Review on Classification of Web Log Data using CART Algorithm

Jagriti Chand
Computer Science & Engg.
NIIST Bhopal

Abhishek Singh Chauhan
Computer Science & Engg
NIIST Bhopal

Ashish Kumar Shrivastava
Computer Science & Engg.
NIIST Bhopal

ABSTRACT
Web Usage Mining (WUM) is the process of extracting knowledge from Web user behavior on web, who are actively involved in accessing the web data by exploiting Data Mining techniques. This knowledge can be used for various purposes such as personalization, system development and site improvement. This knowledge discovery is also useful for web designer to quickly respond to the web user needs. The researcher has developed no of classification technique to extract the user log data to find the interested web user. This paper reviews the various existing classification techniques for web user data mining and presents a comparative analysis of these classification techniques.

Keywords
Web Mining, Data Mining, Classification, Classification Algorithm, CART, Naive Bayes.

1. INTRODUCTION
The ways in which users interact with a World Wide Web (Web) site provide enormous data processing on the usefulness and effectiveness of Web design elements and content built in it. Yet the informative log files captured by Web servers, and client logs, provide potentially important data about users Web site interactions. These data may be segregated and studied to generate inferences about Web site is used to design, test prototypes of Web sites or their modifications over time, and to test hypothetical hypotheses about the effects of different design variables on Web user behavior.

Data mining is the process of extracting useful information from databases. Many approaches to temporal data mining have been proposed to extract useful information, such as time series analysis, temporal association rules mining, and sequential pattern discovery. Several core techniques that are used in data mining describe the type of mining and data recovery operation.

1.1 Classification
Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may aspiration to use categorization to predict whether the weather on a particular day will be "sunlit", "wet" or "hazy". Popular categorization techniques include decision trees and neural networks. In this paper proposed classification algorithm of web log data and studying the interested users from them. Due to the involvement of uninterested users in the web log, the original log cannot be used as a process in the web usage mining procedure. Thus in the first phase the web log data is preprocessed, to extract the interested data and then to proceed with the extracted data. During this phase, the actual size of the database will be minimized to certain extent. The second phase consists of segregating the data using CART algorithm.

1.2 Web Mining Subtasks
Web mining is used to understand customer performance, estimate the usefulness of a specific Web site, and help measure the success of a marketing crusade. Web mining can be decayed into the subtasks, like as.

a) Resource finding: The task of retrieving intended Web credentials. By resource judgment we mean the procedure of retrieving the data that is either online or offline from the text sources available on the web such as electronic newsletters, electronic newswire, the text inside of HTML documents obtained by removing HTML tags, and also the physical selection of Web resources.

b) Information selection and pre-processing: Automatically selecting and pre-processing specific information from retrieved Web resources. It is a type of alteration processes of the original data retrieved in the IR process. These alteration could be either a kind of pre-processing that are mentioned above such as discontinue words, twigging, etc. or a pre-processing intended at obtaining the desired representation such as finding phrases in the training amount, transforming the illustration to relational.

c) Generalization: It automatically discovers general patterns at individual Web sites as well as across multiple sites. Machine knowledge or data mining techniques are typically used in the process of simplification. Humans play a vital role in the information or knowledge discovery process on the Web since the Web is an interactive medium [1].

1.3 Web Mining Approaches
Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and effectively. The third approach is to find out the information which can be derived from the activities of users, which are stored in log files for example for predictive Web caching. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined. These three categories are shown [2].

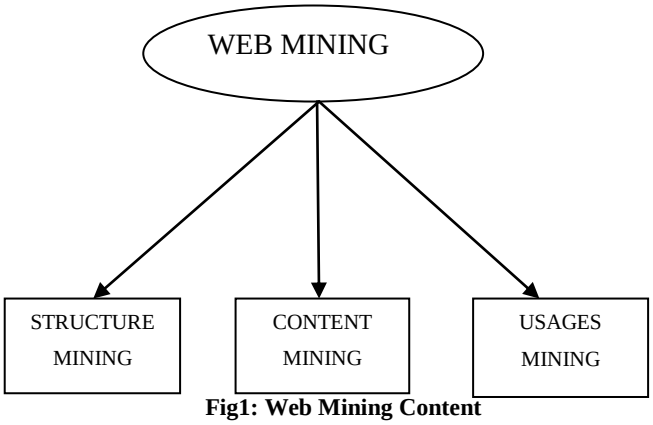

Fig1: Web Mining Content

2. BACKGROUND

Web usage mining involves with the application of data mining methods to discover user access patterns from web data. The main task of web usage data is to capture web browsing behavior of users from a specified web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is web log data, which maintains the information regarding the user navigation. This work concentrates on web usage mining [3].

3. RELATED WORK

K. Santra, S. Jayasudha, CARE School of Computer Applications, Trichy, present a Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification [3]. Objective of this paper is to review the behaviour of the interested users instead of spending time in overall behaviour. The present model used increased version of decision tree formula C4.5. During this we tend to propose to use the Naive Bayesian Classification formula for classifying the interested users and also we present a comparison study of using enhanced version of decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying. Classification of web log data using naïve Bayesian method is one of the well-known approaches that improve the overall performance of the web server. Naive Bayesian Classification shows good result in the improvement in time and memory utilization, it can be applied to any web log files. From the experiments conducted, many attributes are not used for classifying as they are irrelevant. We have given a mathematical evaluation of maximum likelihood for the Naive Bayesian Classification which provided a more efficient implementation with a performance increase compared to enhanced C4.5 decision tree. This method can be used in e-commerce applications, such as Web Caching, Web page proposal, and Web personalization.

G.Sathyadevi Department of Computer science and Engineering Anna University of Technology, Tiruchirapalli, proposed application of cart algorithm in Hepatitis disease diagnosis[4] motivated by domain-driven data mining, paper attempts to maximize the utility of domain experts (oracles) in active learning process. This study uses CART Algorithm to examine hepatitis disease diagnosis. From the given training datasets, only relevant attributes are selected using decision tree algorithm CART. The missing values in the given datasets can be easily handled by CART algorithm. Identification and selection of relevant attributes that Contribute to hepatitis disease is a challenging task. Thus, in this paper, our empirical Studies with UCI hepatitis patient datasets show that the proposed active learning algorithm is more effective than the other state-of-the-art algorithms for active learning. We propose a CART decision tree Algorithm against the biomedical hepatitis patient datasets and compare the results with other data mining techniques. Among these algorithms, CART algorithm always generates a binary decision tree. That means the decision tree generated by CART algorithm has exactly two or it has no child. However the decision tree which is created by other two algorithms may have two or more child. Also, in respect of accuracy and time complexity CART algorithm performs better than the other two algorithms. Dynamic queries and their certain answers from ORACLE are examined using the learning model.

S.Vijayalakshmi, Dr.V.Mohan studied a wide range of methods of Extracting Sequential Access Pattern from Pre-processed Web Logs [5]. Sequential pattern mining is an important data mining problem with broad applications, counting the analysis of client purchase behavior, patterns of web access, scientific experiments, and treatment of disease, natural disaster, and protein formations. In this paper, they expand the horizon of frequent sequence pattern mining by analyzing an efficient algorithm for mining frequent and systematically explore a pattern-growth approach for efficient mining of sequential patterns in large sequence database. Based on this philosophy, we first examine a straightforward pattern growth method, Free Span for Frequent pattern-projected Sequential pattern mining, which help to minimize the efforts of candidate subsequence production. They check another and more well-organized technique, called Prefix Span for Prefix-projected Sequential pattern mining, which offers ordered growth and minimize projected databases. The Prefix Span consumes a much smaller memory space in comparison with GSP. They also check whether one can fix the order of item protuberance in the generation of a probable database.

L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai Presented the Effective Personalized Web Mining by Utilizing The Most Utilized Data [6]. Concept-based user profiling methods aim at capturing users' conceptual requirements. Users' browsed ID and search histories are mapped into a set of contemporary categories. User profiles are created based on the users' preferences on the extracted topical categories. The user profiling method was introduced. For performing search in a database or the web the user information details which is provided by the user. A split data base table is maintained for the user. Whenever the user logs inside with his/her user name and password he would be able to do a personalized search. The search key given by the user would provide the list of links that are related. The search key matching the key words present in the database will get listed.. The link that was clicked and used by the user would get the higher weightings, this gets into the first spot in the list. Each and every time the search made by the user gets updated in the database. In related to this system not only the concerned links but also the uninterested links gets displayed. But this can contain the last few spaces in the list. In this system the utilization concept is provided for a definite set of data set. If the web server is connected and utilize a wide range of data it would be even more efficient. The time duration is calculated in terms of minutes if they are done in seconds it would give an even more fine output.

N.Senthil Kumar, P.M. Durai Raj Vincent considers Web Mining – An Integrated Approach [7]. Web personalization is increasing and more imminent to eradicate the difficulties by taking the content and entire structure of websites to the requirements of the web users and understand the web users access activities and their behavior. To match that, the promising research work has been carried out in Web Mining. Web mining related projects to search the essential and most seek information or patterns from the web hyperlinks structure, page content and web usage log. Web structure functions on the hyperlinks structure and produce the graph structure which provide information about a page ranking. It infers knowledge from the web and links between references in the web. Web usage mining analyses results of user interaction including web logs, click streams and other transactions held. Recent research binds content and structure mining to leverage the technique more strength and to yield high productivity. This system extends the capabilities of traditional web content mining approaches in order to analyze constantly changing web sites containing information about multiple topics (such as online news sites). With the continued

growth of the Web as an information source and as a medium for providing web services, Web Mining continues to play an ever expanding and inevitable role. Web mining has adapted techniques from the field of data mining, database mining and extraction of information, as well as developing some new method of its own like as path analysis.

R.Khanchana and M. Punithavalli is a research scholar, Department of Computer science proposed Web Usage Mining for Predicting Users' Browsing Behaviours by using FPCM Clustering [8]. Objective of web usage mining is to make out the useful data from web data or web log files. The additional purposes are to recover the usability of the web data and to apply the method on the web applications like, perfecting and caching, personalization etc. For assessment management, the outcome of web usage mining can be utilized for target advertisement, enhancing web design, enhancing satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc. Predicting the users' browsing pattern is one of web usage mining technique. For this purpose, it is required to recognize the customers' browsing behaviors by means of analyzing the web data or web log files. Predicting the exact user's next needs is according to the earlier related activities. There are several merits to employ the forecast, like, personalization, structure proper web site, enhancing marketing strategy, promotion, product supply, receiving marketing data, forecasting market trends, and increasing the competitive strength of enterprises etc.

Chhavi Rana studies various research tools on A Study of Web Usage Mining Research Tools [9]. Web usage mining deals with user interacts with the Web. It tries to form sense of the info generated by the online surfer's sessions or actions. There's an attempt to offer an overview of the state of the art within the analysis of web usage mining, whereas discussing the most relevant tools available within the sphere likewise as the niche needs that the present type of tools lack. They offer an outlook on the prevailing tools, their specialized focus with respect to a practical objective and also the want for a additional comprehensive new entrant during this sphere within the light of the present scenario. In the end, they concluded by listing some challenges and future trends during this analysis area. Overall the main focus of the paper is going to be to present a survey of the recent developments during this area that is getting an excessive amount of attention from web development arena. The approach is multi-disciplinary, involving Software Engineering and Artificial Intelligence techniques. There is a strong relation between structured documents (such as Web sites) and a program; the Web is a good candidate to experiment with some of the technologies that have been developed in software engineering. Web Mining has been an important topic in data mining research in recent years from the standpoint of supporting human-centered discovery of knowledge.

4. CONCLUSION

Although Naive Bayesian Classification shows good result in the improvement in time and memory utilization, it can be applied to any web log files but the analysis over various web log data shows that many attributes are not used for classifying as they are irrelevant. With the help of CART the number of irrelevant attributes can be reduced so that the performance can also be proved efficient. The Naïve Bayesian has low time complexity but is not very efficient as per the error rate and the classified instances of the attribute values have been concerned. Also the decision tree created using Naïve Bayesian algorithm provides more number of levels and so is the search for any data from the web data. Additionally, a new strategy that can dynamically determine the grid to assign for a new GMSCP jobs may be helpful. Several powerful grids will be included in this grid environment. We will configure PC clusters and workstations with multi-cores processors to serve as efficient computing grids [10].

5. REFERENCES

[1] Chintandeep Kaur , Rinkle Rani Aggarwal " Web mining tasks and types", International Journal of Research in IT & Management (IJRIM),Volume 2, Issue 2 ,February 2012.

[2] Renta Ivancsy, Istvan Vajk "Frequent Pattern Mining in Web Log Data", Journal of Applied Science at Budapest Tech Hungary,Vol. 3,No-1,pp 77-90,ISSN 1785-8860 2006.

[3] A. K. Santra, S. Jayasudha "Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification", International Journal of Computer Science Issues (IJCSI), Vol. 9, Issue 1, No 2, January 2012.

[4] G.Sathyadevi," application of cart algorithm in Hepatitis disease diagnosis", International Conference on Recent Trends in Information Technology, ICRTIT-IEEE , 978-1-4577-0590,june 2011.

[5] S.Vijayalakshmi,Dr.V.Mohan,"Extracting Sequential Access Pattern from Pre-processed We b Logs", Proceeding in International Conference on IEEE – PACC ,Pp – 2011.

[6] L.K. Joshila Grace,V.Maheswari, Dhinaharan Nagamalai "Effective Personalized Web Mining by Utilizing The Most Utilized Data", International Journal of Database Management Systems (IJDMS), Vol.3, No.3, August 2011.

[7] N.Senthil Kumar, P.M. Durai Raj "Vincent considers Web Mining – An Integrated Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012.

[8] R.Khanchana and M. Punithavalli,"Web Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering", International Journal of Engineering and Technology IACSIT, Vol. 3, No. 5, October 2011.

[9] Chhavi Rana, "A Study of Web Usage Mining Research Tools", Int. J. Advanced Networking and Applications, Volume: 03 Issue: 06, pp1422-1429, 2012.

[10] Chih-Hung Wu, Yen-Liang Wu, Yuan-Ming Chang "Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment",Volume-6,Pp 2909- 2914 IEEE, July 2010

.