Journal of Computer Science 8 (6): 835-840, 2012
ISSN 1549-3636

Efficient Web Usage Miner Using Decisive Induction Rules

[1]Poongothai, K. and [2]S. Sathiyabama
[1]Department of Information Technology,
Selvam College of Technology, Namakkal, Anna University, Tamilnadu, India
[2]Department of the Computer Science,
Thiruvalluvar Govt. Arts and Science College, Rasipuram, Periyar University Tamilnadu, India

Abstract: Problem statement: Web usage mining is the technique of extracting useful information from server logs (user's history) and finding out what users are looking for on the Internet. This type of web mining allows for the collection of Web access data for Web pages. Scope: The web usage data provides the paths leading to accessed Web pages with prefrences and higher priorities. This information is often gathered automatically into access logs through the Web server. **Approach:** In this study we propose Induction based decision rule model for generating inferences and implicit hidden behavioral aspects in the web usage mining which investigates at the web server and client logs. The decision based rule induction mining combines a fast decision rule induction algorithm and a method for converting a decision tree to a simplified rule set. **Results:** The experimentation is conducted by weka tool and the performance of proposed Induction based decision rule algorithm is evaluated in terms of mined decisive rules, Execution time, root mean square error and mean absolute error. Proposed induction rule mining needs 400 ms of execution time for decisive rule generation, whereas previous work expectation maximization algorithm needs 600ms. **Conclusion:** Web usage mining is evaluated with decisive rules of user page navigation and preferences. Decisive rule provide the web site developers and owners to known the site presentation likeness and demands of the web users.

Key words: Decision rule, based decision, absolute error, mined decisive, execution time, root mean square, web pages, induction based

INTRODUCTION

The World Wide Web (WWW) is rapidly emerging as an important communication means of data related to a wide range of topics (e.g., education, business, government). It has created an environment of abundant user needs, where organizations should provide importance to enhance customer loyalty. To reorganize a website in terms of structuring links and attractive design of web pages, organizations must understand their user's behavior, preferences and future requirements. This imperative leads many vendors to design more e-service systems for data collection and analysis. The web usage patterns of users generally gathered by the web servers and stored in server access logs. Analysis of server access log data give information to restructure a web site to enhance effectiveness, better management of work group communication and to target ads to specific users.

Web mining (Yiming *et al*., 2011) is a popular method for analyzing customer's activities in e-service systems. It includes (i) Web Content Mining-discovering knowledge from the content of documents, (ii) Web Structure Mining extracting knowledge from Internet links (iii) Web Usage Mining-inferring interesting patterns from web access logs. Our work focuses on Web Usage Mining.

Web usage mining is a technique of data mining to extract usage patterns and behavior from Web log data. In general, Web usage mining can be classified into preprocessing, pattern discovery and pattern analysis. Preprocessing will process untreated site files and user profile data into page classification, site topology and server session files. Pattern discovery will process a server session file into rules, patterns and statistics information. Pattern analysis looks into the rules, patterns and statistics information obtained from pattern discovery of results that will be of interest to the management professionals.

Building accurate and efficient web data for large dataset is one of the essential tasks of web usage mining. In this study, we develop a new method, Induction based decision rule model for accurate and efficient extraction. This method focuses on extracting

Corresponding Author: Poongothai, K., Department of Information Technology, Selvam College of Technology, Namakkal, Anna University, Tamilnadu, India

decisive rules from web log data. Due to large amount of irrelevant information in the web log, the original log cannot be directly used in the web log mining procedure. Thus in the first step the web server log data is preprocessed, to extract useful data and then to map these data in to the abstract data necessary for pattern discovery. In this step, the original size of the database will be reduced. In another step, web data is classified using decision rule depending on splitting attributes.

Literature review: The rapid growth of e-commerce has made both business community and customers face a new situation (Yiming *et al*., 2011). Because of intending competition on the one hand and the customer's option to choose from several alternatives, the business community has realized the needs of intelligent marketing strategies and relationship management (Alessandro *et al*., 2011). Web usage mining attempts to extracts useful knowledge from the secondary data obtained from the interactions of the users with the Web (Eric *et al*., 2011). Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization and network traffic flow analysis and so on (Krol *et al*., 2008). The important concepts of Web usage mining and its various practical applications are presented in (Prasad *et al*., 2010). The approach called\intelligent-miner" (i-Miner) was presented in (Suneetha and Krihnamoorthi, 2009). I-Miner could optimize the concurrent architecture of a fuzzy clustering algorithm (to discover web data clusters) and a fuzzy inference system to analyze the Web site visitor trends.

Accurate Web usage information (AbuJarour and Awad, 2011) could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, track leaving customers and find the most effective logical structure for their Web space (Grace *et al*., 2011). User profiles could be built by combining users' web paths with other data features, such as page viewing time, hyper-link structure and page content (Susanne *et al*., 2011). What makes the discovered knowledge interesting had been addressed by several works (Tantan and Agrawal, 2011) and (Chitraa and Davamani, 2010). Results previously known are very often considered as not interesting. So the key concept to make the discovered knowledge interesting will be its novelty or unexpected appearance.

In this study, we develop a new method, Induction based decision rule model. The most popular classification method is the decision rule induction which builds a decision tree and performs classification on the given data using it .A decision tree is a tree in which each non-leaf node denotes a test on an attribute of cases, each branch corresponds to an outcome of the test and each leaf node denotes a class prediction (Rawat and Rajamani, 2010).

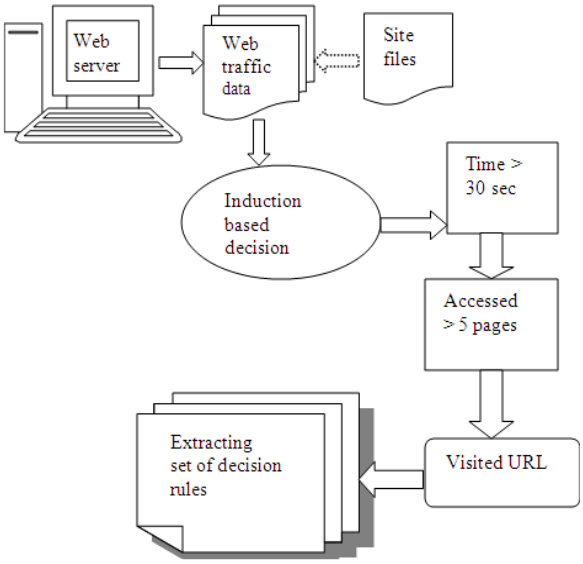

Fig. 1: Decision based induction rule mining on web data

Decision based induction rule algorithm for web usage mining: Web usage mining is attained initially by providing visitors traffic information based on Web server log files and other source of traffic data. Web server log files were used first by the webmasters and system administrators for the reasons of "how much traffic they are receiving, how many requests fail and what kind of errors are being produced". However, Web server log files can also record and trace the visitors' on-line behaviors. Web log file is one way to gather Web traffic data.

Figure 1 says that, after the Web traffic data is got, it may be joined with other relational databases, over which the data mining models are implemented. Through the data mining technique, Induction based decision rule model, visitors' behavior patterns are identified and interpreted.

Decision rules: Decision rules are used in classification and prediction. It is simple yet a powerful way of knowledge representation. The models generated by decision rules are represented in the form of tree structure. A leaf node indicates the class of the examples. The instances are classified by sorting them down the tree from the root node to leaf node. In our work we have used Induction based decision rule algorithm.

Induction based decision rule algorithm: Input: training samples, represented by discrete attributes; the set of candidate Attributes, attribute-list.

Output: Decision rules

Method:

Step 1: Create a node P

Step 2: If samples are the entire same cluster P, then Return P as a leaf node labeled with the cluster L
Step 3: If attribute list is empty then Return P as a leaf node labeled with the most common cluster in samples (majority voting)
Step 4: Select test attribute, the attribute among attribute-list with the highest information gain ratio
Step 5: Label node P with test-attribute;
Step 6: For each known value t_i of test-attribute
Step 7: Grow a branch from node P for the condition test-attribute = t_i
Step 8: Let d_i be the set of samples in samples for which test-attribute = a_i
Step 9: If d_i is empty then
Step 10: Attach a leaf labeled with the most common clusters in samples;
Step 11: Else attach the node returned by generate decision rule
Step 12: Update rules into knowledgebase

The information gain measure Induction based decision rule algorithm is used to select the test attribute at each node in the tree. Such a measure is referred to as an attribute selection measure of the goodness of split. The attribute with the highest gain is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessary the simplest) tree is found.

Decision rules:

R1: If (Time<30 secs and Pages <5) = "Not Preffered Users"
R2: If (Time <30 secs and Pages >5) = "Either Preffered or Not Preffered Users"
R3: If (Time >30 secs and Pages <5) = "Either Preferred or Not Preffered Users"
R4: If (Time >30 secs and Pages >5) = "Preffered Users"

Information gain: Let D be a set of training Dataset samples with their corresponding labels. If there are k classes and the training set contains d_i samples of class C and s is the total number of samples in the training set. Expected information needed to classify a given samples is calculated by Eq. 1:

$$C(D1, D2\ldots\ldots Dk) = \sum_{i=1}^{k} (Di/D)\ \log^2 (Di/D) \quad (1)$$

A feature Q with values { q1,q2,…qn} can divide the training set into n subsets {d1,d2,…dn} where dj is the subset which has the value qj for feature Q. Furthermore let Dj contain Dij samples of class i. Entropy of the feature Q is Eq. 2:

$$E(Q) = \Sigma(D1j + \ldots\ldots\ldots + Dkj) \sum_{J=1}^{n} D * C(D1j\ldots\ldots Dkj) \quad (2)$$

Information gain for Q can be calculated as Eq. 3:

$$\text{Gain}\ (Q)\ = C\ (D1,\ D2\ldots,\ Dk) - E(Q) \quad (3)$$

In our experiment, information gain is obtained for cluster labels by employing a binary discrimination for each cluster. That is, for each cluster, a dataset instance is considered in-cluster, if it has the same label: outclass, if it has a different label. Consequently, as opposed to measuring one information gain as a general measure on the relevance of the feature for all clusters, we calculate an information gain for each cluster. Thus this signifies how well the feature can discriminate the given cluster from other clusters.

Gain ratio criterion: The notion of information gain established earlier tends to favor attributes that have a greater number of values. For example if we have an attribute A that has a distinct value for each record, then Info (A, R) is 0, thus Gain (A, R) is maximal. To compensate for this, it was suggested to use the following ratio instead of gain.

SplitInfo is the information due to the split of R on the basis of the value of the categorical attributes A, which is defined by Eq. 4:

$$\text{SplitInfo}(X)\ = -\sum_{i=1}^{n} x\log 2 \frac{|Ri|}{|R|}\ \frac{|Ri|}{|R|} \quad (4)$$

And Gain Ratio is then calculated by Eq. 5:

$$\text{Gain Ratio}\ (A,R) = \frac{\text{Gain}\ (A,R)}{\text{SplitInfo}\ (A,R)} \quad (5)$$

The gain ratio, expresses the propotion of useful information generation split, i.e., that appears helpful for classification. If the split is near trivial, split information will be small and this ratio will be unusable. To avoid this, the gain ratio criterion selects a test to maximize the ratio above, subject to the constraint that the information gain must be large, at least as great as the average gain over all tests examined.

Performance evaluation: The experimental evaluation was conduced using car data sets of web traffic data. The data is in the original arff format used by Weka tool. The characteristics of the dataset used are given in the Table 1. Induction based decision rule algorithm is used for User Modeling in Web Usage Mining System.

Table 1: Car dataset used in the experiments

Country	Car	MPG	Weight	Drive ratio	Horse power	Displacement	Cylinders
U.S	Buick estate wagon	16.9	4.360	2.73	155	350	8
U.S	Ford Country squire wagon	15.5	4.054	2.26	142	351	8
U.S	Chevy malibu wagon	19.2	3.605	2.56	125	267	8
U.S	Chrysler lebaron wagon	18.5	3.940	2.45	150	360	8
Japan	Toyota corona	27.5	2.560	3.05	95	134	4
Japan	Datsun 510	27.2	2.300	3.54	97	119	4
Japan	Honda accord LX	29.5	2.135	3.05	68	98	4
Japan	Mazda GLC	34.1	1.975	3.73	65	86	4
Germany	Audi 5000	20.3	2.830	3.90	103	131	5
Germany	BMW 320i	21.5	2.600	3.64	110	121	4
Germany	VW rabbit	31.9	1.925	3.78	71	89	4
Germany	VW dasher	30.5	2.190	3.70	78	97	4
Sweden	Volvo 240 GL	17.0	3.140	3.50	125	163	6
Sweden	Saab 99 GLE	21.6	2.795	3.77	115	121	4
Italy	Fiat strada	37.3	2.130	3.10	69	91	4

All evaluation tests were run on a dual processor Intel CPU 2.5 GHz Pentium Core 2 Duo with 4GBytes of RAM, operating system Windows XP. Our implementations run on Weka tool, a data mining software for evaluation part of the system. In this study, there are two steps of data converting before applying induction based decision rule algorithm. There are around 800 URLs in car dataset. Assigning each URL address in the session to sequential numeric values is the first step. It is impossible to assign 800 attributes to Weka so for reducing the number of attributes; each eight sequence of attributes is assigned to one attribute based on bitmap algorithm.

MATERIALS AND METHODS

Decision based induction rule algorithm for web usage mining: By using Induction based decision rule algorithm, car dataset (website file) is evaluated. For example, a car-buying decision tree might start by asking whether customer want a 1999 or 2000 model year car, then ask what type of car, then ask whether customer prefer power or economy and so on. Ultimately it can determine what might be the best car for customer. Decision trees systems are incorporated in product-selection systems offered by many vendors. They are great for situations in which a visitor comes to a Web site with a particular need. But once the decision has been made, the answers to the questions contribute little to targeting or personalization of that visitor in the future.

RESULTS AND DISCUSSION

Figure 2 depicts the number of rules mined based on the number of transactions is done. As number of transactions increases, a mined rule also gets increased. Comparing with the existing EM algorithm, proposed Induction based decision rule algorithm mines the rule better.

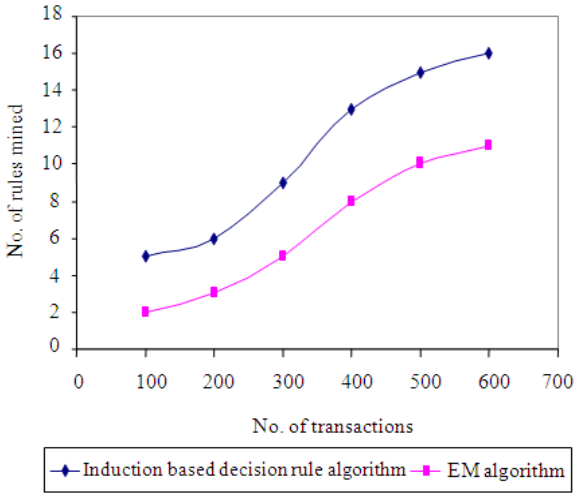

Fig. 2: Mining number of rules

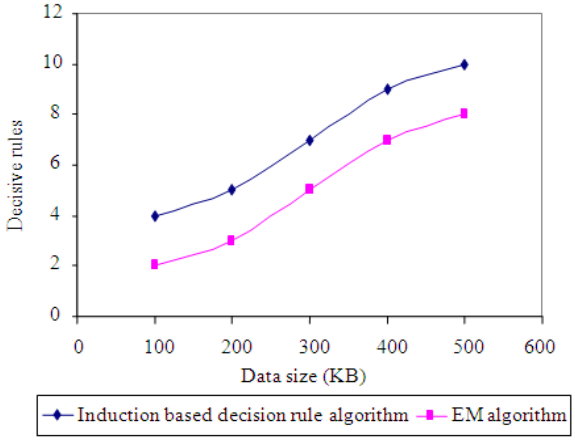

Fig. 3: Decisive rules

Figure 3 shows the number of decisive rules mined according to the data size is taken. Data size is taken in terms of Kilobytes (KB). Mined decisive rules are proportional to the data size.

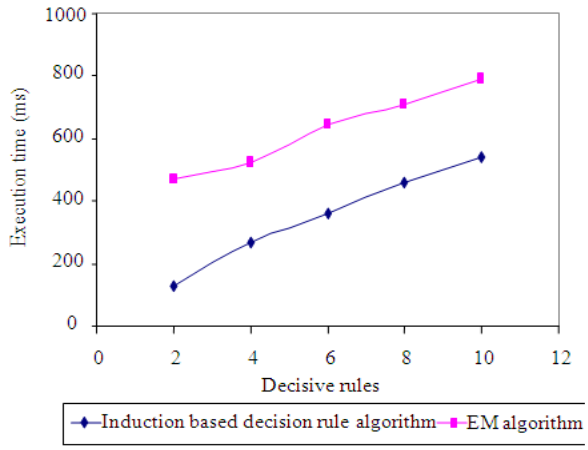

Fig. 4: Execution time

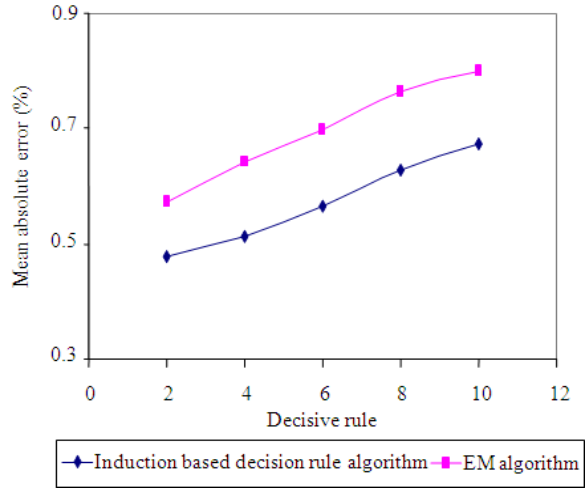

Fig. 5: Mean absolute error

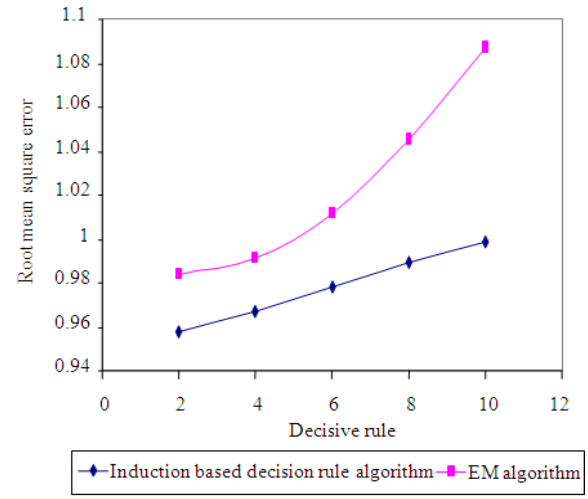

Fig. 6: Root mean square error

So when datasize increased, mined decisive rules also increased. Comparing with the existing EM algorithm, proposed Induction based decision rule algorithm performs well.

Figure 4 gives the comparative study of EM algorithm and Induction based decision rule algorithm based on the Execution time. Execution time is measured in terms of milliseconds (m sec). Induction based decision rule algorithm is having lesser execution time when compared with the EM algorithm.

Figure 5 gives the Mean Absolute Error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. As the name suggests, the mean absolute error is an average of the absolute errors $e_i = f_i - y_i$, where f_i is the prediction and y_i the true value.

The mean absolute error is relatively low in the proposed Induction based decision rule algorithm.

The root mean square is a statistical measure of the magnitude of a varying quantity. When two data sets, for instance are compared, the Root mean square of the pairwise differences of the two data sets can serve as a measure how far on average the error is from 0. Figure 6 shows that our proposed Induction based decision rule algorithm gives low error rate.

CONCLUSION

In this study we have implemented an Induction based decision rule algorithm to induce the knowledge that extracts the decisive rules in web usage mining framework. By linking the Web logs with cookies and forms, it is further possible to analyze the visitor behavior and profiles which could help an e-commerce site to address several business questions. Based on this web traffic data the induction process is done. The number of decisive rules is extracted by inducing only rules that are relevant to user's need. Relevancy is guided by query predicates. The performance of Induction based decision rule algorithm is measured in terms of number of decisive rules mined based on the number of transaction made and data size took, Execution time and finally, mean absolute error and root mean square error. The results indicate the Induction based decision rule algorithm can improve accuracy of rule extraction to 10-13% more.

REFERENCES

AbuJarour, M. and A. Awad, 2011. Discovering linkage patterns among web services using business process knowledge. Proceedings of the IEEE International Conference on Services Computing, July 4-9, IEEE Xplore Press, Washington, pp: 314-321. DOI: 10.1109/SCC.2011.54

Alessandro, B., M. Brambilla, S. Ceri and S. Quarteroni, 2011. A framework for integrating, exploring and searching location-based web data. IEEE Int. Compu., 15: 24-31. DOI: 10.1109/MIC.2011.136

Chitraa, V. and. A.S. Davamani, 2010. A survey on preprocessing methods for web usage data. Int. J. Comput. Sci. Inform. Security, 7: 78-83.

Eric, M., A. Bartoli, G. Davanzo and A. De Lorenzo, 2011. Automatic face annotation in news images by mining the web. Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Aug. 22-27, IEEE Xplore Press, Lyon, pp: 47-53. DOI: 10.1109/WI-IAT.2011.101

Grace, L.K.J., V. Maheswari and D. Nagamalai, 2011. Analysis of web logs and web user in web mining. Int. J. Netw. Security Appli., 3: 99-110.

Krol, D., M. Scigajlo and B. Trawinski, 2008. Investigation of internet system user behaviour using cluster analysis. Proceedings of the International Conference on Machine Learning and Cybernetics, July 12-15, IEEE Xplore Press, Kunming, pp: 3408-3412. DOI: 10.1109/ICMLC.2008.4620993

Prasad, G.S., N.V.S. Reddy and U.D. Acharya, 2010. Knowledge discovery from web usage data: A survey of web usage pre-processing techniques. Commun. Comput. Inform. Sci., 70: 505-507. DOI: 10.1007/978-3-642-12214-9_88

Rawat, S.S. and L. Rajamani, 2010. Discovering potential user browsing behaviors using custom-built APRIORI algorithm. Int. J. Comput. Sci. Inform. Technol., 2: 28-37.

Suneetha, K.R. and R. Krihnamoorthi, 2009. Identifying user behavior by analyzing web server access log file. Int. J. Comput. Sci. Netw. Security, 9: 327-332.

Susanne, G., I. Terrizzano, A. Lelescu and J. Sanz, 2011. Systematic web data mining with business architecture to enhance business assessment services. Proceedings of the Annual SRII Global Conference, March 29, IEEE Xplore Press, San Jose, pp: 472-483. DOI: 10.1109/SRII.2011.99

Tantan, L. and G. Agrawal, 2011. Active learning based frequent itemset mining over the deep web. Proceedings of the IEEE ICDE Conference, Apr. 11-16, IEEE Xplore Press, Hannover, pp: 219-230. DOI: 10.1109/ICDE.2011.5767919

Yiming, L., D. Xu, I.W.H. Tsang and J. Luo, 2011. Textual query of personal photos facilitated by large-scale web data. IEEE Trans. Patt. Anal. Mach. Intell., 33: 1022-1036. DOI: 10.1109/TPAMI.2010.142