Discovering Usage Patterns from Web Data Using Web Usage Mining

1L.Mohan  
Research Scholar  
Rayalaseema University, Kurnool, India  
lavmohan@gmail.com

2Dr.T.Venu Gopal  
Associate Professor  
JNTU Jogiyal, India  
t_venugopal@rediffmail.com

3B.Ranjitha  
Assistant Professor  
Adam’s Engineering College, Paloncha, India  
Ranjitha.bandi@gmail.com

Abstract - This paper work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. Web Usage Mining is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data. Web Usage mining involves mining the usage characteristics of the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc., in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users and is the main input.

Keyword – Web Data, Application Server, Web Mining.

I. INTRODUCTION

The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view, users, web service providers, business analysts. The users want to have the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users’ behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the user/consumers’ needs. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the problems encountered on the Web. Therefore, Web mining becomes a popular active area and is taken as the research topic for this investigation.

Web Usage Mining is the application of data mining techniques [3] to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses [6], page references, and access time of the users and is the main input to the present Research. This paper work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files.

A. Web Usage Mining

Web Usage Mining is a part of Web Mining, which, in turn, is a part of Data Mining [7]. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data [9], Web Usage mining involves mining the usage characteristics of the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc.

Typical sources of data:
- Automatically generated data stored in server access logs, referrer logs, agent logs and client-side cookies
- User profiles
- Metadata: Page attributes, content attributes, usage data

The main processes in Web Usage Mining are:

Preprocessing: Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

The different types of preprocessing in Web Usage Mining are:
• **Usage Pre-Processing**: Pre-Processing relating to Usage patterns of users.
• **Content Pre-Processing**: Pre-Processing of content accessed.
• **Structure Pre-Processing**: Pre-Processing related to structure of the website.

**Pattern Discovery**: Web Usage mining can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data.

The following are the pattern discovery methods. 
1. Statistical Analysis
2. Association Rules
3. Clustering
4. Classification
5. Sequential Patterns
6. Dependency Modeling

**Pattern Analysis**: This is the final step in the Web Usage Mining process. After the preprocessing and pattern discovery, the obtained usage patterns are analyzed to filter uninteresting information and extract the useful information. The methods like SQL (Structured Query Language) processing and OLAP (Online Analytical Processing) can be used.

---

**II. ASSOCIATION RULES**

The goal of mining association rules is to derive multi-feature (attribute) correlations from a database table [1]. A simple yet interesting commercial application of the association rules algorithm is to determine what items are often purchased together by customers, and use that information to arrange store layout[1]. Formally, given a set of records, where each record is a set of items, an association rule is an expression \( X \Rightarrow Y \); confidence; support. \( X \) and \( Y \) are subsets of the items in a record, \( support \) is the percentage of records that contain \( X + Y \), where as \( confidence \) is \[
\frac{support(X+Y)}{support(X)}
\]

**Apriori Algorithm:**

- \( L_k \): Set of frequent item sets of size \( k \) (with min support)
- \( C_k \): Set of candidate item set of size \( k \) (potentially frequent item sets)

\[
L_k = \{ \text{frequent items} \};
\]

\[
\text{for } (k = 1; L_k ! = \emptyset; k++) \text{ do}
\]

\[
C_{k+1} = \text{candidates generated from } L_k;
\]

\[
\text{for each transaction } t \text{ in database do}
\]

increment the count of all candidates in
Apriori Algorithm Module

The main goal of the proposed system is to identify usage pattern from web log files of a website. Apriori Algorithm[2] is used for this purpose. The main goal of the proposed system is to identify usage pattern from web log files of a website. For this purpose, the usage of apriori algorithm is proposed. Both are influential algorithms for mining frequent item sets for boolean association rules. In computer science and data mining [4], Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases [5] containing transactions (for example, collections of items bought by customers, or details of a website frequency).

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k - 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

The key concepts in this algorithm are:

- **Frequent Item sets**: The sets of item which has minimum support (denoted by L_i for i-th item set).
- **Apriori Property**: Any subset of frequent item set must be frequent.
- **Join Operation**: To find L_k, a set of candidate k-item sets is generated by joining Lk-1 with itself.

The advantages of using apriori algorithm are:

- Uses large item set property.
- Easily parallelized
- Easy to implement

This paper work is designed to operate on log files. The algorithm [9] attempts to find subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the item sets. The system operates in the following three modules.

- Preprocessing module
- Apriori Algorithm Module
- Association Rule Generation
- Results

The preprocessing module converts the log file, which normally is in ASCII format, into a database like format, which can be processed by the Apriori algorithm.

The second module is performed in two steps.

- Frequent Item set generation
- Rules derivation

This paper introduces the problem of “mining" a large collection of basket data type transactions for association rules between sets of items with some minimum specified confidence, and presents an efficient algorithm for this purpose. An example of such an association rule [10] is the statement that 90% of transactions that purchase bread and butter also purchase milk. The antecedent of this rule consists of bread and butter and the consequent consists of milk alone. The number 90% is the confidence factor of the rule.

A. Discovering Usage Patterns

This Paper focus in mining association rules among web usage data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered [8]. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users and is the main input to the present Research. This Research work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files.

III. SYSTEM DESIGN METHODOLOGY

A. Input Design

The input design is the process of converting user oriented input to a computer-based format. The goal of the input design is to make the data entry easier, logical and error free. In the present paper work, the input is taken from the web log file. The web log file has the extension .log and contains ASCII characters. A log file is a text file in which every page request made to the web server is recorded. For each request the corresponding log file contains the following information:
IP address of the computer making the request;
- User ID, (this field is not used in most cases);
- date and time of the request;
- a status field indicating if the request was successful;
- size of the file transferred;
- Referring URL, that is, the URL of the page which contains the link that generated the request; name and version of the browser being used.

This information can be used to reconstruct the user navigation sessions within the site from which the log data originates. In an ideal scenario, each user is allocated an unique IP address whenever an access is made to a given web site. Moreover, it is expected that a user visits the site more than once, each time possibly with a different goal in mind. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For instance, the format of the data for the discovery task may be different than the format necessary for mining sequential patterns. Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints.

B. Output Design

The main objective of any system is the generation of reports. It has various uses. Some of them are:
- For the users, reports provide source of information required.
- They provide permanent hard copy of the results of transactions.

Careful consideration has been given in the designing of the reports as it helps in decision-making process. In the present work, the performance of the system is judged using two metrics. The first one is the amount of memory used and the second one is the time taken for the algorithm to create the association rules.

IV. CONCLUSION

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. This Research work implements each of these phases. One of the algorithms which is very simple to use and easy to implement is the Apriori algorithm. This algorithm is used in the present Research work to generate association rules that associates the usage pattern of the clients for a particular website. The output of the system was in terms of memory usage and speed of producing association rules.

REFERENCES