Web Recommender System based on Consumer Behavior Modeling using Fuzzy Representation

V. T. Chavan\textsuperscript{1}, Prof. S. K. Shirgave\textsuperscript{2}
\textsuperscript{1}Assistant Professor, Department of Computer Science \& Engineering, Sanjay Ghodawat Institute, INDIA
\textsuperscript{2}Professor, I.T Department, DKTES’ Textile and Engineering Institute, INDIA

ABSTRACT
Among various web related activities, web surfing, online purchases are more popular activities. Day by day World Wide Web proved to be the biggest source of information and contains a huge amount of data. But many times, most of the received or recommended data is irrelevant and inaccurate from users’ point of view. For users, it becomes necessary to use recommendation systems to discover and extract the desired information and resources. Web recommender systems predict the information needs of users. Recommender systems provide relevant recommendations to the users. This paper describes a Web recommender system that constructs a knowledge base using fuzzy temporal Web access patterns as input. The user’s Web access habits and behavior is modeled using the knowledge base. The Knowledge base is used to generate association rules which are used to provide personalized recommendations to the user. The fuzzy representation is used to construct a knowledge base. Fuzzy logic is applied to requested resources of periodic pattern-based Web access activities. These fuzzy patterns are used to generate association rules for Recommendation System. Experimental results on the session data of different Web sites showed the effectiveness of this approach.

Keywords----- consumer habits, personalization, recommender system, Web log mining, knowledge discovery, semantic Web, fuzzy temporal pattern.

I. INTRODUCTION
The World Wide Web is the vast source of information. Everyone can easily access this existing information at any time. The Web has become a very popular medium for consumer to exchange or find ideas, opinions, experiences on products and services. Along with online information sharing Many consumers actually perform purchases on the Web. The increasing availability and popularity of portable Web enabled handheld devices, looks set to fuel further growth in the volume of consumer Web traffic.

It is often hard for Web users seeking the information they really want. The problem of the user is how to divide his attention between the numerous sources of information available to him. These types of problems are described by using two new terms. The first term is “Information overload” which “because of having too much information, decision making about a topic becomes difficult”. The second term is “Lost in hyperspace” which describes the tendency of users to lose their way because of the nonlinear nature of browsing in a hypertext environment [1]. Consequently, for users to utilize automated tools become increasingly necessary in order to discover and extract the desired information and resources. Now days it is ordinary for the Web users to come across Web sites that offer recommendations for products and services, banner advertising, and individualized link selections. Web recommender systems predict the information needs of users and provide them with relevant recommendations. Recommender systems are custom-built information agents that provide recommendations for items likely to be of use to a user [2]. Recommender systems have been attracting more attention as a suitable approach to work against information overload and help the users of the Web information space. For example, in [3, 4, 5], the various authors described a number of personalized systems for TV program recommendations, or recommendations for multimedia channel selection. Some researchers have performed behavior profiling to enhance the quality of the recommendations. However, as customers tastes and desires are transient and fractal, it is very hard to recommend to the users.

This paper focuses on effective and timely personalized Web content recommendation using consumer behavior modeling as a natural extension of the types of behavior profiling. This work describes Web content recommender system that generates personalized recommendations to consumers based on a model of
consumers web access habits and behaviors. Further, the model is stored in a knowledge base prepared in advance. This system mines periodic access patterns, and temporal information to construct the knowledge base so that subsequent applications will provide recommendations that are both useful and timely. Specifically, fuzzy logic is applied to better represent real-life temporal concepts and requested resources attributes of periodic pattern-based Web access activities. Knowledge base of the consumer’s Web access habits and behaviors is constructed using fuzzy representation. This Knowledge base is then used to provide useful personalized recommendations to the consumer.

II. RELATED WORK

This section presents a survey of work pertinent to personalized Web content recommendation and consumers’ Web access pattern mining based on periodicity information. This section therefore serves to lay the foundation for further discussion of the proposed approach in Section III.

A. Web Content Recommendation Using Fuzzy Association Rule Mining

Web usage mining is the application of data mining techniques. Web usage mining is used to discover usage patterns from Web data. Web Usage Mining applies a series of diverse techniques, like clustering, classification, discovery of association rules or sequential patterns. Association rule mining discovers relation of association between existing records in a dataset. Association rules can help to discover relationships between Web resources accessed by a user that would otherwise be missed, especially if the resources are disjoint. They can also be used to find groups of people with similar interests. There are various techniques for efficient mining of association rules. One of the widely used technique is apriori algorithm. Apriori algorithm is used to generate a boolean association rule that associates the usage pattern of the clients for a particular website. Apriori is very influential algorithms for mining frequent item sets for association rules [6]. However above Crisp Association Rule Mining (ARM) algorithms can mine only binary attributes. It requires any numerical attributes to be converted to binary attributes. Until now the method for this conversion process is to use ranges, and try to fit the numerical attribute values in these ranges. However, using ranges introduces uncertainty at the boundaries of ranges, leading to loss of information.

A major problem of traditional association rule mining is that each item in a record is considered only either to exist or not. Thus, the user’s preference and interest on each transaction item cannot be precisely represented. Crisp association rules use sharp partitioning to transform numerical attributes to binary valued attributes, and can potentially introduce a loss of information due to these sharp ranges.

A better way to solve this problem is to use fuzzy pattern. Fuzzy pattern have attribute values represented in the interval [0, 1], instead of having binary 0 and 1, and have transactions with a given attribute represented to a certain extent (in the range [0, 1]). Fuzzy logic used to convert numerical binary attributes to fuzzy attributes, thus maintaining the integrity of information conveyed by such numerical attributes. In this way binary attributes are replaced by fuzzy attributes [6]. The concepts of preference and interest are fuzzy data, fuzzy logic [8] can be applied. These Fuzzy attributes serve input to fuzzy association rules Mining Algorithm. Wong et al. [9] combine fuzzy association rule mining and case-based reasoning (CBR) [7] to improve the quality of Web access pattern prediction. The fuzzy rule set was found to perform better in prediction accuracy and rule coverage than traditional rule set.

B. Periodicity-Based Pattern Mining

Most of the previous work focused only on mining common sequential access patterns of Web access events, which frequently occurred within the entire duration of all user’s Web access transactions. In practice, many useful sequential access patterns frequently occur during a particular time interval, such as morning or late evening. Such a sequential access pattern called as periodic sequential access patterns. Discovering such periodic patterns from access log records is an important Web mining task for many applications.

Recommender system Methods focuses on the different characteristics of the user. The fundamental requirement of an effective, Web content recommendation system is to present the most relevant suggestions to the user. Thus, both context and temporal information are important. Currently systems tend to lack focus on temporal information.

Timeliness in the recommendation of relevant resources is important in many situations. For example, a user may have a tendency of reading the financial news and traffic information between 9 am and 10 am on weekdays. Periodic association rules associated with a set of events that occur periodically. Periodic association rules hold only during certain time intervals of transactions, but not others. Calendar information is critical in describing the time intervals, the authors studied the problem of discovering cyclic association rules that show regular cyclic variation over time [12]. However, this work is unable to describe real-life time concepts and deal with multiple granularities of time, such as the morning of every weekend. In [13], the authors proposed calendric association rules that extend beyond cyclic association rules for handling multiple units of time. In [10], [11], Calendar schemas were proposed for discovering temporal association rules. By using calendar schema as a framework for mining temporal patterns, it requires less prior knowledge of data than the previous approaches in [12], [13].

III. KNOWLEDGE BASE USING FUZZY PATTERN
have been cleaned, the next step in the data pre-processing need to apply on web access log files. Pre-processing is the technique used to transform the raw web access log data into a format that will be more easily and proficiently processed to do the required task. The phases of pre-processing include data cleaning, user identification and session identification. The process of data cleaning is to remove erroneous, redundant references and irrelevant data. Also, all log entries with file name suffixes such as GIF, JPEG, JPG can be eliminated since they are irrelevant. Once HTTP log files have been cleaned, the next step in the data pre-processing is the identification of users using IP address. The Session identification is a process of partition the access log of each user into distinct access sessions. If the duration of a session exceeds a certain limit, then it is considered as another access session of the user. Discovered from empirical findings, a 30-min threshold for the total session duration has been recommended. From this preprocessed data, it needs to extract semantic attributes and periodic patterns. Then generate Fuzzy pattern and knowledge base. Use this Fuzzy representation to generate fuzzy Association rules. These rules serve input to recommendation system to generate Recommendations.

In order to obtain meaningful results, Web usage mining must exploit the semantics of the pages visited along user paths. The URLs recorded in Web usage logs contain little semantic information about the Web content accessed by users. This makes it difficult to be used for understanding users' actual access behaviors, interests and intentions. Some form of semantic enhancement is therefore required to make the Web log data really useful. Annotate each requested URL in Web usage logs with the corresponding semantic information, i.e., One or more predefined topic, concepts or categories, such as Product News, Current Affairs, Sports and Entertainment. These attributes further referred as resource attributes which are used for the generation of fuzzy patterns.

This work deals with Web recommender that models user behaviors and habits by constructing a knowledge base using temporal Web access patterns and Web semantics as input. Using the semantic Web usage logs as inputs, it first identifies a set of periodic attributes (i.e., temporal concepts such as morning, afternoon) and a set of resource attributes (i.e., semantic information such as topic, concepts or categories) to represent periodic pattern-based Web access activities.

Fuzzy logic is applied to periodic attributes and resource attributes to represent real-life temporal concepts and requested resources of periodic pattern-based Web access activities. This representation is used to generate a knowledge base of the user’s Web access habits and behaviors, which is used to provide recommendations to the user.

If the user accesses specific resources periodically, then the user has a Web access activity (i.e., periodic Web access pattern). Use a set of periodic attributes $M_p$ and resource attributes $M_r$ to represent Web access activity. This proposed approach recommends to define eight real-life temporal concepts as Early Morning, Late Morning, Noon, Early Afternoon, Late Afternoon, Evening, Night and Late Night, as periodic attributes. More generally, days of the week (e.g., Monday, Tuesday, etc.) or other real-life temporal concepts (e.g. weekdays, weekends, etc.) can be also used as periodic attributes.

From a semantic Web usage logs as input, it first need to identify periodic attributes (i.e., temporal concepts such as morning,afternoon) and a set of resource attributes (i.e., useful domain ontological concepts) to represent periodic pattern-based Web access activities. A user’s Web access session $S = (U_{URL_1}, t_1); (U_{URL_2}, t_2), \ldots, (U_{URL_n}, t_n)$ is a sequence of requested URL with timestamp $t_i$ ($1 \leq i \leq n$). In general, the time spent by a user for a URL may indicate the level of interest that the user has in the content of that URL. The duration $d_i$ of $URL_i$ can be estimated simply as $d_i = (t_i+1 - t_i)$. For the last requested URL $URL_n$ in each user access session that does not have $t_{n+1}$, use the average duration of the current session for estimating its duration, i.e., $d_n = (d_1 + d_2 + \ldots + d_{n-1})/(n-1) = (t_n - t_1)/(n-1)$. To compute $d_n$, need $n > 1$, i.e., retain user access sessions that contain more than one requested URL [14]. Furthermore, evaluate the start time and end time of $S$ as $t_1$ and $t_n + d_n$ respectively.

A period of a user access session $S$ is defined as a continuous time interval with a session start time $t_s(S)$ and a session end time $t_e(S)$, denoted as

$$P(S) = \begin{cases} \{t_s(S), t_e(S)\}, & \text{if } t_s(S) \leq t_e(S) \\ \{0, t_e(S)\} \cup [t_s(S), 24], & \text{otherwise} \end{cases}$$

Define a fuzzy periodic Web Usage Context $K$ based on the preprocessed access sessions [14]. From the definition of $K$, we can further define the set of attributes common to user access sessions, the set of user access sessions having the same attributes, fuzzy support of a set of attributes, and the notion of Web access activity.

$$K = (G, M_p, M_r, I)$$
where $G$ is a set of user access sessions and $I = R \left( G \times (M_p \cup M_r) \right)$ is a fuzzy set in the domain of $G \times (M_p \cup M_r)$ to represent fuzzy relations between access sessions $g \in G$ and attributes $m \in (M_p \cup M_r)$. Each fuzzy relation is represented by a membership value between 0 and 1.

$$
\mu(g, m) = \begin{cases} 
\mu_p(g, m), & \text{if } m \in M_p \\
\mu_r(g, m), & \text{if } m \in M_r
\end{cases}
$$

From this definition, each user access session $g \in G$ can also be denoted as a fuzzy set on the domain of $M_p \cup M_r$ i.e., $g = \{m, \mu(g, m) \mid m \in (M_p \cup M_r)\}$

The membership value $\mu_p(g, m_p)$ for a periodic attribute $m_p \in M_p$ in a user access session $g \in G$ can be computed using the period of $g$, i.e., $p(g)$.

Here, the membership function is defined as $\mu_p(g, m_p) = \max_{t \in p(g)} (\mu_p(t, M_p))$, where $\mu_p(t, M_p)$ is defined in Figure 2 which is modified from [14].

![Figure 2. Member function $\mu_p(t, M_p)$](image)

The membership value $\mu_r(g, m_r)$ for a resource attribute $m_r \in M_r$ in a user access session $g \in G$ can be computed using the total duration of $m_r$ i.e., $d(g, m_r)$ [14]. The membership value of the resource attribute is defined as

$$
\mu_r(g, m_r) = \begin{cases} 
0, & \text{if } z(g, m_r) < \frac{1}{2} Z(m_r) \\
\frac{z(g, m_r)}{Z(m_r)} - 1, & \text{if } \frac{1}{2} Z(m_r) \leq z(g, m_r) \leq Z(m_r) \\
1, & \text{if } z(g, m_r) > Z(m_r)
\end{cases}
$$

where $z(g, m_r) = \frac{d(g, m_r)}{\sum_{g \in G} d(g, m_r)}$, $Z(m_r)$ is the proportion of the total duration of accessing the resource $m_r$ in all Web access sessions of the user, which indicates the user's global interest of the resource $m_r$, $z(g, m_r)$ is the proportion of the duration of accessing the resource $m_r$ within the user access session $g$, weighted by an emotional influence factor $e_r = 1$.

Table 1 shows an example Web Usage Context of a user, which consists of five user access sessions, three periodic attributes “P1 (night)”, “P2 (Early morning)”, and “P3 (Morning)”, and three resource attribute “R1”, “R2” and “R3”.

In order to construct user behavior knowledge base, the system identifies the activity relationships [14]. Given a Web Usage Context $K = (G, M_p, M_r, I)$ the fuzzy support of a set of attributes $B \subseteq M_p \cup M_r$ and $B \neq \emptyset$

$$
\text{Sup}(B) = \frac{\sum_{g \in B} (\mu_p(g) \times \mu_r(g))}{|G|}
$$

Where

$$
\mu_p(g) = \begin{cases} 
\min_{m_p \in B \cap M_p} \left( \mu_p(g, m_p) \right), & B \cap M_p \neq 0 \\
1, & \text{otherwise}
\end{cases}
$$

$$
\mu_r(g) = \begin{cases} 
\min_{m_r \in B \cap M_r} \left( \mu_r(g, m_r) \right), & B \cap M_r \neq 0 \\
1, & \text{otherwise}
\end{cases}
$$

Web access activity $v(B) = \{m, \mu(B, m) \mid m \in B\}$, where $\mu(B, m) = \max_{g \in B} (\mu(g, m))$. The fuzzy support of $v(B)$ is defined as $\text{Sup}(v(B)) = \text{Sup}(B)$ and the fuzzy confidence of $v(B)$ is defined as $\text{Conf}(v(B)) = \frac{\text{Sup}(B)}{\text{Sup}(B \cap M_p)}$.

The next step is to construct a knowledge structure to represent all relevant activity relationships. For example, suppose there are two Web access activities for a user known as $v(B_3)$, $v(B_4) \in W_p$ where $W_p$ is the set of all Web access activity of a user, $v(B_i)$ is sub-activity of $v(B_j)$ if and only if $B_i \subseteq B_j$ Virtual Web access activity $v(\cdot)$ represents the super-activity of all other Web access activities and therefore sits at the root of the knowledge structure with $\text{Sup}(v(\cdot)) = 1.0$ and $\text{Conf}(v(\cdot)) = 1.0$. Figure 3 shows Segment of user behavior knowledge base with the support and confidence of Web access activities [14].

![Figure 3. Segment Of User Behavior Knowledge Base](image)
IV. RECOMMENDER SYSTEM BASED ON ASSOCIATION RULE DISCOVERY

One of the most important, effective and widely used data mining techniques is Association rule mining. Association rules capture the relationships between items based on their patterns of co-occurrence across transactions [15]. Association rules that reveal similarities between the items derived from user behavior can be simply utilized in recommender systems. This recommendation system suggests to the current user some Web pages that appear to be useful [16].

The problem of association rule mining is defined as: Let \( I = \{i_1, i_2, i_3 \ldots i_n\} \) be a set of \( n \) binary attributes called items. Let \( D = \{t_1, t_2, t_3 \ldots t_m\} \) be a set of transactions. A rule is defined as implications of the form \( X \Rightarrow Y \) where \( X, Y \subseteq I \) and \( X \cap Y = \phi \) where \( X, Y \) are the items. To select interesting rules from the set of all possible rules, different constraints on measures of significance and interest can be used. The best-known and widely used constraints are minimum thresholds on support and confidence [17].

Generation of fuzzy association rules algorithm is not a straightforward process. Firstly, Crisp dataset must be converted into a fuzzy dataset. Second, fuzzy ARM algorithms need to take into account the fuzzy membership of an itemset in a set of transaction, in addition to its presence or absence, whereas the crisp ARM algorithm calculate the frequency of an itemset just by considering for its presence or absence in a transaction of the dataset.

Fuzzy ARM, like many ARM algorithms uses the support-confidence framework [6], a feature of which is that for an item set to be "frequent" its supersets must also be frequent. The support for a 1-itemset is simply the sum of the membership values divided by the number of records in the data set. The support for N-item sets, for each record containing the item set, is the sum of the products of the membership values in each record divided by the number of records in the data set. Thus, given 4 data set of the form having letters as attribute identifiers, and numerical value (in the range [0, 1]) representing membership value:

\[
\begin{align*}
&<c,1.0> \\
&<a,0.5> <b,0.5> \\
&<a,0.5> <c,0.5> \\
&<a,0.5> <b,0.5>
\end{align*}
\]

The calculated support will be:

\[
\begin{align*}
&\{a\}=0.375 \\
&\{c\}=0.375 \\
&\{a \ c\}=0.0625 \\
&\{a \ b\}=0.125
\end{align*}
\]

Fuzzy Association rules generated from the previous step are used as input for the recommendation engine to generate recommendations. The recommendation engine extracts the resource attributes from page views of the active user session window and then uses association rules in order to search associated resource attributes. All these input resource attribute and associated resource attributes are used to generate recommendation set. Pages which contain the maximum of these recommended attributes are added in the final recommendation set.

V. IMPLEMENTATION

Web access log data is preprocessed as shown in figure 4. Sessionization is implemented to find useful sessions from the preprocessed web access log. To generate fuzzy Association rules, Apriori-T package is used. Association Rule generation using fuzzy Apriori T is shown in the Figure 5.

![Figure 4. Preprocessing of log data](image)

![Figure 5. Sample of Generated Association Rules Using Fuzzy ARM](image)

VI. PERFORMANCE EVALUATION

This subsection briefly describes the general results of the Recommendation system. In order to evaluate the recommendation effectiveness, the performance is measured using three different standard measures, namely, precision, coverage and F1-measure, hit ratio.

Assume that transaction \( t \) is viewed as a set of pageviews, and a window \( w \) of \( t \) (of size \( |w| \) is used to produce a recommendation set \( R \) using the recommendation engine. Then the precision of \( R \) with respect to \( t \) is defined as:

\[
\text{Precision}(R, t) = \frac{|R \cap (t-w)|}{|R|} \tag{9.1}
\]
The coverage of $R$ with respect to $t$ is defined as:

$$Coverage(R, t) = \frac{|R \cap (t-w)|}{|t-w|} \quad (9.2)$$

Figure 7. Coverage Line Chart for 50 Sessions

Figure 6 and 7 shows Line chart for precision and Coverage of the sample 50 sessions, for further analysis $F1$-measure is calculated as

$$F1(R, t) = \frac{2 \times \text{precision} \times \text{Coverage}}{\text{precision} + \text{Coverage}} \quad (9.3)$$

In this context, precision measures the degree to which the recommendation engine produces accurate recommendations i.e., the proportion of relevant recommendations to the total number of recommendations. On the other hand, coverage measures the ability of the recommendation engine to produce all of the pageviews that are likely to be visited by the user. The $F1$ measure attains its maximum value when both precision and coverage are maximized. $F1$ can be viewed as a measure of similarity between $R$ set and $(t-w)$ set.

The effectiveness of the system is evaluated using Recommendation performance. The one more metric called hit ratio is used to measure the performance. From a given user session in the test set, extract the first $k$ pages as an active session to generate top N recommendations via the recommendation system using fuzzy association rule. Then compare the $(k+1)$ page of the testing session with recommendation list. if $(k+1)$ page is appeared in the recommendation set, then it is considered a hit. Count total number of hits and calculate the hit ratio by averaging it by the total number of testing sessions i.e. $\text{hitratio} = \frac{\text{hits}}{|T|}$ where $hit$ represents total number of hits and $T$ represent number of the testing sessions in whole test set. Figure 8 shows hit ratio of the current system for three different datasets.

![Hit ratio chart for three data sets](image)

VII. CONCLUSION

Generally Internet usage has seen tremendous growth. In the recent years, the number of Internet users globally increased from million to billion. Most of the percent of the world's population had access to computers with 1 billion searches every day. Reason behind this is the massive availability of information onto the World Wide Web has facilitated users in retrieving information. This huge availability data on the Web has made it necessary to find the ways to retrieve the information needed. Web usage mining is one of the ways to find and provide the necessary data. Web Usage Mining is used to discover the interesting user navigational patterns. These user navigational patterns are used to build the Web page recommender systems. Web page recommender systems predict users’ needful information and recommend the next page to the user to facilitate their navigation. Web usage mining uses different techniques to build the Web page recommender systems.

This work focused on the recommender system. This system firstly applies preprocessing on Web access logs and extracts semantic information from Web page contents. From preprocessed Web access logs fuzzy periodic attributes and fuzzy semantic attributes with fuzzy membership values are calculated. These fuzzy attributes are given as input to fuzzy ARM to generate association rules. Recommendations are generated by using these association rules. The experimental results have shown that this System can achieve effective periodic Web personalization. In Future this approach will be extended to provide real-time recommendation on portable Web-enabled devices that are becoming increasingly popular among consumers (e.g. mobiles, Smartphone).

REFERENCES


