Association Rule Mining for Web Recommendation

R. Suguna
Assistant Professor
Department of Computer Science and Engineering
Arunai College of Engineering,
Thiruvannamalai – 606 603
sugunarjasekar@yahoo.co.in

D. Sharmila
Professor and Head
Department of Electronics and Instrumentation Engineering
Bannari Amman Institute of Technology
Sathyamangalam- 638 401
sharmiramesh@rediffmail.com

Abstract - Web usage mining is the application of web mining to discover the useful patterns from the web in order to understand and analyze the behavior of the web users and web based applications. It is the emerging research trend for today’s researchers. It entirely deals with web log files which contain the user website access information. It is an interesting thing to analyze and understand the user behavior about the web access. Web usage mining normally has three categories: 1. Preprocessing, 2. Pattern Discovery and 3. Pattern Analysis. This paper proposes the association rule mining algorithms for better Web Recommendation and Web Personalization. Web recommendation systems are considered as an important role to understand customers’ behavior, interest, improving customer convenience, increasing service provider profits and future needs.

Key words - web mining; web logs; clustering; association rule mining; web recommendation

I. INTRODUCTION

In today’s world, information is growing much fast and rapid manner. It is challenging and interesting task to discover the interesting pattern and analyze them in proper manner [1][2][3]. The use of internet and World Wide Web is increasing in a dense manner. Everyday tremendous volumes of user browser detail are stored in the form of web log files in the web server. So, careful investigations on the web server log are important to analyze the user behavior and personalize the website which attract the user and satisfy the user needs in a fastest manner. But it is complex task to handle the numerous volumes of web logs without preprocess them [1][4].

The server logs are increased in the dense manner because every day number of users using the internet. The server logs are stored in the web server in the form of unformatted text files. It is too complex to manipulate the web logs with properly arrange them in some order. Preprocessing is applied in the web logs to reduce the volume of web log files and eliminate the unwanted data in the log files [5]. It is always better to group the web logs for applying any kind of operation. In data mining terminology, this grouping is called clustering [6][7]. Later, association rule mining is applied for fining the relative measure of the website users for better web recommendation and personalization. Recommendation systems are the effective way to connect consumers with products and services that they most likely would purchase based on their past actions and interests [8][9].

Recommender systems as a specific type of information filtering (IF) technique that attempts to present information items such as movies, music, books, news, images, web pages, etc. that are likely of interest to the user. The recommendations are generally based on an information item called the content-based approach or the user's social environment called the collaborative filtering approach.

The main four approaches for recommendations:

Personalized recommendation - recommend things based on the individual's past behavior.
Social recommendation - recommend things based on the past behavior of similar users.
Item recommendation - recommend things based on the item itself.
A combination of the three approaches above.
The proposed system is organized as follows. Section II deals with Related Works, Section III describes Preprocessing and Clustering of Web Logs and Section IV deals with Association Rule Mining. Section V describes Web recommendation and Personalization. Finally Conclusion and Future Work is given in Section VI.

II. RELATED WORKS

Researchers were always very enthusiastic to find out the efficient ways to execute the web usage mining because of the limited computing resources and the large data over which queries are executed.

The author Przemysław Kazienko (2009) has done the work on Classical association rules, mentioned his paper as “direct”, reflect relationships existing between items that relatively often co-occur in common transactions. In the web domain, items correspond to pages and transactions to user sessions. The main idea is to discover indirect associations existing between pages that rarely occur together but there are other, “third” pages, called transitive, with which they appear relatively frequently. Two types of indirect associations rules are described in the paper: partial indirect associations and complete ones. His new algorithm extracts complete indirect association rules with their important measure confidence which use the pre-calculated direct rules. Both direct and indirect rules are joined into one set of complex association rules, which may be used for the recommendation of web pages.

The authors Maja Dimitrijevic and Zita Bosnjak (2010) have focused on applying association rules as a data mining technique to extract potentially useful knowledge from web usage data. They conducted a comprehensive analysis of web usage association rules found on a website of an educational institution. The open source (Weka 3) data mining software was used for discovering association rules in web log data. However, Weka does not support web log mining in an efficient and natural way, while it is better suited for relational database mining.

The authors Maja Dimitrijevic and Zita Bosnjak (2011) have implemented a system for the discovery of association rules in web log usage data as an object-oriented application and used it to experiment on a real life web usage log data set.

The authors Huan Wu et al., (2009) have introduced a modified Apriori algorithm called, an Improved Apriori Algorithm IAA. IAA adopts a new count-based method to prune candidate itemsets and uses generation record to reduce total data scan amount. The new algorithm given better result than the original Apriori and some other existing ARM methods.

The author Huiping Peng (2010) done the research on finding the interesting association rules effectively from the Web logs after been preprocessed. The author proposed the FP-growth algorithm for processing the web log records, obtaining a set of frequent access patterns, then used the combination of browse interestingness and site topology interestingness of association rules for web mining, discovered a new pattern to provide valuable data for the site construction.

Kavita Das and O P Vyas (2011) have presented a model for web personalization approach using web mining. The server side and browser side details are taken for consideration. The author proposed bottom-up approach for achieving web personalization from personalized websites. The websites are personalizes for individual users by analyzing the user’s browsing history.

K R Suneetha and R Krithnamoorti (2011) have developed an Intelligent Recommendation System (IRS) to determine pages that are most likely to be visited by the user in future. IRS assists site owners in optimization, improving user satisfaction etc. The proposed IRS uses Classification and frequent pattern mining methods for recommendations.

III. PREPROCESSING AND CLUSTERING OF WEB LOG DATA

The preprocessing steps are considered as the initial process of the proposed approach. Secondly the preprocessed web logs are grouped to discover the pattern before applying the association rule mining to find the interesting measure of the web site visitors and users.

Most of the researchers [6][7] done their research on web usage mining as preprocessing is the initial step of their work. Later clustering techniques are applied on the preprocessed web logs to make a group for better processing of the web logs. The familiar clustering algorithms like k-means, modified k-means and Harmony k-means algorithms are used by some of the authors [6] [7][9].

In our previous work [27-28] we have proposed the effective algorithm for preprocessing and bird flocking algorithm [10] for clustering the web logs. The new algorithm effectively preprocesses the web logs which fit for the biological based algorithm called bird flocking algorithm. Since the web logs are growing in a rapid manner every day. So the web logs are dynamic nature. The bird flocking algorithm group the web logs in an efficient manner.
The web logs from various sources like server logs, browsers logs, etc as input to the Bird Flocking Algorithm. The aim of the proposed approach is to extract the user’s interest to visit particular web pages. The web log contains the data of websites visited user, which includes URL, web session duration, date, user activity duration etc. The web logs are updates each time a user starts a new session. In order to work with our algorithm the web logs has to formatted as to fit as boid. Initially the log file contains each and every detail regarding the user, the Ip address, website name, time stamp and other details. But these details are generated based on each and every second, so to make the log files light which we obtained from different sources, some preprocessing steps are first taken into action. According to the proposed approach, we define the boid as tuple of 5 values,

\[ b = \langle \text{ip}, \text{user}, \text{url}, \text{session}, \text{frequency} \rangle \]

Where, b is the representation of the boid, which has values, “ip” the ip address, “user” user name, “url” web address, “session” session duration of the user, “frequency” the number of visits by the user. There are many techniques by which we can reduce the density of log content in a log file. In this paper we are considering only five entities and they are Ip Address, user name, website name, session and frequency. The Extraction process of the session timing and the frequency is calculated by taking the time difference and the total number of clicks on a particular web site given in a log file.

To label the Session, we have calculated the time duration between two nearby website visited by the particular user. It is calculated each and every time when a user switches from one website to another and the amount of time he spends in each website.

\[
\text{session} = \sum \text{time}(\text{site}_i \rightarrow \text{site}_j)
\]

\[
\text{frequency} = \sum w_v
\]

The session is calculated as the time taken to traverse from on site to another site by the user, and the proposed approach take the whole sum of the duration of particular web site. The sum is taken as the total session duration collected for a website and in the frequency equation; \( w_v \) represents the visit of the user to a website \( w \).

**Bird Flocking Algorithm**

The bird flocking algorithm [10] is swarm intelligence algorithm based on the behavior of bird flock. The bird flocking was introduced by C. Reynolds as boids model. The following Algorithm shows the structure of a typical Boids model. A boid is the representation of a bird in the bird flocking algorithm. The main features affecting a boid are the speed and course of the flock. The speed and course of the boids are controlled Reynolds rules. Let us consider the basic algorithm defined by Reynolds.

Data: A group of boids.

**Algorithm 1. Bird flocking algorithm**

```
for each boid do
  Separation (boid);
  Cohesion (boid);
  Alignment (boid);
End

for each boid do
  boid.x ← cos(boid.course) * b.velocity * dT ime;
  boid.y ← sin(boid.course) * b.velocity * dT ime;
End
```

Reynolds have defined three rules for updating the speed and course of the boids, the rules are defined as, **Cohesion rule** is the rule that keeps the flock together, without it there would not be any flocking at all,
research efforts in Web personalization correspond to the evolution of extensive research in Web usage mining with the creation of new index pages. Most of the collected data include content-based filtering, collaborative filtering, rule-based filtering and Web usage mining for the determination of the actions that should be performed. The ways that are employed in order to analyze the categorization of these data (pre-processing phase), (c) the analysis of the collected data and (d) the mining of constraint association patterns. Here, we assign a weight to be associated with each page in a transaction to reflect interest of each page within the transaction. Some of the significance parameters are, (1) time of stay, (2) Quality rating. The associative patterns will be mined from the clustered data records by considering the significance of the web pages. Here, we assign a weight to be associated with each page in a transaction to reflect interest of each page within the transaction.

**IV. ASSOCIATION RULE MINING**

Association rule mining [10] is the technique of data mining, which is used to extract the interesting correlations, frequent patterns, associations among sets of items in the transaction databases or other data repositories. Many efficient association rule mining algorithms were proposed in the last few years [11][12][13][17]. But, the Apriori algorithm is commonly used by most of the researchers. Many researchers attempted to improve and optimize the efficiency of the Apriori algorithm [17][12].

Association Rule [10] is a one of the useful technique in data mining to find the relationships among the items present in large number of transactions. Given \( I = \{i_1, i_2, i_3 \ldots i_n\} \) is a set of items, a transaction may be defined as a subset of \( I \), and a dataset may defined as a set \( D \) of transactions. \( X \) and \( Y \) are non-empty subsets of \( I \).

The association rules are mainly defined by two metrics: support and confidence. The support of an item set \( X \) in a dataset \( D \), denoted as support \( D(X) \), is defined as \( \text{count} \ D(X)/|D| \), where \( \text{count} \ D(X) \) is the number of transactions in \( D \) containing \( X \). An itemset is said to be frequent (large) if its support is larger than a user-specified value (also called minimum support (\( \text{min} \) _sup_)). An association is an implication of the form \( \{X \rightarrow Y\} \), where \( X \cap Y = \emptyset \). Support \( S = \text{Number of sessions that contain A and B} / \text{Total number of sessions.} \)

Confidence \( C = \text{Number of sessions that contain A and B} / \text{Number of sessions that contain A}. \) Both support and confidence are fractions in the interval. The support is a measure of statistical significance, whereas confidence is a measure of the strength of the rule. The rule is said to be “interesting” if its support and confidence are greater than user defined threshold \( \text{Supmin} \) and \( \text{Conmin} \) respectively. There are two thresholds: \( P_\text{s} \) is a lower bound on the support of the rule and \( P_\text{a} \) is a lower bound on the accuracy of the rule.

**V. WEB PERSONALIZATION AND RECOMMENDATION**

Web personalization [4][14][15] is defined as any action that adapts the information or services provided by a Web site to the needs of a user or a set of users, taking advantage of the knowledge gained from the users’ navigational behavior and individual interests, in combination with the content and the structure of the Web site. The steps of a Web personalization process include: (a) the collection of Web data, (b) the modeling and categorization of these data (pre-processing phase), (c) the analysis of the collected data and (d) the mining of constraint association patterns. Here, FP-growth algorithm will be modified by incorporating the above constraints in its algorithmic procedure to mine constrained and useful associative patterns. **Recommendation**: In this phase, a usage model will be developed for predictions of users based on the mined constraint associative patterns. The proposed system will help the users to find their exact web pages that they want to visit.

**VI. CONCLUSION AND FUTURE WORK**

Many researchers were conducted research in the area of web personalization and web recommendation. In our previous work, we have done preprocessing the web log files and grouping the web logs based on the user’s interest. In this paper, association rule mining algorithm is proposed to find the user’s interest and for better web personalization and web recommendation.
References