Constraint-based Web Log Mining for Analyzing Customers' Behaviour

Anagha Shastri
Assistant Professor
DBIT, Mumbai, India

Dipti Patil
Assistant Professor
MITCOE, Pune, India

V.M.Wadhai
Professor and Dean of Research,
MITSOT, MAE, Pune, India

ABSTRACT
Analysis of Web logs is one of the important challenges to provide Web intelligent services. Association rule mining algorithms are used widely to track users' web behaviour. Due to large amount of data many times the rules formed by these algorithms are very long and redundant. Recently Constraint-based mining approaches have received attention to deal with these big and redundant association rules. In this paper we discuss the Constraint based web mining approach used to reduce the size of association rules derived from Web log. The approach proves effective in reducing the overlap of information and also improves the efficiency of mining tasks. Constraint-based mining enables users to concentrate on mining their interested association rules instead of the complete set of association rules.

Keywords
Data Mining, Association rules, Constraint Based Web Mining

1. INTRODUCTION
Today, millions of visitors interact daily with Web sites around the world and massive amounts of data about these interactions are generated. Web sites are generating a big amount of Web logs data that contain useful information about the user behaviour. The term "Web Usage Mining" was introduced by Cooley in 1997 [1]. He defines Web mining as the "discovery and analysis of useful information from the World Wide Web". Web usage mining is the "automatic discovery of user access patterns from Web servers". The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals [2] or to improve the Web structure and Web server performance. Improving the Web page content and structure increases the number of visitors and thus increases sales and generates revenue. It helps analyze the system performance and network communications or even build adaptive Web sites. Different Web Mining approaches that exploit Web logs given in [3] are association rules, frequent sequences and frequent generalized sequences. Some commonly used data mining algorithms for Web usage mining are association rule mining, sequence mining and clustering [4]. The developed system for tracing user's buying pattern on web applies constraint based association rule mining. An association rule '*r'* is a relation between itemsets and an expression of the form $X \Rightarrow (Y - X)$, in which X and Y are items and $X \subset Y$. Association rules are used in order to reveal correlations between items accessed together in any transaction or during a server session. The problem of finding associations among itemsets in transaction databases is similar to problem of finding web pages visited together. A session in web log is same as a transaction in transaction database. Association mining in case of online purchase captures patterns relating to itemsets irrespective of the order in which they occur in a transaction. Some important Applications of association rules are Market Basket Analysis, Cross-Marketing, Catalog Design, Product assortment decision etc. Association rules are applicable whenever a customer purchases multiple things in proximity related to credit cards, services of telecommunication companies, banking services, medical treatments etc. Bit maps are used for transaction database representation. Statistical method is adopted to reduce the number of attributes. Interesting association rules are then derived on the basis of user defined support value. Section-2 discusses concepts of constraint based association rule mining. Section 3 describes the architecture of the web mining system. Section 4 gives implementation details of the system. Section 4 evaluates results of implemented system. Section 5 represents concluding remarks.

2. RELATED WORK

2.1 Web Mining
Web mining can be categorized into three different classes based on which part of the Web is to be mined. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining.

The aim of Web content mining is to provide an efficient mechanism to help the users to find the information they seek. Web content mining includes the task of organizing and clustering the documents and providing search engines for accessing the different documents by keywords, categories, contents etc.

Web structure mining is the process of discovering the structure of hyperlinks within the Web. Practically, while Web content mining focuses on the inner-document information, Web structure mining discovers the link structures at the inter-document level. A page having a lot of referencing hyperlinks means that the content of the page is useful, preferable and maybe reliable. Mining the structure of the Web supports the task of Web content mining.

Web usage mining, from the data mining aspect, is the task of applying data mining techniques to discover usage patterns. As every data mining task, the process of Web usage mining also consists of three main phases as shown in Figure 1.

(i) Preprocessing,
(ii) Pattern discovery and
(iii) Pattern analysis.

In the first phase, Web log data are preprocessed in order to identify users, sessions etc. In the second phase, statistical

methods, as well as data mining methods (such as association rules, sequential pattern discovery, clustering, and classification) are applied and interesting patterns are extracted. These patterns are stored so that they can be further analyzed in the third phase of the Web usage mining process.

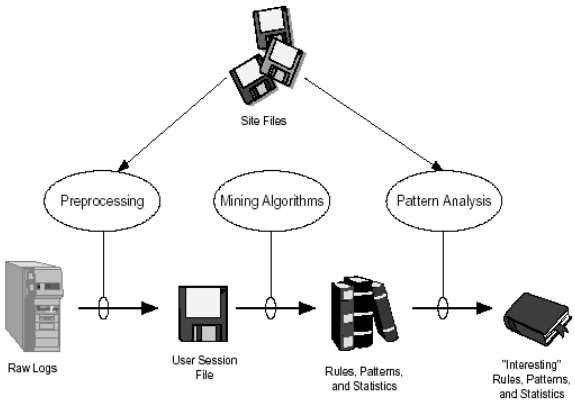

Figure 1: Process of Web Mining

2.2 Association Rule Mining

2.2.1 Association Rule Definition

It's the process of finding frequent patterns or associations within the data of some database or some set of log files.
Association rule takes the following form :

Body ⇒ Head [Support, Confidence]
e.g. buys(*X*, "diapers") ⇒ buys(*X*, "beers") [0.5%, 60%]
major(*X*, "CS") ^takes(*X*, "DB") ⇒grade(*X*, "A") [1%, 75%]

2.2.2 Database Representation

In the system developed Bitmap representation of the transaction database is used in context with association rule mining and the following conventions are considered.

$I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called items.
D be a database consisting of transactions.
|*D*| is no. of transactions stored in *D*.
t be a transaction such that $t \in D$ and $t \subseteq I$.

Each transaction *t* has a unique identifier called as TID. Presence of an item in transaction is represented by 1 and absence of item is represented by 0 in columns of a table. e.g. Table1 shows a database of transactions where the set of items is *I* = {A, B, C}. Each row in the table represents a transaction. There are three items and six transactions.

Given a database *D* of set of transactions, an association rule is an expression of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \phi$. The *X* and *Y* are called as ***body of the rule and head of the rule*** respectively. The support and the confidence are two parameters to measure the association rules.

Table1: Sample Database D with Bitmap Representation

TID	A	B	C
T1	1	1	1
T2	1	1	1
T3	1	0	1
T4	1	0	0
T5	1	0	0
T6	1	1	1

The support of the rule is the percentage of transactions that contains both *X* and *Y* in all transactions and is calculated as $|X \cup Y| / |D|$. The support of the rule measures the significance of the correlation between itemsets.

The confidence is the percentage of transactions that contain *Y* in the transactions that contain *X*. The confidence of a rule measures the degree of correlation between the itemsets and is calculated as $|X \cup Y| / |X|$. **The support is a measure of the frequency of a rule and the confidence is a measure of the strength of the relation between the sets of items.**

Table 2: Frequent Itemsets obtained from D

Itemset	Support
A	100%
B	50%
C	66%
AB	50%
AC	66%
BC	50%
ABC	50%

2.2.3 Two Step Process of association rule mining [5] :

i) Frequent itemset identification (Support as the criterion): Discover the large (frequent) itemsets that have transaction support above a predefined minimum threshold.

ii) Rule generation from frequent itemset (Confidence as the criterion): Use the obtained large itemsets to generate the association rules that have confidence above a predefined minimum threshold.

2.3 Constraint Based Mining

In addition to support and confidence other constraints such as *syntactic constraints*, may also be added to rules [5]. These constraints involve restrictions on items that can appear in the rule. *Syntactic constraints* are applied to filter out uninteresting items.

To improve the effectiveness and efficiency of mining tasks, constraint-based mining enables users to concentrate on mining their interested association rules instead of the complete set of association rules [6]. An approach to mine association rules with multiple constraints constructed by multi-dimensional attribute values [7] consists of three phases.

i) To collect the frequent items and prune infrequent items according to the Apriori property.

ii) To exploit the properties of the given constraints to prune search space or save constraint checking in the conditional databases.

iii) For each itemset possible to satisfy the constraint, generate its conditional database and perform the three phases in the conditional database recursively.

Constraint-based mining enable user to focus on a subset of transactions. It also eliminates generation of unwanted rules thereby saving processing time and resources. Thus, constraints can be used to optimize the algorithm used for mining. It can be done either by pushing constraints as deep as possible inside the frequent set computation or it can be applied after the frequent itemsets are generated. The constraints can be of two types : Data constraint and Rule constraint.

i) Data constraints e.g. SQL-like queries - find product pairs sold together on weekly holiday, or OLAP-like queries - in relevance to region, price, brand, customer category.

ii) Rule constraints e.g. specify the form or property of rules to be mined.

2.3.1 Rule Length and Item Constraint

Rules found by the Associations mining function can be controlled by various means. These include changing the values for minimum support, for minimum confidence, or for both. Other ways are also used to control results. These are [8] :

i) Maximum rule length constraint
This determines the maximum number of items that occur in an association rule. For example, maximum rule length is specified as 3 then association rules with no more than two items in the rule body and one item in the rule head will be obtained.

Example :

[Swimsuit][Beach towel] => [Sunglasses]

ii) Item constraint
This determines which rules are to be included in the results or excluded from the results. If you decide to *include* the specified items, the rules that contain at least one of the specified items are generated. Including specified items is called a *positive item constraint*. If specified item is excluded, the rules that contain any of the specified items are discarded from the results. Excluding specified items is called a *negative item constraint*.

Item constraints can be applied to the rule body, to the rule head, or to the complete rule. Positive and negative item constraints can also be combined that apply to the rule body, the rule head, or to the complete rule by using the boolean predicates `AND` or `OR`. For name mappings, the original item names or their name mapping to define item constraints can be used. You can specify the maximum rule length and the item constraints as part of defining mining settings.

While analyzing association rules, one has to read the rules and decide if they are:

i) Chance correlations: For example, two items were on sale at half price on the same day, and therefore were correlated by chance.
ii) Known correlations: For example, the paint and paintbrush correlation is something that is already known.
iii) Unknown but trivial correlations: For example, a correlation between red gloss paint and red non-gloss paint may be unknown but unimportant.
iv) Unknown and important correlations: For example, a correlation between paint and basketballs may be previously unknown. It may also be very useful in both the organization of advertising and product placement within the store.

3. SYSTEM DESCRIPTION

The architecture of the system is shown in Figure 2. In this project, the approach consists of four phases. They are as follows:

a. The web log is preprocessed and a transaction file is generated.

b. In second phase, instead of collecting the frequent items and pruning infrequent items are pruned according to the Apriori property using support threshold, user is given a choice of selecting products of his interest for sales promotion scheme. Thus user voting method is adopted to select condition attributes to prune search space.

c. Third, for each itemset possible to satisfy the constraint, its conditional database is generated and prediction algorithm is performed in the conditional database to find most beneficial decision attribute. The decision table is then created and rules are generated.

d. Lastly, support and confidence threshold values are applied independently to get the desired rule set. Rules thus generated are pruned and final rule set is displayed.

Constraint-based mining is employed to focus on the transactions of interest for a particular objective. Here the constraints used are of two types. i) ***Maximum rule length constraint*** to determine the maximum number of items that occur in an association rule. And ii) ***Item constraint*** to determine which items are to be included in the result. This eliminates generation of unwanted rules thereby saving processing time and resources and also improves effectiveness and efficiency of mining tasks. It is done by pushing constraints as deep as possible before the frequent set computation.

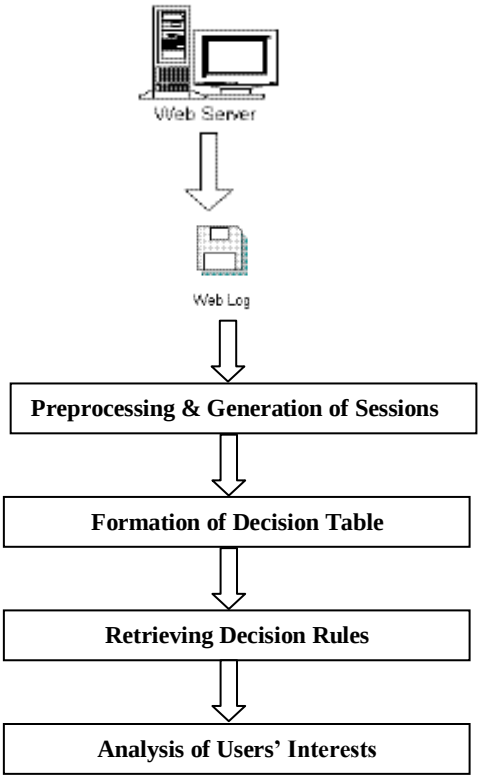

Figure 2: Overall System Architecture

4. IMPLEMENTATION

4.1 Session Generation & Data Preprocessing

i) Web log considered here consists of six parts : (a) Shop identification, (b) Access time, (c) User's IP address, (d) Session identification, (e) URL of the visited page and (f) Referrer.

ii) Visited page lists 21 categories of page accesses like product category, product sheet, list of brand names, online advice etc.

iii) Since the numbers of files were many, deciding on file naming conventions was an important task. A convention of creating a database for an hourly log was adopted and databases created thereby were used for statistical summary.

4.2 Preprocessing

The log tables were converted to multicolumn format as per the input requirement of the mining algorithm.

4.3 Creation of Decision Table

In conventional terminology while finding associations *items* are the units among which associations are identified whereas *groups* are units that contain items. Typically, groups contain the result of a single business *transaction* where several items are involved. Here, we have considered, an *item* is an individual article bought by a customer in a session and a *group* is the set of articles bought by a customer in one transaction.

User voting method is adopted to restrict the number of items that would be included in rules. Equivalence classes were generated by employing SQL queries.

The attributes were divided into two groups: condition attributes and decision attributes. The most beneficial decision attribute for the said selection of condition attributes was obtained by two approaches. In the first method count of each item was gathered for those items that were not present as condition attributes. TopCount method was used to display the suggested top five decision attribute. In second method, frequency of the items in decision set in only those transactions where a particular pattern was repeated was found. Thus decision attribute was predicted. The user was also allowed to select any other decision attribute of his/ her choice. The decision table was created based on condition attributes and a decision attribute selected by the user.

4.3.1 Algorithm for Predicting Decision Attribute

Input: A set of equivalence classes $L = (e_1, e_2, \ldots, e_k)$ set of equivalence classes
Output: A decision attribute for each pattern
Assumption: *A* be a set of attributes

A pair $S = (U, A)$ is an *information table*
L be a set of equivalence classes already created
k be number of equivalence classes for a particular condition selection condition attributes are already selected so attributes in decision set are known.

Begin
1. for each object e_i in L do
2. The transaction data that match each object is stored
3. for each attribute in R_Item
4. find count of all the attributes in decision set
5. put the counts in the same table R_Items
6. find topcount
7. display the topcount as decision attribute for e_i

end for
end for
End

These decision rules formed are then applied to trace the user behavior and its interest for further analysis.

5. RESULTS

Figure 3 shows snapshot of log file after preprocessing the web log data and session information.

ID	IP_Address	Digital_Cameras	Digital_Camcord	Zoom_Lenses	Flash_Lights
1	12.104.180.254	1	0	1	0
2	12.105.86.217	1	1	0	0
3	12.109.0.51	0	0	0	1
4	12.45.97.57	0	0	0	0
5	128.194.135.80	1	0	0	0
6	129.11.157.69	1	0	1	0
7	129.244.28.142	0	0	0	0
8	129.247.120.20	0	0	0	0
9	13.16.137.11	0	0	0	0
10	130.149.4.40	1	0	0	0
11	130.239.70.44	0	0	0	0
12	130.242.58.25	0	0	0	0
13	130.37.210.210	1	0	0	0
14	130.54.130.227	0	0	0	0
15	130.67.166.73	0	0	0	0

Figure 3: Processed Log File

This log file is used for generating the decision table. Figure 4 shows snapshot of decision table without putting any constraint while generating rules and figure 5 depicts the decision table after attribute reduction which makes rules formed more clear and simplified.

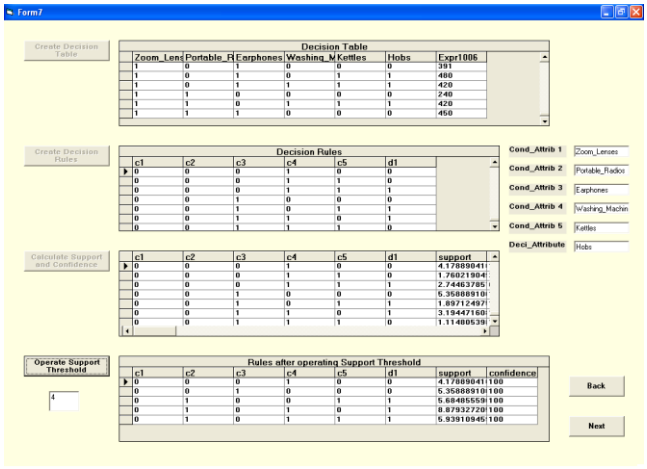

Figure 4: Decision Table and Support and Confidence for Rules

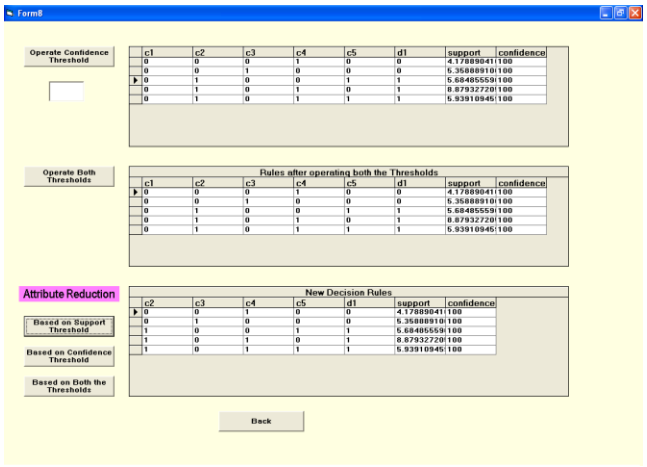

Figure 5: New Decision Table After Attribute Reduction

6. CONCLUSION

When customer database consume several gigabytes and contain millions of records, effective target marketing campaign or effective sales promotion strategy is required where constraint-based mining plays an important role. Constraint-based association rule mining eliminates generation of many useless association rules and improves the quality of association rule mining.

The method used in this project can reduce the attributes in both the condition and the decision granules. To set up a decision table, user voting is adopted on condition attributes. In particular, the generation of decision granules depends on the processed condition granules, where the uninteresting attributes and classes are removed by threshold. However, attribute reduction phase clearly depends on the selection of attributes by the user and it is also data dependent.

7. REFERENCES

[1] R. Cooley, J. Srivastava, and B. Mobasher, "Web mining: Information and pattern discovery on the World Wide Web". In *9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997.

[2] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization". In *ACM* Trans. *Inter. Tech.*, vol. 3, no. 1, pp. 1-27, 2003

[3] Mathia Gery, Hatem Haddad, "Evaluation of Web Usage Mining Approaches for User's Next Request Prediction". In *WIDM '03,* ACM November 2003.

[4] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining world wide web browsing patterns" In *Knowledge and Information Systems*, vol. 1, no. 1, pp. 5-32, 1999.

[5] Xin Chen and Yi-fang Brook Wu, "Web Mining from Competitors' Websites", Research track poster, New Jersey Institute of Technology

[6] W. Yang, Yuefeng Li, Yue Xu, Hang Liu, "Rough Set Model for Constraint-based Multi-dimensional Association Rule" In journal Software Engineering and Data Communication of *Queensland University of Technology, Brisban.*

[7] R.T.Ng, L.V.S. Lakshmanan, J. Han, A. Pang, "Exploratory Mining and Pruning Optimizations of Constrained Associations Rules". In Proceedings of ACM-SIGMOD, 1998, pp. 13-24.

[8] Rajendra K.Gupta and Dev Prakash Agarwal, "Improving the performance of Association Rule Mining Algorithms by Filtering Insignificant Transactions dynamically", Asian Journal of Information Management, pp.7-17. 2009 Academic Journals Inc.

[9] V.Umarani, Dr.M. Punithavalli, "A Study on Effective Mining of Association Rules from Huge Databases", IJCSR International Journal of Computer Science and Research, Vol 1 Issue 1, 2010