International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Web Usage Mining Using Association Rules

Dr.K. Prabha & T. Suganya
Assistant Professor
Ph.D Research scholar, Department of Computer Science
Periyar University PG Ext. Centre, Dharmapuri, Tamil Nadu, India.
Email:suganyacs59@gmail.com

Abstract

The explosive growth of Internet has given rise to many websites which maintain large amount of user information. The very important is to utilize this information, identifying usage pattern of users. The process of finding out this usage pattern and has many practical applications. The paper discusses how association rules can be used to discover patterns in web usage mining.Starts with preprocessing of the given weblog, followed by grouping them and finding the association rules. These rules help to improve website design, in advertising, web personalization etc. The web designers can restructure their web sites efficiently with the help of presence or absence of the association rules.

Index Terms— **Association rule hiding, Data mining, Privacy preserving data mining.**

I Introduction

One of the data mining tasks which can be used to uncover relationship among data is the association rule. Association rule identifies specific association among data and its techniques are generally applied to a set of transactions in a database. The amount of data handled is extremely large, current association rule techniques are trying to prune the search space according to support count.

Applications of association rule include health insurance, fraudulent discovery and loss-leader analysis, telecommunication networks market and risk management, inventory control etc., Association rule mining aims at extraction, hidden relation and interesting associations between the existing items in a transactional database. “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. 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.

Rules discovery finds common rules in the format A->B, meaning that, when page A is visited in a transaction, page B will also be visited in the same transaction. These rules may have different values of the confidence and support [3].

Confidence is the percentage between the number of transactions containing both items of the rule and the number of transactions containing just the antecedent. Support is the percentage of transactions in the rule is true.

In the context of Web Usage Mining, association rules refers to set of pages which are accessed together with a minimum support value which can help in organizing Web space efficiently.

For example: Consider if 70% of the users who accessed
get/programs/courses/x.asp also accessed
get/programs/courses/y.asp, but only 30% of those
who accessed get/programs/courses accessed
get/programs/courses/y.asp,
then it shows that some information in x.asp is
making the clients accessy.asp.

This inference helps the designers to decide on designing a link between the above two pages. The task of association rule mining has received a great deal of attention. Association rule mining is still one of the most popular pattern-discovery methods in KDD.

Hence, we would like to use association rules for pattern discovery analysis of Web Server Logs.

A. Web Server Log

Web Servers are used to record user interactions whenever any request for resources are received. A server log is a log file automatically created and maintains a history of page requests. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added [4]. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information [2].

B. Pattern Discovery

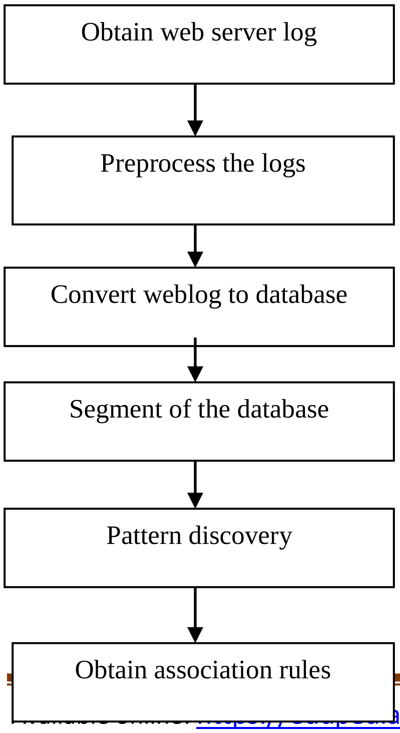
This is the key component of the Web mining. Pattern discovery converge the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. there are several kinds of access pattern mining that can be performed depending on the needs 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.

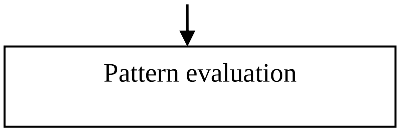
a. Design of Pattern Discovery

Flow Diagram

The flowchart for pattern discovery using association rules is given in fig 1[6].

Figure 1 Flow diagram for pattern discovery of weblogs





Each of these blocks is explained in detail as follows

1) *Obtain Web Server logs*

Web server log is a file which is created and maintained by the web server. We are analyzing the log file of the site: www.cs.depaul.edu. It is a text file. The file follows the extended log file format.

2) *Preprocessing the logs*

The weblog created by the web server contains details of all requests. It contains lot of irrelevant, incomplete data. Preprocessing involves removing such data.

3) *Conversion of log file to database*

The weblog cannot be directly used for data mining. The dataset is converted to a database. This involves creating a database and then importing the log file to the MySQL database table.

4) *Segmenting the database*

In this step, the database is segmented into clusters depending on the support count. After this a number of small clusters are obtained. Depending on the need, these clusters can be analyzed. Clustering web usage data allows the Web master to identify groups of users with similar behaviors for which personalized versions of the Web site may be created.

5) *Pattern Discovery*

The next step is pattern discovery. Once the clusters are formed they are studied to recognize patterns within the entries of the clusters.

6) *Association rules*

Association rules show relationship among different items. In case of Web mining, an example of an association rule is the correlation among accesses to various web pages on a server by a given client. Such association rules are obtained in this step.

7) Pattern Evaluation

The association rules obtained in the earlier step help in establishing relationships among data items. These association rules are evaluated to understand the information they provide. The interpretations of the rules provide useful knowledge.

II 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. [5]

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]
- **'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.

Sample log file converted into database is shown in table:[2]

Table 1

T no	Client IP	Date time	Method	Server IP	Port	URI Stem
0	202.185.122.151	1/23/2018 4:00:01 PM	GET	202.190.126.85	80	/index.asp
1	202.185.122.151	1/23/2018 4:00:08 PM	GET	n202.190.126.85	80	/index.asp
2	210.186.180.199	1/23/2018 4:00:10 PM	GET	202.190.126.85	80	/index.asp
3	210.186.180.199	1/23/2018 4:00:13 PM	GET	202.190.126.85	80	/tutor/include/style03.css
4	210.186.180.199	1/23/2018 4:00:13	GET	202.190.126.85	80	/tutor/include/detectBrowser_cookie.js

		PM				

III Web usage association rule interestingness measure

An interestingness measure is a function that assigns a value to each association rule, which corresponds to the web usage rule interestingness[5].

Various interestingness measures can be used when the association rules are discovered to qualify the potential interestingness of the association rules and highlight potentially interesting rules for the data analyst.

A. Confidence

Confidence is an interestingness measure of an association rule that refers to conditional probability of the rule consequent given the rule antecedent.

The rule $X \rightarrow Y$ holds in a set of sessions D with confidence c if c% of sessions in D that contain X also contain Y. If Supp denotes candidate support, the following formula can be used to calculate the confidence of a rule.

$$\text{Conf}(X \rightarrow Y) = \text{Supp}(X \cap Y) / \text{Supp}(X)$$

B. Lift

Lift is an interestingness measure of an association rule that compares the rule confidence to the expected rule confidence.

The expected confidence of a rule $X \rightarrow Y$ in a set of sessions D is the probability of the rule con-sequent Y in D. If the probability P(Y) is equal to the conditional probability P(Y|X), the item sets X and Y are not correlated in D.

If Supp denotes candidate support, the following formula can be used to calculate the lift of the rule $X \rightarrow Y$.

$$\text{Lift}(X \rightarrow Y) = \text{Conf}(X \rightarrow Y) / \text{Supp}(Y)$$

IV Web usage association rule mining

The problem of mining association rules is to generate a set of potentially interesting association rules in a data set of sessions that have support higher than the specified minimum support threshold and assign an interestingness value to all rules based on an interestingness measure.

A. Web usage association rule true interestingness

We consider a web usage association rule truly interesting if it is expected to cause a web master to take an action to change the structure of the website, based on the knowledge acquired through the association rule.

We define three categories of the rule true interestingness according to this Definition. The rules in the first category are those that are expected to cause a web master to take an action to change the website structure. The rules in the second category are those based on which a web master might take such action. The rules in the third category are those that are not expected to cause a web master to act.

B. Interestingness based on Confidence

We sorted the discovered association rules according to Confidence only and show how the top 10 rules are distributed over the three categories of true rule interestingness. Each column in Table 2 corresponds to an association rule algorithm run where the minimum support threshold is set to the specified value. Each cell in the table contains the number of rules that fit into the rule.

	Supp=0.007	Supp=0.0085	Supp=0.01
Expected	6	6	6
Might	4	3	3
Not expected	0	1	1

Table 2: True interestingness of Confidence

In Figure 2 we present the data in the histogram format, with the rule interestingness categories listed on the side. The vertical axis corresponds to the number of top 10 rules in each category.

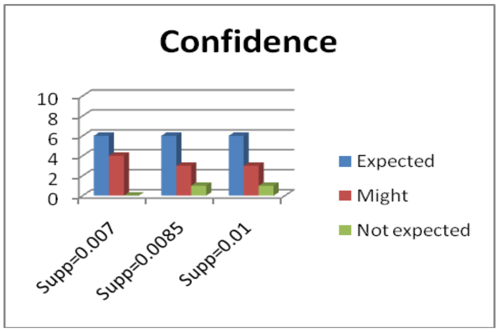


Fig 2

When the rules are sorted according to confidence, most of the top 10 rules are expected to, or might cause the web master to act in all three experiments with different support values.

C. Interestingness based on Lift

We sorted the discovered association rules according to Lift only and show how the top 10 rules are distributed over the three categories of true rule interestingness.

Each column in Table 2 corresponds to an association rule algorithm run where the minimum support threshold is set to the specified value. Each cell in the table contains the number of rules that fit into the rule interestingness category according to Definition 3.

Table 3: True interestingness of Lift

	Supp=0.007	Supp=0.0085	Supp=0.01
Expected	9	9	7
Might	1	1	3
Not expected	0	0	0

Figure 3 presents the distribution of the top 10 rules according to Lift, over the true interestingness categories. The vertical axis corresponds to the number of top 10 rules in each category.

Figure 3: Top 10 rules according to Lift

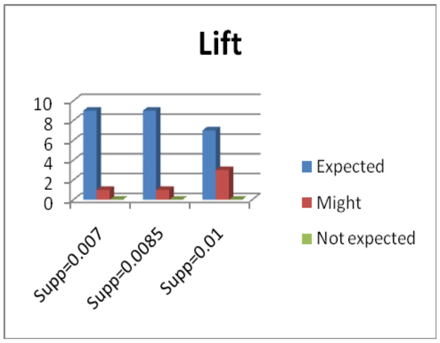


Fig 3

When the rules are sorted according to lift almost all of the top 10 rules fit into the first category of the true interestingness in all three experiments with different support values. There were no rules in the top 10, based on which a web master would not be expected to act in all three experiments.

D. Optimal interestingness measures

In our experiments, minimum confidence threshold is set to 0.4 and the rules with lower confidence are pruned out of the rule set prior to sorting all remaining rules according to either lift or confidence. We showed that Lift then gives better results than Confidence with respect to the true rule interestingness according to the Definition 3. Lift almost perfectly accurately measures interestingness in our experiments and no other measure needs to be considered.

However, in our initial experiments when the confidence threshold had been set lower (for example 0.2), lift alone did not perform that well. Neither Confidence nor Lift proved to be reliable in determining true interestingness of the association rules under these conditions.

We conclude that Lift and Confidence need to be considered in combination when deciding on the association rule interestingness. In this research we decided to combine them by assigning minimum confidence threshold value of 0.4 before applying Lift as the rule interestingness measure. While this way of combining Lift and Confidence gave excellent results in our experiments, it would be worth conducting experiments on other web log usage data sets in order to re-evaluate this method, or find other ways to use web usage association rule interestingness measures[6].

V SCOPE AND APPLICATIONS OF ASSOCIATION RULE

The user access log has very significant information about a Web server. A Web server access log contains a complete history of web pages accessed by clients. By analyzing these logs, it is possible to discover various kinds of knowledge, which can be applied to improve the performance of Web services. Web usage mining has several applications and is used in the following areas:

1) It offers users the ability to analyze massive volume of click stream or click flow data.

2) Personalization for user can be achieved by keeping track of previously accessed pages which can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages[6].

3) By determining access behavior of users, needed links can be identified to improve the overall performance of future accesses.

Web usage patterns are used to gather business intelligence to improve customer attraction, customer retention, sales, marketing, and advertisements cross sales. Web usage mining is used in e-Learning, e-Business, e-Commerce, e-Newspapers, e-Government and Digital Libraries.

VI. PROPOSED SYSTEM

Steps involved in the proposed system would involve the following steps.

1) The input is a set of Weblogs for which we have to find association rules. We have chosen University Web server logs from www.cs.depaul.edu site.

2) The server logs contain entries that are redundant or irrelevant for data mining tasks.

3) The Data cleaning process will select a subset of fields that are relevant for the task.

4) These selected attributes are then stored into a database.

5) Using a simple clustering approach these entries are divided into clusters or segmented.

6) Now, association rule mining is applied on these clusters, to obtain association rules having minimum support and confidence.

7) As a result of association rule mining, interesting patterns can be discovered and client's web usage can be evaluated.

VII CONCLUSION

Web usage mining is a kind of mining to server logs.The improvement of customer's relations and improving the requirement of system performance and so on. Web usage mining provides the support for the website design, providing personalization server and other business making decision, etc. Its main aim is getting useful to users for easy access information in logs to make sites perfect with effectual. Data preprocessing is the data pretreatment work.This experiments showed that interestingness measures can successfully be used to discovered association rules after applying the pruning method. Most of the rules that ranked highly according to the interestingness measures proved to be truly valuable to a web master.

REFERENCES

[1] Agrawal, R. and Srikant, R. (1994). Fast Algorithm for Mining Association Rules. Proc. of the 20th VLDB.

[2] Bamshad Mobasher, Namit Jain, Eui-Hong Han, Jaideep Srivastava (1996). Web mining: pattern discovery from www transactions.

3] M. Henri Briand, M. Fabrice Guillet, M. Patrick Gallinari, M. Osmar Zaaiane, "Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support", World Academy of Science, Engineering and Technology 48 2008.

[4] Mr. Sanjay Bapu Thakare, Prof. Sangram. Z. Gawali, "A Effective and Complete Preprocessing for Web Usage Mining", Expert Systems with Applications, 36(3), 6635-6644.

[5] Maja Dimitrijevic, Zita Bosnjak, "Web Usage Association rule mining system", International journal of information, knowledge and management.

[6] Ms Kiruthikrena, Rahul jaudhav, Dipa Dixit, Rashmi, Anjali Nehete, Trupti Khodar, "Pattern Discovery using Association rules", (IJACSA) Vol.2 No.12,2011.

[7] Mr.Mohan, Venu gopal, "Association rule mining among web pages for discovering usage patterns in web log data", (IJARCS) Vol.4, No.5 May 2013(Special issue).

[8] Shanthi, "Survey on Web Usage Mining Association Rule mining", (IJICSE) Vol. 4 issue 3 2017.