ISSN No. 0976-5697

Volume 4, No. 5, May 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Association Rule Mining among web pages for Discovering Usage Patterns in Web Log Data

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

T.Venu Gopal[2]
[2]Associate Professor,JNTU Jagityal, India,
t_venugopal@rediffmail.com

Abstract: The objective of this project is to find the associations among different web pages in a web log file of a website, and then divide the page of each user accessed, who is likely to visit the Web site more than once and create individual sessions. This can be done by computing the association rules in relation to the web pages by using the information from the web log file maintained by its web server. Such information containing association rules gives a broad view of user sequences. It can be extended to any field of activity that involves large bulks of data. This paper is concerned with web usage mining of web mining. The server maintains the information of usage data sequences done by different users. The web log file has the information about the different pages that have been accessed by the users and the secondary information about each URL. Considering that information, in order to perform an association rule analysis, which is also known as market basket analysis, there's the need to define the basket; in the web environment this is not as clear as in a real supermarket. A phase of transaction identifiers is needed. Finally, after the application of one of these methods, the transaction file is created, the basket entities are defined and the discovery process of association rules can go on with the application of an Association Rule algorithm such as the Apriori one. To find the confidence among the web pages accessed, the user gives a user specified threshold value such as minimum support and minimum confidence values, which determines the probability of accessing a web page when a set of web pages are accessed.

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 [1] 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 [2], 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 [9] of websites from the server log files.

A. Motivation

Data Mining has attracted to great deal of attention in the information industry in recent years due to the availability of huge amounts of data and the imminent need for turning such data in to useful information and knowledge. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases. Active themes of research focus on the Association rule mining methods, the effective methods for mining large databases. Association rule mining is a challenging field of research where its potential applications pose their own special requirements. The following are typical requirements of Association rule mining in data mining:

- Scalability
- Ability to deal with large databases
- Efficient Discovery of Association rule
- Minimal requirements for domain knowledge to determine Input parameters
- Ability to deal with high dimensionality
- interpretability and usability

B. Web Mining

"It is the term of applying data mining techniques to automatically discovery and extract useful information from the World Wide Web documents and services". Data mining efforts associated with the web called web mining. Web data mining can be broadly categorized into three areas of interest based on which part of the web to mine.

Web content mining	: mining the content data on the Web
Web structure mining	: mining the Web structure data
Web usage mining	: mining the Web log data

Web Data Classification:

1) *Content data:* any complete or synthetic representation of the resource (the real data) such as HTML documents, images, sound files, etc.

2) *Structure data:* data describing the structure and the organization of the content through internal tags (intra-page) or hyper-links (inter-page).
3) *Usage data:* collection of available data describing the usage of web resources (e.g., access logs of web servers).

In web mining, data can be collected at the server side, client-side, proxy servers or a database. The information provided by the data sources described above can be use to construct several data abstractions, namely users, page-views, click-streams and server sessions.

C. Conclusion

Data mining: With the wide application of computers and automated data collection tools, massive amounts of data have been continuously collected in databases which pose a great demand for analyzing such data and turning them to useful knowledge. Consequently, data mining has become a research area with increasing importance. Data mining is the confluence of many fields such as Database, Statistics, and Artificial Intelligence etc.

Web mining: Web is used in several ways. We interact with the web for the following purposes:

- ➢ *Finding Relevant Information*
- ➢ *Discovering New Knowledge from the Web*
- ➢ *Personalized Web Page Synthesis*
- ➢ *Learning about Individual Users*

Web mining provides a set of techniques that can be used to solve the above problems. Sometimes, they provide direct solutions to above problems; on the other hand, they can be a part of a bigger application. Other related techniques form different research areas, such as databases information retrieval and natural language processing can also be used.

II. CONCEPT OF ASSOCIATION RULE AND APRIORI ALGORITHM

A. *Association Rule Discovery Techniques on Web Transactions*

Association rule discovery techniques are generally applied to databases of transactions where each transaction consists of a set of items. In such a framework the problem is to discover all associations and correlation among data items where the presence of one set of items in a transaction implies the presence of other items. In the context of web usage mining, this problem counts to discovering the correlations among references to varies files available on the server by a given client. Each transaction is comprised of a set of URL's accessed by a client in one visit to the server.

For example, using association rule discovery techniques we can find correlations such as the following.

- 40% of clients who accessed the web page with URL /company product1, also accessed /company/product2;or
- 30% of clients, who accessed/company/special placed an online order in /company/product1.

Since usually such transactions databases contain extremely large amount of data, current association rule discovery techniques try to prune he search space according to support for items under consideration. Support is a measure based on the number of occurrences of user transactions with in transaction logs.

Association Rules Basic Concepts On The Web:

Item = Web resource: When considering the process of discovering association *rules* from data, everyone comes up with the classical example of the supermarket and the basket. A market basket is filled of items. It is not a coincidence if the association rules analysis is also known as *market basket analysis*. On a Web domain, an item can be easily associated to a Web resource, that is to say an URL. However, the basket (or *transaction*) identification is not as easy as this one. "Unlike market basket analysis, where a single transaction is defined naturally and don't have a natural definition of transaction for the task of mining association rules [6]". When purchasing products in a supermarket, every transaction is defined, at the moment of paying the cashier. If a product has been picked up before another one (in the Web domain we have this information, giving the possibility of mining sequences of patterns), but have the exact definition of the basket.

In Web terms, don't know when a user leaves the site; and never warns a Web server that his *visit* has ended. A *visit* or *server session* is "a collection of user clicks to a single Web server during a user session" [10]. The only criterion for identifying a server session is to take notice that the client has not surfed through the site for a reasonably long time interval (e.g. 30 minutes).Thus, by examining next entries in the access log.

User =... : The identification of the user is important because the user is the character which creates a transaction, which chooses what items or resources go into the basket; therefore, in the preprocessing phase, the user identification is applied before the transaction one.

B. Association Rule

"Simply it's a relationship between two or more attributes" The goal of the mining association rules is to generate all possible rules that exceed some minimum user specified support and confidence threshold. The problem of deriving Association Rules from data was first formulated in [3] and is called the "market-basket problem". The problem is that we are given a set of items and a large collection of transactions which are sets (baskets) of items. The task is to find relationships between the containments of various items within those baskets. The task in Association Rules mining involves finding all rules that satisfy user defined constraints on minimum support and confidence with respect to a given dataset. Most commonly used Association Rule discovery algorithm that utilizes the frequent item set strategy is exemplified by the Apriori algorithm [3]. Apriori was the first scalable algorithm designed for association-rule mining algorithm. Apriori is an improvement over the AIS and SETM algorithms [4].

The *Apriori* algorithm searches for large item sets during its initial database pass and uses its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent item sets and those below are called small item sets. The algorithm is based on the large item set property which states: *Any subset of a large item set is large and if an item set is not large and then none of its supersets are large* [4].

Apriori Principle:

Any subset of a frequent item set must be frequent.

*Mining frequent item sets: the key step***:** The sets of items that have minimum support.

1. A subset of a frequent itemset must also be a frequent itemset.
2. i.e., if {A, B} is a frequent item set, both {A} and {B} should b a frequent item set.
3. Iteratively find frequent item sets with cardinality from 1 to k (k-itemset).
4. Use the frequent itemsets to generate association rules.

Apriori Algorithm:

- *Input:* Database D, of transactions; minimum support threshold, *min_sup*.
- Output: L, frequent itemsets in D.
- *Method:*

```
L1=find_frequent_itemsets(D):
for (k=2;Lk-1!=empty;k++) {
    Ck=apriori_gen(Lk-1,min_sup);
    for each transaction t ∈ D  {   //scan D for counts
        Ct=subset(Ck,t);             //get the subsets of t that are candidates
        for each candidate c ∈ Ct
            c.count++;
    }
    Lk={c ∈ Ck / c.count>=min_sup}
}
return L= Uk Lk;
```

Procedure: apriori_gen (Lk-1: frequent (k-1) itemsets; minsup: minimum support threshold)

```
for each itemset l1 ∈ Lk-1
    for each itemset l2 ∈ Lk-1 {
        if(l1[1] = l2[1]) ^ (l1[2] = l2[2]) ^….^ (l1[k-2] = l2[k-2]) ^  (l1[k-1] = l2[k-1]) then
            C=l1 (join) l2;           //join step: generate candidates
            if has_infrequent_subset(c,Lk-1) then
                Delete c;             //prune step: remove unfruitful candidate
            else add c to Ck;
    }
Return Ck;
```

Procedure: has_infrequent_subset(c: candidate k-itemset, Lk-1: frequent (k-1)-itemsets);
//use apriori knowledge

```
for each (k-1)-subset s of c
    if s does not belong to Lk-1 then
        return true;
return false;
```

Apriori Algorithm (working):

a. Test the support for items of length 1, called 1-itemsets by scanning the database. Discard those that do not meet minimum required support.
b. Extend the large 1-itemsets into 2-itemsets by appending one item each time to generate all candidate itemsets of length two. Again, test the support of all candidate itemsets.
c. Repeat the above steps; at step k, the previously found (k-1) itemsets extending into k-itemsets are tested for minimum support

Apriori employ breadth first search and uses a hash tree structure to count candidate item set efficiently. The algorithm generates candidate item sets of length k from k-1 length item sets. Then, the patterns which have an infrequent sub pattern are pruned. According to down ward closure lemma, the generated candidate set contains all frequent k length item sets. Following hat, the whole transaction data base is scanned to determine frequent item sets among the candidates. For determining frequent items in a fast manner, the algorithm uses a hash tree to store candidate item sets. This hash tree has items sets at the leaves and hash tables at internal node. Apriori is designed to operate on databases containing transactions (e.g., collection of items bought by customers of details of website frequentation). Other algorithms are designed for finding association rules in data having no transactions or having no time stamps.

III. THEORETICAL ANALYSIS AND TECHNICAL DESCRIPTION

Problem Definition

This project is concerned with web usage mining as discussed. The objective of project is to mine access patterns using web log file of a website. The data containing different user access patterns form the usage sequences done on the website in a particular week are collected. A web server maintains the information of usage sequences done by different users. That information is in the form of a web log file(record). The web log file has the information about the different items, that have been accessed by the users and the secondary information about each URL. This information obtained from the website kdd.ics.uci.edu/databases/msweb/msweb.html and by downloading the file anonymous_msweb_data.gz. This website is well known for its various knowledge discovery techniques. This site also has information about different usage sequences on a variety of applications that can be useful for providing insight into different fields of research.

Considering that information, find a set of association rules. Then find the confidence for each rule which determines the probability of accessing a web page when a set of web pages are accessed. Then those association patterns are to be applied onto that web server. The Web Usage Mining process proposed [5] becomes a major guide line upon project implementation**.** Fig.3.1 shows the general flow of the project methodology.

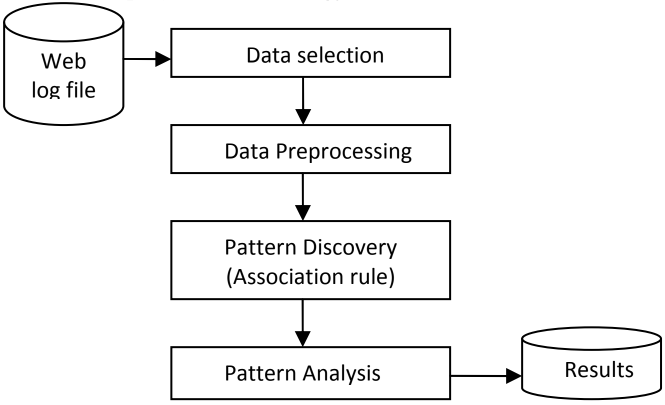

Fig 1: Project Methodology

A. Web Log File

The web log file has been selected for further analysis. The server log files are retrieved from the anonymous web data server, kdd.ics.uci.edu/databases/msweb/msweb.html and by downloading the file anonymous-msweb-data_gz. The total amount of the server log file between that duration is about 1.35 MB and the large amount of data becomes the most challenging problem to handle during the Data Preprocessing phase. The server log file consists of several attributes in the single line of record.

File Information: The input file anonymous_msweb_data.gz has two types of information. The information about different items accessed by the user is of the following format.

“A”, <web-id>,”web-page-title>”, <URL>

Example: A, 1287,"International AutoRoute","/autoroute",

- **‘A’** indicates that the line is an attribute.
- **<web-id>** refers to the webpage identity number. Each webpage has an identity number associated with it. That number is unique for that web page only. It is the representation of unique space occupied by it in the internet. The meaning of that webpage id is out of scope of the user.
- **‘1’** is ignored.
- **<webpage-title>** refers to the title of a webpage.
- **<URL>** refers to the URL accessed by that user.

“C” , “<user-id>” , “<user-id>”

“V” , “<web-id>” , “1”

Example: C, "10019" , 10019

- V, 1017, 1
- V, 1004 ,1
- V, 1018, 1
- V, 1029, 1
- V, 1008, 1
- V, 1030, 1

- **‘C’** indicates that it is a case line. It represents that a user has entered and accessed the set of pages.
- **<user-id>** represents unique user identity number. It may also represent those registered users for that website.
- **‘V’** indicates that the user with his registered number **<user-id>** has been accessing he webpage **<web-id>.**
- **‘1’** can be ignored.

B. Preprocessing

Data preprocessing phase is one of the most challenging phase in this study. The major task in this phase are includes handling missing values, identifying outliers, smooth out noisy data and correct inconsistent data [5]. Data Preprocessing consists of all the actions taken before the actual Pattern Analysis phase process starts. Since the data abstraction is very important in the data preprocess, it’s necessary to clarify the definitions of the related data abstractions before the description of the different type of the data converting.

C. Pattern Discovery

This is the key component of the Web mining. Pattern discovery converge the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. Once user transactions or sessions have been identified, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst, such as path analysis, discovery of association rules and sequential patterns and clustering and classification [7].This paper is related with association rule.

D. Pattern Discovery - Association Rules

In the Web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions [4]. The authors of [8] pointed that in the term of the Web usage mining, the association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. The support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. When loading a page from a remote site, association rules can be used as a trigger for pre-fetching documents to reduce user perceived latency. Given a server log files that represent anonymous web data file, the main purpose of Association Rules is to generate all Association Rules that have support and confidence greater than the user specified minimum support (called *min_sup*) and minimum confidence (called *min_conf*) respectively. An algorithm for finding all Association Rules, henceforth, referred to as the Apriori algorithm [3]. The selected of *Apriori* algorithm is because of the performance where it able to run the mining process in short period. Currently, *Apriori* algorithm is commonly used for generating the Association Rules for Web Usage Mining and this experimental study focus on exploratory of Web Usage Mining in web data portal (anonymous data file).

E. Pattern Analysis

During the Pattern Analysis phase, the descriptive method is being used analyze the data such as general summary of the Web usage and customer behaviors. This general summary includes the most active users using the portal from anywhere. The analysis also tries to find out the visitors for each facility or option that being provided by the Anonymous Web log File. The sever log files trace the information of documents that was downloaded. Finally filter out uninteresting rules or patterns from the set found in the pattern discovery phase.

Input: The input file anonymous_msweb_data.gz downloaded from kdd.ics.uci.edu/databases/msweb/msweb.html; an item file containing all user transactions; user specified threshold values: minimum support threshold, ***min_sup*** and minimum confidence threshold***, min_conf***.

Output– A file that contains all types of association rules and a confidence threshold value associated with each rule. Association rule contains frequent web pages that are accessed by the users.

IV. CONCLUSION

This paper has attempted to give an overview to the process of association rule mining from a Web server data. Due to the heterogeneous nature of the Web, it is almost impossible to get accurate knowledge from the common server data. Better results can be obtained by the use of dynamic Web sites, developed using server side programming languages with information stored in *ad-hoc* databases and with HTTP communications that take advantage of both temporary (session) and permanent (with a long life time) *cookies*. The contribution of the paper is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's sessions.

Web usage mining and data mining to find patterns is a growing area with the growth of Web-based applications. Application of web usage data can be used to better understand web usage, and apply this specific knowledge to better serve users. Web usage patterns and data mining can be the basis for a great deal of future research.

V. REFERENCES

[1] Tan, P. N., M. Steinbach, V. Kumar, .Introduction to Data Mining", Addison-Wesley, 2005, 769pp.

[2] R. Jin and G. Agrawal, "An Efficient Implementation of Apriori Association Mining on Cluster of SMPs," Proc. Workshop High Performance Data Mining (IPDPS 2001), Apr. 2001.

[3] Agrawal, R. and Srikant, R. (1994). Fast Algorithm for Mining Association Rules. Proc. of the 20th VLDB Conference. Pp 487-499

[4] Agrawal, R., Imielinski, T. and Swami, A. (1993). MiningAssociation Rules between Sets of Items in Large Databases. In Proceedings of the International ACM SIGMOD Conference, pages 207–216.

[5] Srivasta, J., Cooley, R., Deshpande, M., and Tan P. N. (2000). Web Usage Mining*:* Discovery and Application of Web Usage Pattern from Web Data.

[6] Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques*.* Morgan- Kaufmann Academic Press.

[7] Bamshad Mobasher, Namit Jain, Eui-Hong Han, Jaideep Srivastava (1996). *Web* mining: pattern discovery from World Wide Web Transactions.

[8] R. Cooley, B.Mobasher, and j.Srivastava. Grouping web page references into transactions for mining World Wide Web browsing patterns. Technical Report TR 97-021, University of Minnesota, dept. of Computer Science, Minneapolis, 1997.

[9] R. Cooley, B.Mobasher, and j.Srivastava. Web mining: information and pattern discovery on the World Wide Web. Technical Report TR 97-027, University of Minnesota, dept. of Computer Science, Minneapolis, 1997.

[10] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide Web browsing patterns. Knowledge and Information Systems, 1(1), 1999.