Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 12 (14): 1495-1500, 2012
ISSN 1812-5654 / DOI: 10.3923/jas.2012.1495.1500
© 2012 Asian Network for Scientific Information

Integration of Clustering and Rule Induction Mining Framework for Evaluation of Web Usage Knowledge Discovery System

[1]K. Poongothai and [2]S. Sathiyabama
[1]Department of Information Technology, Selvam College of Technology, Namakkal, Tamil Nadu, India
[2]Department of Computer Science, Thiruvalluvar Government Arts and Science College,
Rasipuram, Tamil Nadu, India

Abstract: With the increasing popularity of WWW, some data such as users addresses or URLs which are being requested by the user are repeatedly collected by Web servers and collected in access log files. Investigating a server access data provides valuable information for performance improvement such as, reorganizing a web site for improving efficiency. Determining the path leading to accessed web pages which are often gathered into access log files is generally termed as web usage mining. In this connection, the concerned techniques mainly focus on the customer behavioral patterns discovered from a Web server log file in order to mine relationships within gathered data. The proposal in study, presents a novel framework, Integration of Clustering and Rule Induction Mining (ICRIM) which evaluate the performance of web usage knowledge discovery system. ICRIM framework incorporates the clustering model and Induction based decision rule model. The proposed evolutionary clustering model discovers web data clusters and analyzes the web site visitor behavior and optimally segregate similar user interests. Induction based decision rule model generates inferences and implicit hidden behavioral aspects in the web usage mining to investigate at the web server and client logs. Experimentation is carried out on ICRIM framework to evaluate the performance of web usage knowledge discovery system. Performance results of ICRIM framework are compared with existing clustering algorithm and induction based decision rule model.

Key words: Rule induction mining, clustering, web usage mining

INTRODUCTION

World wide web is rising at an incredible pace as an information gateway and acts as an inter-mediator for performing business. Web mining is categorized into three domains namely, content mining, structure mining and usage mining. Web content mining extracts information from the content present in the form of authentic web documents (text content, multimedia etc.). Web structure mining extracts the related information from the web structure, hyperlink references, etc. Web usage mining determines practical information from the secondary data obtained by way of communicating with the users through the web. Web usage mining has become essential for efficient web site administration, constructing adaptive web sites, commerce and sustained services, personalized and network traffic flow analysis.

Technique discussed earlier mainly focused on the customer behavioral pattern discovery from the Web server log file in order to mine relationships within gathered data. In this aspect, the user accessed pattern is presented and matched with one or more chronological pattern mined and navigational hints are further added to the pages accessed. Considering the growing number of web sites' visitors and the rapidity at which behavior of users' vary each and every day, the outcome attained by investigating an access log file loses its relevance, if it is oppressed too long.

In order to keep away from the obsolescence of the end result, Integration of Clustering and Rule Induction Mining (ICRIM) framework is proposed in this study which intends to obtain frequent behavior patterns, countering the web usage mining crisis in real time. Moreover, using ICRIM framework we discuss and solve several problems met by existing web usage mining techniques.

Whenever, the data is stored in the access log file, the occurrence of the pattern is repeatedly appeared, only during particular time period. This type of frequent pattern matches is not exhibited by the classical approach. When many rules correspond to the user's behavior, they are considered at the same level. The data mining process is

Corresponding Author: K. Poongothai, Department of Information Technology, Selvam College of Technology, Namakkal, Tamil Nadu, India

based on the resultant obtained from access log file. Smaller the number of clients obtained in the file rapid will be the data mining process.

One of the major problems faced in frequent item sets mining is the size of the pattern generated. Researchers have developed a handful of methods to provide solution to the problem. But the size obtained was so large that the algorithms required to mine identified larger number of solutions. As a result our proposal is based on the idea that as far as one uses the conceptual framework used in this study, the computing capacity on the network remains unused and the size of pattern generated remains unaffected.

Inductive learning method is one type of machine learning technique which is used to infer rules of classification by analyzing examples from a domain (Bozzon *et al*., 2011). The resultant knowledge obtained by learning techniques has various representation forms, including parameters in algebraic expressions, decision trees, formal grammar, production rules, formal logic-based expressions, graphs and networks (Nasraoui *et al*., 2008). As a means of classification or prediction, decision-tree induction techniques construct decision trees to discriminate among the various classes of objects. In contrast to neural networks, decision-tree induction techniques represent rules that can be readily expressed in English. As a result, humans can understand and easily map to the sets of rules (Medvet *et al*., 2011). Indeed, an inducted decision tree is a set of nested if-then statements.

New relational clustering techniques with robustness to noise were introduced (Liu *et al*., 2011) to discover user profiles that can overlap, while a density based evolutionary clustering technique was proposed (Abraham, 2003) to discover multi-resolution and robust user profiles. The K Means algorithm was used (Kerdprasop *et al*., 2008) to segment user sequences into different clusters. An extensive and in depth survey of different approaches to Web usage mining can be found (Glissman *et al*., 2011). It is interesting to note that an incremental way to update a Web usage mining model was proposed (AbuJarour and Awad, 2011). In this approach, the user navigation records are modeled by a Hypertext Probabilistic Grammar (HPG) whose higher probability generated strings correspond to the user's preferred trails. The model has the advantages of being self-contained as well as compact.

Decision-tree induction tools allow users to create decision trees and produce decision rules for both types of variables namely the continuous and the discrete target variables. In an inducted decision tree, each intermediate node represents a condition or a test and each leaf is assigned to a class or a vector of class probabilities. Instances or cases are typically represented as attribute-value vectors that give either numerical or nominal values of a fixed collection of properties (Chakrabarti, 2003).

A classification or regression tree is formed by successively partitioning a data set into discrete subgroups, based on one of the independent variables, until splitting is no longer feasible. The final result is a set of splits containing observations. Each split leads to one classification. The need to abstract information in an automated manner from large quantities of data emphasizes the need for the usage of machine-learning principles (Kerdprasop *et al*., 2008). Machine learning helps to acquire knowledge about a specific domain from available data in an automated manner.

Business-to-consumer electronic commerce is a newly emerging application. Internet stores can easily gather data regarding customer or sales data because they have built-in databases for customers, sales and inventory. With the help of tracking or monitoring techniques in a Web environment, Internet stores accumulate customer-behavior data in the Internet stores (Liu and Agrawal, 2011). Since Internet stores usually have many data sets, there is potential for applying machine-learning techniques. Knowledge acquired from learning techniques can be a valuable way to understand customers' online behavior in Internet stores and to gain competitive strength. The knowledge is utilized from an operational level to a strategic level in Internet store management.

ICRIM FRAMEWORK FOR EVALUATION OF WEB USAGE KNOWLEDGE DISCOVERY SYSTEM

Web has been a relentless generator of data that ranges in different forms, ranging from Web content data that contains the core substance of most Web documents, left by visitors as they surf through a Website, also known as Web usage data. Based on the hidden data, there exists certain interesting pattern. Using the hidden data certain objectives related to customer management and customization of user's website is achieved.

The ICRIM framework (Fig. 1) is used to extract usage patterns from Web log data. In ICRIM framework, clustering model segments the user sessions into clusters or profiles which in a later session form the basis for personalization. ICRIM works on the basis of modeling using Induction rule mining.

Discovering web data clusters using ICRIM framework: Discovering web data clusters using ICRIM framework presents the distinguished concepts of Web usage mining and its varied practical applications. Further a novel framework called Integration of Clustering and Rule Induction Mining is presented. ICRIM framework

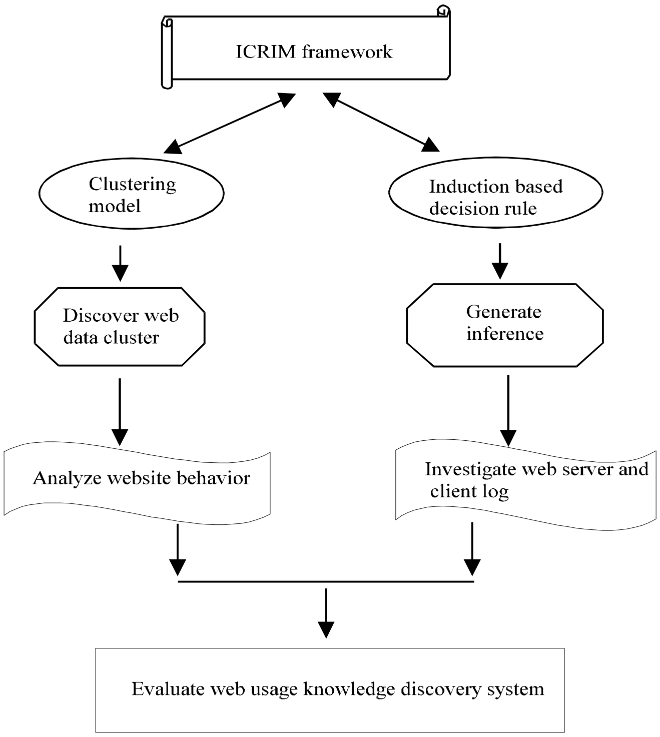
ICRIM framework
Clustering model
Induction based decision rule
Discover web data cluster
Generate inference
Analyze website behavior
Investigate web server and client log
Evaluate web usage knowledge discovery system

Fig. 1: Architecture diagram of ICRIM framework

analyzes the Web site visitor trends and optimally segregates similar user interests. Proposed approach is compared with hierarchical patterns (to discover patterns) and several function approximation techniques.

The ICRIM framework works on the basis of an evolutionary algorithm that optimizes the fuzzy clustering algorithm. The process starts with the pre-processing of raw data from the log files that are cleaned, pre-processed and number of clusters are identified using Fuzzy C Means (FCM) algorithm. The next step consists of analyzing the trend patterns using the clusters developed and the resultant is fed to the fuzzy inference system. The ICRIM framework uses evolutionary algorithm and back propagation algorithm. The if-then rule structures are determined by using learning procedure of an evolutionary algorithm and the rule parameters are fine tuned using a back propagation algorithm.

The optimization of clustering algorithm progresses at a faster time scale by using inference method. The number of cluster centers is initialized randomly and the role of FCM algorithm is to minimize the objective function which leads to local minima. No guarantee is ensured that FCM converges to an optimum solution. Henceforth an evolutionary algorithm is used to decide the optimal number of clusters and their cluster centers. The values are initialized by constraining them within the limit of vectors to be clustered.

ICRIM framework is used for clustering in the context of mixture models and estimates missing parameters of probabilistic models. Generally, this is an optimization approach which is performed iteratively using two steps. They are; (a) the cluster expectation step and (b) the maximization step. The cluster expectation step calculates the values expected for the cluster probabilities. The maximization step computes the distribution parameters and their likelihood of the data. The iteration process continues repeatedly until the optimization parameter reaches the fixed point or the parameters reach the maximum limit.

To simplify the discussion we first briefly describe the EM algorithm for ICRIM framework. The algorithm is proceeded until a desired convergence threshold value is achieved. In the context of mixture model, all attributes are assumed to be independent random variables. The mixture model for ICRIM framework consists of N probability distributions where each distribution represents a cluster. The process of algorithm proceeds in a manner that calculation for mean and standard deviation for ‘N’ clusters are determined. Then the mixture model for ICRIM framework evaluates the sampling probability ‘P’ for first cluster, ‘P’ for second cluster and so on until threshold value is reached. The next section discuss about mining induction rule using ICRIM framework.

Mining induction rule by ICRIM framework: Visitor traffic information is maintained by the web server log files and other source of traffic data. Using web server log files, diversified information such as the traffic received, request failure and error produced are measured and recorded and trace the on-line behaviors of users. After the receipt of web traffic data, they are 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. The models generated by decision rules by ICRIM framework are represented in the form of tree structure. The leaf node in ICRIM framework indicates the class of the examples. The processing continues by sorting them down the tree from the root node to leaf node.

PERFORMANCE ON ICRIM FRAMEWORK

In order to evaluate the resultant quality of the presented integration of clustering and rule induction mining framework experiments are conducted to show the resultant value. The experimental evaluation was conducted using UCI repository data sets of Syskill Webert and Car. The data is in the original ARFF format used by Weka tool. Integration of clustering model and rule induction mining model are used for user modeling in web usage mining system.

All evaluation tests were run on a dual processor Intel CPU 2.5 GHz Pentium Core 2 Duo with 4 GBytes 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 conversion before applying induction based decision rule algorithm. There are around 800 URLs in Syskill Webert and 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. The performance of ICRIM Framework is evaluated in terms of:

- Cluster precision
- Number of rules mined
- Mean absolute error
- Root mean square error

RESULTS AND DISCUSSION ON ICRIM FRAMEWORK

Evaluation is made for ICRIM framework by way of using Syskill Webert and Car data set. For example, in Car dataset, a car-buying decision tree might start by asking whether customer want a 2005 or 2010 model year car, then ask what type of car, then ask whether customer prefer power or economy or style and so on.

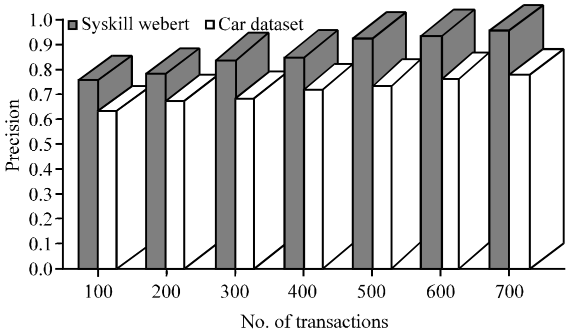

Fig. 2: Performance of cluster precision On ICRIM framework

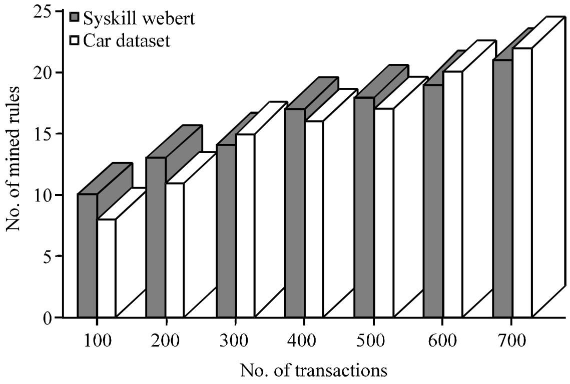

Fig. 3: Number of rules mined on ICRIM framework

Ultimately, it determines what might be the best car for customer. Decision trees systems are incorporated in product-selection systems and when applied to ICRIM framework provides with better results. The next section discusses about the results of ICRIM framework.

Figure 2 shows the performance result of ICRIM Framework in terms of Cluster precision using Syskill Webert and Car datasets. The results obtained from the graph show that precision value for each transaction of Syskill Webert dataset is higher than the Car dataset. The above figure also indicates that the precision rate even increases for increasing number of transactions for both the data sets.

Figure 3 illustrates the number of rules extracted using ICRIM Framework with the help of Syskill webert dataset and car dataset. As the number of transactions increases the rule mined also gets increased in both the datasets. Comparing both datasets, with the transaction size as 100, 200, 400 and 500 Syskill Webert dataset extracted higher number of rules whereas the transactions in 300, 600 and 700 shows better results for Car dataset.

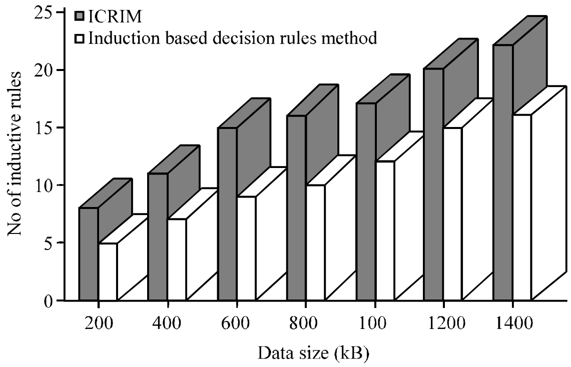

Fig. 4: Comparative result of inductive rules on proposed ICRIM framework against induction based decision rule method

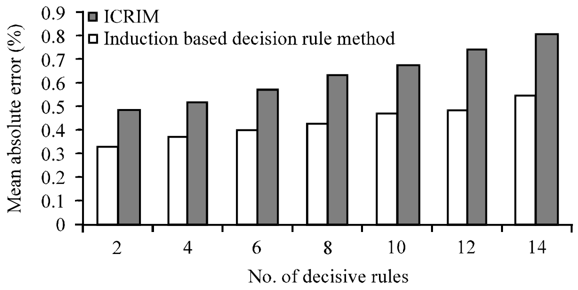

Fig. 5: Measure of mean absolute error-ICRIM

Figure 4 depicts comparative result of proposed ICRIM Framework against Induction based decision rule method. The figure shows the result of number of rules mined with the data size ranging from 200 to 140 kB. If given data size increases, the rules discovered also gets increased gradually.

The mean absolute error in ICRIM measures the average magnitude of the errors for the number of visitors visited in a website considering their direction. The mean absolute error for ICRIM presents with a linear score in which all individual differences are weighted equally in average. Figure 5 shows that our proposed ICRIM framework has low mean absolute error compared with existing induction based decision rule method.

The root mean square error for ICRIM framework consider quadratic scoring rule which measures the average magnitude of the error. In ICRIM framework, the square root value of the average is taken. As the errors are squared before they are averaged, the root mean square error for ICRIM framework gives relatively high weight to large errors. From Fig. 6, we can say, root mean square error is relatively low in ICRIM Framework compared to the counterfeit model Induction based Decision rule method.

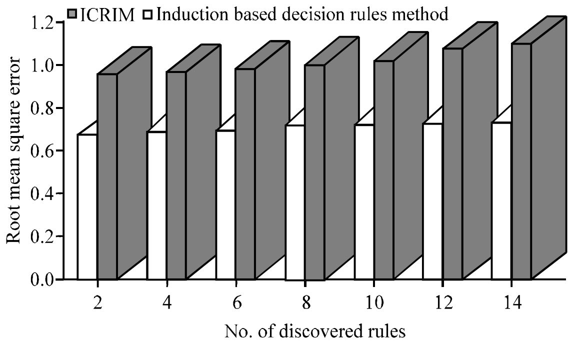

Fig. 6: Root mean square error-ICRIM

As a whole the proposed ICRIM Framework performs better in the evaluation of web usage knowledge discovery system. From the Performance results, ICRIM Framework gives 15 to 26% better result in Mean Absolute Error and 18 to 27 % better result in Root Mean Square Error.

CONCLUSION

In this study, we have implemented a new framework, Integration of Clustering and Rule Induction Mining (ICRIM) that evaluate the performance of web usage knowledge discovery system using integrated clustering model and Induction based decision rule model. Web data clusters and behavior of web site visitor is analyzed and similar user interests are segregated optimally. Whatever be the number of clients, ICRIM framework is designed in such a way that the inferences generate behavioral aspects of web usage mining at the web server and client logs. Experimentation is carried out using ICRIM framework to evaluate the performance of web usage knowledge discovery system. Performance results of ICRIM framework are compared with existing clustering algorithm and Induction based decision rule model. It gives 15 to 26% better result in Mean Absolute Error and 18 to 27% better result in Root Mean Square Error.

REFERENCES

Abraham, A., 2003. Business intelligence from web usage mining. J. Inform. Knowl. Manage., 2: 375-390.

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, 2011, Washington, DC USA., pp: 314-321.

Bozzon, A., M. Brambilla, S. Ceri and S. Quarteroni, 2011. A framework for integrating, exploring and searching location-based web data. IEEE Internet Comput., 15: 24-31.

Chakrabarti, S., 2003. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers, California.

Glissman, S., 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-April 2, 2011, San Jose, CA USA., pp: 472-483.

Kerdprasop, N., N. Muenrat and K. Kerdprasop, 2008. Decision rule induction in a learning content management system. Proceedings of the World Academy of Science, Engineering and Technology, August 13-15, 2008, Vienna, Austria.

Liu, T. and G. Agrawal, 2011. Active learning based frequent itemset mining over the deep web. Proceedings of the IEEE 27th International Conference on Data Engineering, April 11-16, 2011, Hanover, Germany, pp: 219-230.

Liu, Y., D. Xu, I.W.H. Tsang and J. Luo, 2011. Textual query of personal photos facilitated by large-scale web data. IEEE Trans. Pattern Anal. Machine Intell., 33: 1022-1036.

Medvet, E., 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, August 22-27, 2011, Lyon, France, pp: 47-54.

Nasraoui, O., M. Soliman, E. Saka, A. Badia and R. Germain, 2008. A web usage mining framework for mining evolving user profile in dynamic web sites. IEEE Trans. Knowl. Data Eng., 20: 202-205.