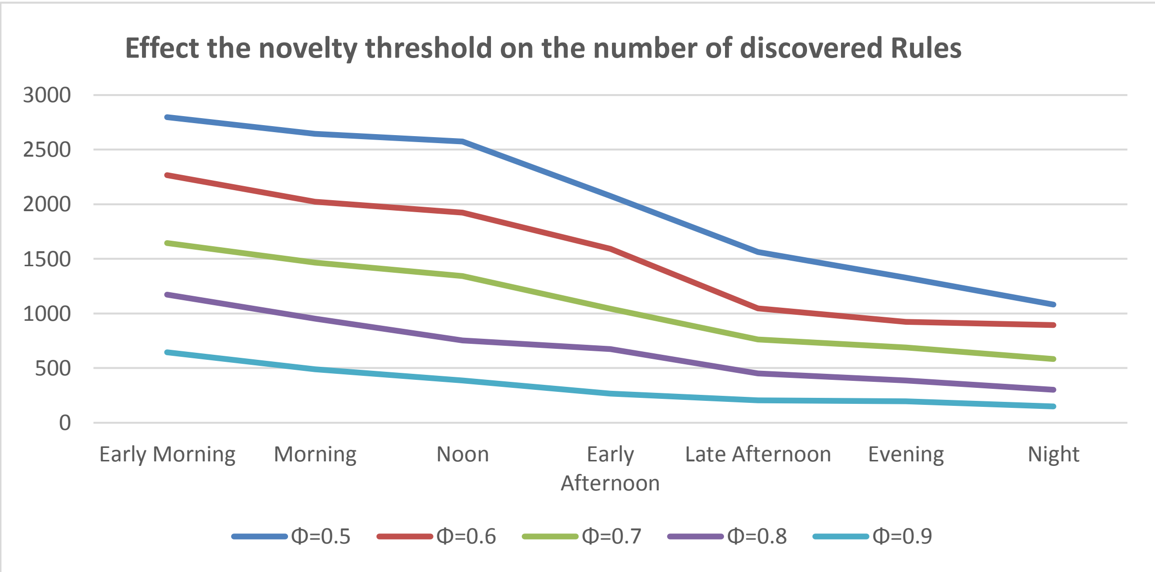


Figure 4.3 Effectiveness of the novelty threshold on the number of discovered Rule

Intervals		Effect the novelty threshold on the number of discovered Rules				
		Φ=0.5	Φ=0.6	Φ=0.7	Φ=0.8	Φ=0.9
Time T1	Early Morning	2798	2265	1645	1173	645
	Morning	2645	2023	1465	952	489
	Noon	2573	1923	1344	753	388
	Early Afternoon	2077	1592	1043	673	266
	Late Afternoon	1562	1046	762	452	205
	Evening	1329	923	689	388	196
	Night	1082	893	583	301	149
Time T2	Early Morning	852	801	509	267	133
	Morning	599	507	480	212	112
	Noon	363	345	430	190	103
	Early Afternoon	308	288	218	179	101
	Late Afternoon	287	276	203	170	90
	Evening	131	120	106	94	82
	Night	84	69	52	41	26

Table 4.3 : Effect the novelty threshold on the number of discovered Rules

The number of interesting rules decreases in a proposed approach at eight intervals. The interesting rules discovered by the a proposed approach at first interval is no more interesting at the second interval and the interesting rules discovered at at the second interval is no more interesting at third interval. As per our expectations. The results shown in Table 4.3 and Figure 4.3

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

In this work , we have proposed approach for incremental association rules mining that integrates interestingness criterion during the process of building the model. One of the main features of the proposed approach is to capture the user background knowledge, which is monotonically augmented. The proposed approach is a self-upgrading filter that utilizes interestingness criterion to reflect the user subjectivity and extract patterns, incrementally, from datasets arrive . The proposed approach makes use of interestingness measure as the basis of extracting interesting patterns. This important feature of the proposed approach is attractive and desirable in many real life applications as the volume of data keeps on growing and changing over the time and therefore the user background knowledge is monotonically augmented.

We have proposed a generalized fuzzy data mining algorithm to extract interesting patterns. The proposed approach uses static membership functions to fuzzify the quantitative web usage data along with predefined membership function. We also use predefined support and divided whole database into different partitions based on intervals . At each interval partition, we apply separately fuzzy mining algorithm to extract association rules. Finally, all intervals association rules combined to declare total number of rules one for given database.

5.2 FUTURE WORKS

The future work will aim to further develop mechanism for building user profile dynamically that is better suited for personalization systems.

We plan a comprehensive user evaluation will be conducting user trials testing , by a method to show the entire process in a graphical user interface, this will make the system easier to use, and investigate both the efficiency and impact of the personalization and adaptation components on users.

Reference

[1] K. S. Rao, A. R. Kalyani, M. Krishnamurthy, "A New Approach of Inferring User Interest for User Search Goals with Web Log Sessions", Journal of Web Development and Web Designing, Volume-1 Issue-2, Page 1-15, (2016).

[2] T. Th. S. Nguyen, H. Y. Lu, and J. Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 10,(October 2014).

[3] M. Dhandi, R. K. Chakrawarti, "A Novel Approach for Web Personalization", International Journal of Computer Applications (0975 – 8887) Volume 151 – No.8, (October 2016).

[4] B. Mobasher, R. Cooley and J. Srivastava, "Automatic personalization based on Web usage mining," Communications of the ACM, Vol.43, pp.142-151, No.8, (1997).

[5] B. Sh. Suryavanshi, N. Shiri and S. P. Mudur, "Incremental Relational Fuzzy Subtractive Clustering for Dynamic Web Usage Profiling," In Proc. of ACM Workshop on Knowledge Discovery in the Web, (2005).

[6] B. Xu, J. Bu, C. Chen and D. Cai, "An Exploration of Improving Collaborative Recommender Systems via User-Item Subgroups," In Proc. of the IEEE 21st international conference on World Wide Web, ACM, pp.21-30 , (2012).

[7] B. M. Kim, Q. Li, J.-W. Kim, and J. Kim, "A New Collaborative Recommender System Addressing Three Problems, "Springer, LNCS, Vol.3157, pp495-504, (2004).

[8] C. Shahabi, A. Zarkesh, M. Abidi, J. and V. Shah, "Knowledge discovery from users Web-page navigation," In Proc. of IEEE 7th International Workshop on Research Issues in Data Engineering, pp.20-29, (1997).

[9] D. Pierrakos, G. Paliouras and Y. Ioannidis, "OurDMOZ: A System for Personalizing the Web," In Proc. of 6th International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases, (2012).

[10] D. Pierrakos, G. Paliouras, Ch. Papatheodorou and Constantine D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: A Survey," ACM, User Modeling and User-Adapted Interaction, Vol.13, No.4, pp.311-372, (2003).

[11] B. Mobasher, R. Cooley and J. Srivastava, "Automatic personalization based on Web usage mining," Communications of the ACM, Vol.43, No.8, pp.142-151, (1997).

[12] http://www142.ibm.com/software/products/us/en/personalized-productrecommendations

[13] A. P. Warade , V. M. Natarajan and S. Sh. Chandak, "How to Develop Online Recommendation Systems that Deliver Superior Business Performance," Cognizant 20-20 Insights.

[14] V. Dongre, J. Raikwal, "An Improved User Browsing Behavior Prediction Using Web Log Analysis", International Journal of Advanced Research in Computer Engineering & Technology, Volume 4 Issue 5, (May 2015).

[15] M. Jafari, F. S . Sabzchi, A. J. Irani, "Applying Web Usage Mining Techniques to Design Effective Web Recommendation Systems: A Case Study", ACSIJ International Journal, Vol. 3, Issue 2, No.8, pp. 78-90,(March 2014).

[16] G. Stumme, A. Hotho, and B. Berendt, "Usage mining for and on the Semantic Web", In Data Mining: Next Generation Challenges and Future Directions. Menlo Park, CA, USA: AAAI/MIT Press, pp. 461–480, (2004).

[17] D. Pierrakos, G. Paliouras, Ch. Papatheodorou and C. D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: A Survey," ACM, User Modeling and User-Adapted Interaction, Vol.13, No.4, pp.311-372, (2003).

[18] R. F. Dell, P. E. Roman, and J. D. Velasquez, "Web User Session Reconstruction Using Integer Programming", IEEE International Conference on Web Intelligence and Intelligent Agent, (2008).

[19] J. Han and Kamber, "DataMining: Concepts and Techniques," Morgan Kaufmann Publishers.Second Edition, (2006)

[20] J. Srivastava, R. Cooley, M. Deshpande and P.-N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data," ACM SIGKDD Explorations Newsletter, Vol.1, No.2, pp.12-23,(2000).

[21] C. I. Ezeife and Y. Lu, "Mining Web log sequential patterns with position coded pre-order linked WAP-tree", IEEE Data Minining Knowledge Discovery, Vol. 10, no. 1, pp.5–38, (2005).

[22] S. Khonsha and M. H. Sadreddini, "New hybrid web personalization framework", In Proc. of IEEE 3rd International Conference on Communication Software and Networks, pp.86-92, (2011).

[23] M. Jalali, N. Mustapha, Md. N. B Sulaiman, A. Mamat, A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems, 12th International Conference Information Visualisation,(2008).

[24] G. Fang, Jia-le Wang, H. Ying, J. Xiong, "A double algorithm of Web usage mining based on sequence number", IEEE, (2009).

[25] M. Eirinaki and M. Vazirgiannis, "Web Mining for Web Personalization," ACM Transactions on Internet Technology, Vol.3, No.1, pp.2-38,(2003).

[26] K. P. Adhiya And S. R. Kolhe, "An Efficient And Novel Approach For Web Search Personalization Using Web Usage Mining", In Journal Of Theoretical And Applied Information Technology, Vol. 73, No. 2, (March 2015).

[27] O. Etzioni, "The World Wide Web: Quagmire or gold mine," Communications of the ACM, Vol.39, No.11, pp.65-68, (1996).

[28] R. Kosala and H. Blockeel., "Web mining research: a survey," ACM SIGKDD Explorations Newsletter, Vol.2, No.1, pp.1-15,(2000).

[29] R. Agrawal and R. k. Shrikant, "Fast Algorithms for Mining Association Rules in Large Databases," In Proc. of 20th International Conference on Very Large DataBases, pp.487-499,(1994).

[30] Kuok, Ch. Man, Fu, Ada; Wong, Man Hon: Mining Fuzzy Association Rules in Databases. SIGMOD Record Volume 27, (1998).

[31] Kuok, Ch. Man, Fu, Ada; Wong, Man Hon: Mining Fuzzy Association Rules in Databases. SIGMOD Record Volume 27, (1998).

[32] V. Ganti, J. Gehrke, and R. Ramakrishnan, “Demon: Mining and Monitoring evolving data”, In Proceeding of the 16th International Conference on Data Engineering, San Diego, USA, (2000).

[33] B. Liu, W. Hsu, and S. Chen, “Using General Impressions to Analyze Discovered Classification Rules,” in Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD 97), USA, pp.

[34] B. Liu, W. Hsu, F. Mun, and Y. Lee, “Finding Interesting Patterns Using User Expectations,”Technical Report TRA7/96, (1996).

[35] P. Markellou, I. Mousourouli, S. Sirmakessis, A. Tsakalidis , “Personalized Ecommerce Recommendations”, Proceedings of IEEE International Conference on e-Business Engineering, (2005).

[36] J.-G. Liu, W.-P. Wu, ”Web Usage Mining for Electronic Business Applications “, Proceedings of the Third International Conference On Machine Leaming and Cyhemetics, Shangha, 26-29, (August 2004).

[37] A. S. Al-Hegami, A. S. Kumar, N. Bhatnagar, V. , " Novelty Framework for Knowledge Discovery in Databases", In Proceedings of the 6th International Conference on Data warehousing and Knowledge Discovery (DaWaK 2004), Zaragoza, Spain, pp 48-55, (2004).

[38] "Server Log Files", http://www.mediacollege.com/ internet/statistics/ logs/

[39] M. P. , S. K. Pani, S. K. Padhi, L. Panigrahy, R. Ramakrishna, "A Review Of Trends In Research On Web Mining", International
Journal of Instrumentation, Control & Automation, Volume 1, Issue 1, pages 37-41, (2011).

[40] G. K. Lekeas, "Data mining the web: the case of City University's Log Files," (2000).

[41] J. Srivastava, R. Cooley, M. Deshpande, P. N. Tan "Web usage mining: Discovery and Applications of usage patterns from web data" SIGKDD Explorations- vol-1, issue-2,pages 12-33, (Jan 2000).

[42] K. Suneetha and R. Krishnamoorthi, "Identifying user behavior by analyzing web server access log file," IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 4, pp. 327–332, (2009).

[43] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for," International Journal of Data Mining Techniques and Applications, vol. 02, no. 01, pp. 141–150, (2013).

[44] P. Shapiro, G. Matheus, C. J., "The Interestingness of Deviations", In Proceedings of AAAI Workshop on Knowledge Discovery in Databases, (1994).

[45] Silberschatz, A.Tuzhilin, A., "On Subjective Measures of Interestingness in Knowledge Discovery", In Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, (1995).

[46] K. B. Patel, J. A. Chauhan, and J. D. Patel, "Web Mining in E-Commerce : Pattern Discovery , Issues and Applications," International Journal of P2P Network Trends and Technology, vol. 1, no. 3, pp. 40–45, (2011).

[47] Borges, J. and Levene, M. "Data mining of user navigation patterns. InWeb Usage Analysis and User Profiling", Lecture Notes in Computer Science, vol. 1836, Springer-Verlag New York, 92–111(1999)

[48] A.G. Buchner, M. Baumgarten, , and, S.S.Mulvenna, M..D., and J.G. Hughes, "Navigation pattern discovery from Internet data. In Proceedings of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99)", Fifth ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining (San Diego, August), 25–30, (1999).

[49] A.M.D. MULVENN, S.S. ANAND, AND A.G. BUCHNER, “Personalization on the net using web mining” , ACM, 43,123–125. (August 2000)

[50] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases” in Proc. of the ACM SIGMOD Int’l Conf. on Management of Data, Washington D.C., (May 1993)

[51] R. Cooley, B. Mobasher, J. Srivastava “Data Preparation for mining World Wide Web browsing patterns”. Knowledge Informa Sys (1999).

[52] S. Kumar, A. Kunjithapatham, M. Sheshagiri, T. Finin, A. Joshi, Y. Peng, and R.S.Cost, “A Personal Agent Application for the Semantic Web, AAAI Fall Symposium on Personalized Agents”, North Falmouth, (2002).

[53] B. Ganter, and R.Wille,“Formal Concept Analysis: Mathematical Foundations”. Springer, Heidelberg, (1999)

[54] P.Y. Lee, S.C. Hui, and A.C.M. Fong, “Neural Networks for Web Content Filtering”. IEEE Intelligent Systems, 17(5) 48-57, (2002).

[55] D. Olaru, and B. Smith “Modelling Daily Activity Schedules with Fuzzy Logic”. In: Proc. Of the 10th Intl. Conf. in Travel Behaviour Research, Lucerne, Switzerland, (2003).

[56] T.P. Hong, and C.Y. Lee, “An overview of mining fuzzy association rules,” In: H. Bustince, F.Herrera, J. Montero (eds.) Studies in Fuzziness and Soft Computing, vol. 220, pp. 397–410. Springer Berlin/Heidelberg (2008)

[57] J. M. Ale, and G. H. Rossi, “An approach to discovering temporal association rules”. In Proc. of the 2000 ACM Symposium on Applied Computing, pages 294–300, (2000).

[58] Z. Lianga, T. Xinming, L. Lin, and J. Wenliang, “Temporal Asocation Rule Mining Based On T-Aprior Algorthim And Its Typical Application”. ,Chine, (2008).

[59] J. Han, and M. Kamber, “Data Mining: Concepts and Techniques”. San Francisco, Morgan Kauffmann Publishers, (2001).

[60] V. Bhatnagar, A. S. Al-Hegami, and N. Kumar, “A hybrid approach for Quantification of Novelty in Rule Discovery”. In Proceedings of International Conference on Artificial Learning and Data Mining (ALDM’05), Turkey, Feb. 25-27, pp 39-42 (2005).

[61] V. Bhatnagar , A. S. Al-Hegami, and N. Kumar, “Novelty as a Measure of Interestingness in Knowledge Discovery”. In International Journal of Information Technology, Volume

2, Number 1, (2005).

[62] A. S. Al-Hegami, “Pushing Novelty Criterion into Incremental Mining Algorithm”, International Journal of Computer Science and Network Security,Korea, VOL.7 No.12, (2007).

[63] A.S. Al-Hegami,“Subjective Measures and their Role in Data Mining Process”, In Proceedings of the 6th International Conference on Cognitive Systems, New Delhi, India, (2004).

[64] A.S. Al-Hegami, “On Quantification of Novelty in Knowledge Discovery Systems”, Ph.D. Dissertation. Department of Computer Science, University of Delhi, India, (2006).

[65] B. Liu and W. Hsu, “Post Analysis of Learned Rules,” in Proceedings of the 13th National Conference on AI (AAAI’96), pp. 1194-2001, (1996).

[66] B. Liu, W. Hsu, and S. Chen, “Using General Impressions to Analyze Discovered Classification Rules”, in Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD 97), USA, pp.73-83, (1997).

[67] E. Al- Yafi , A. S, Alam, M. A., and R. Biswas, “Incremental Mining of Shocking Association Patterns”. In Proceedings of World Academy of Science, Engineering and Technology, Volume 37, Dubai, UAE, (2009).

[68] G. Stumme, A. Hotho, and B. Berendt, "Usage mining for and on the Semantic Web", In Data Mining: Next Generation Challenges and Future Directions. Menlo Park, CA, USA: AAAI/MIT Press, pp. 461–480, (2004).

[69] F. Cai, Sh. Wang, M. de Rijke, "Behavior-Based Personalization in Web Search", Journal Of The Association For Information Science And Technology, 68(4):855–868, (2017).

[70] A. S. Al-Hegami, “A Framework for Incremental Mining of Interesting Temporal Association Rules", International Journal of Computer Applications (0975 – 8887) Volume 131 – No.8,(December2015)

[71] S. Chitra and B. Kalpana Optimum Session Interval based on Particle Swarm Optimization for Generating Personalized Ontology , Journal of Science and Technology, Vol 7(8), 1137–1143, August 2014

[72] A.S Alhegami, H.A Alsaeedi, A framework for Incremental parallel Mining of Interesting Association Patterns for Big Data, International Journal of Computing,19(1) 2020,106-117

[73] A. Shollo The Role of Business Intelligence in Organizational Decision-making Copenhagen Business School [Phd]. PhD series No. 10.2013

[74] S. Jagan, S.P. Rajagopalan, "A Survey on Web Personalization of Web Usage Mining" International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 01, March-2019

[75] D.Kavitha, K.M.Subramanian, Dr.K.Venkatachala SURVEY ON INFERRING USER SEARCH GOALUSING FEEDBACK SESSION , International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 12, December 2013

[76] N. Abdullah1, Yue Xu2, and Sh. Geva3 , Integrating Collaborative Filtering and Matching-based Search for Product Recommendations , Queensland University of Technology, School of Electronic Engineering and Computer Science, Brisbane, Australia, 2013

[77] J. Srivastav , R. Cooley, B. Mobasher,(1997)" Web Mining Information and Pattern Discovery on world wide web" at Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence,1997,ISBN: 0-8186-8203-5

[78] V. Malik1, Dr. S. Singla, A Performance Analysis of Web Usage Mining Tool, International Journal of Research in Advent Technology, Vol.7, No.1, January 2019

[79] R.C. Getek , A usability model for web-basedpersonalization based on privacy and security, PhD dissertation ;University of Maryland, Baltimore County (2010)

[80] M. Spilipoulou and B. Mobasher, B. Berendt,"A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis", INFORMS Journal on Computing
Spring ,2003

[81] R.Cooley, B. Mobasher, and J. Srinivastava, " Web mining: Information and Pattern Discovery on the World Wide Web", In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997

[82] L. Catlegde and J. Pitkow, "Characterising browsing behaviour s in the world wide Web", Computer Networks and ISDN systems, 1995

[83] K. P. Adhiya And S. R. Kolhe, "An Efficient And Novel Approach For Web Search Personalization Using Web Usage Mining", In Journal Of Theoretical And Applied Information Technology, Vol. 73, No. 2, March 2015.