Design and Development of Efficient Algorithm in Web Usage Mining For Web Personalization

Mohit M.Patel[#1], Assistant Prof.Shailendra K.Mishra*[2]

[#]*PG Student, Parul Institute of Engineering and Technology ,Vadodara, Gujarat, India*

[*]*Assistant Professor in CSE Department, Parul Institute of Technology, Vadodara, Gujarat, India*

Abstract— **With the promoting of Cyber Crime informationization process, more and more complex data is accumulated, the data mining techniques used in Cyber Crime. The work, that finding in hidden and useful information to guide Cyber Crime Department from large number of web users data, will be helpful to the Cyber Crime reform and development. This article proposes one kind of improvement ID3 algorithm, this algorithm simplified information entropy solution, which is the standard of attribute selection and reduced complication of calculation. This application motivates us to stabiles different proxy server to decrease the cybercrime and increase the secure way for surfing.**

Keywords— **Web Log Mining; Web Usage Mining, Decision tree; ID3 algorithm; performance analysis**

I. INTRODUCTION

Due to past some years the way of working and information collection is rapidly changed and new methods, tools and techniques are developed and implemented for knowledge extraction. The data mining is an application where data is converted into knowledge and in its sub domain web mining data and knowledge is extracted from World Wide Web. This domain is also known as web mining.

The web mining is defined in three rich fields first web usage mining which is used to get knowledge from web log mining. In second branch of this technology web content mining data is formed and extracted from the web pages. And finally the in web structure mining the structure of web is analyzed.

In this paper our main working domain is web usage mining. Web usage mining is the method of mining useful data from web server logs. Web usage mining is the process of finding out what users are looking for on Internet. In this web usage mining based paper we start from data extraction from proxy server Logs and extract the formal information from log file. In next step we work for user differentiation from the same IP source and web Log Classification analysis, Frequent Patterns Analysis and our proposed algorithms are the core work of the system.

II. BACKGROUND

A **server log** is a log file automatically created and Maintained by a server of activity performed by it. Weblog Expert supports log files of the most popular web servers: **Apache** and **IIS**.

Common Log Format

LogFormat "%h %l %u %t \"%r\" %>s %b" common

CustomLog logs/access_log common

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

127.0.0.1 (%h): This is the IP address of the client (remote host) which made the request to the server. If Hostname Lookups is set to On, then the server will try to determine the hostname and log it in place of the IP address.

- (%l) : The "hyphen" in the output indicates that the requested piece of information is not available. In this case, the information that is not available is the RFC 1413 identity of the client determined by identd on the client's machine.

frank (%u) : This is the user id of the person requesting the document as determined by HTTP authentication. The same value is typically provided to CGI scripts in the REMOTE_USER environment variable. If the status code for the request (see below) is 401, then this value should

not be trusted because the user is not yet authenticated. If the document is not password protected, this entry will be "-" just like the previous one.

[10/Oct/2000:13:55:36 -0700] (%t): The time that the server finished processing the request. The format is: [Day/month/year: hour: minute: second zone]
"GET /apache_pb.gif HTTP/1.0" (\"%r\"): The request line from the client is given in double quotes. The request line contains a great deal of useful information. First, the method used by the client is GET. Second, the client requested the resource/apache_pb.gif, and third, the client used the protocol HTTP/1.0. It is also possible to log one or more parts of the request line independently.
For example, the format string "%m %U%q %H" will log the method, path, query-string, and protocol, resulting in exactly the same output as "%r".

200 (%>s): This is the status code that the server sends back to the client. This information is very valuable, because it reveals whether the request resulted in a successful response (codes beginning in 2), a redirection (codes beginning in 3), an error caused by the client (codes beginning in 4), or an error in the server (codes beginning in 5). The full list of possible status codes can be found in the HTTP specification (RFC2616 section 10).

2326 (%b): The last entry indicates the size of the object returned to the client, not including the response headers. If no content was returned to the client, this value will be "-". To log "0" for no content, use %B instead.

III. PROPOSED METHOD

A log file is a recording of everything that goes in and out of a particular server. The only person who has regular access to the log files of a server is the server administrator and a log file is generally password protected, so that the server administrator has a record of everyone and everything that wants to look at the log files for a specific server.
Data pre-processing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data pre-processing transforms the data into a

format that will be more easily and effectively processed for the purpose of the user.

Web usage logs may be pre-processed to extract meaningful sets of data called user transactions, which consist of groups of URL references. Then we filtered useful data.
Classification models are tested by comparing the predicted values to known target values in a set of test data.

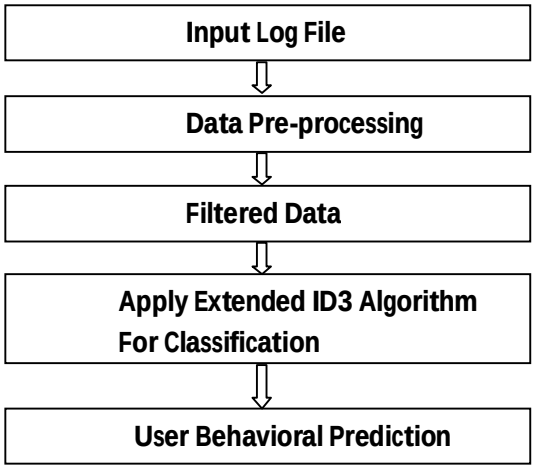


Fig. 1: System Architecture

User sessions may be tracked to identify the user, the Web sites requested and their order and the length of time spent on each one. Once these have been pulled out of the raw data, they yield more useful information that can be put to the user's purposes, such as consumer research, marketing, or personalization.

IV.THE OVERVIEW OF DECISION TREE TECHNOLOGY

Data mining is the process of extract potentially useful, credible information and knowledge from amounts of noisy, fuzzy and random raw-data. Decision tree, an algorithm common used to predict model, can find out some valuable information through huge amounts of data classification.
The decision tree is the basis of learning example inductive learning algorithm, it through the huge amounts of data classification to find some valuable information.

ID3 algorithm is one important method in the technology of decision tree classification and so is widely applied. ID3 algorithm searches through attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n = number of possible values of an attribute) partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices.

The central principle of ID3 algorithm is based on Information theory.

Given a training dataset S, the information entropy of the set S is defined as-

$$I(p,n) = \left(-\frac{p}{p+n}log2\frac{p}{p+n} - \frac{n}{p+n}log2\frac{n}{p+n}\right) \quad (1)$$

Where p is the number of positive example, n is the Number of negative example.

Let us suppose that attribute A ∈{A1,A2……Av}, S is divided into a number of disjoint subsets{ S1,S2…Sv},where Si have pi positive examples and ni negative examples, so the desired information entropy with A as the root is defined as:

$$E(A) = \sum_{i}^{v} \frac{pi + ni}{p + n} I(pi, ni) \quad (2)$$

So the gain of example set S on attribute A is:

$Gain\ (A) = I\ (p, n) - E\ (A)$ (3)

Id3 choose the maximum attribute of Gain (A) as root node.

IV. IMPROVED ID3 ALGORITHM

Every time of choosing a split node, the ID3 algorithm is related to multiple logarithm operation that will obviously affect the decision tree generation efficiency in times of amount of data to operate. Therefore we will consider to change selection criteria from data attribute so that reduce the computational cost of saving decision tree and decision tree generation time. In addition, the choice by ID3 often turn to attributes with more values, because it use each attribute information entropy to judge the value of the data of the division of concentrated properties.

According to the basic principle and algorithm of decision tree based on information theory, we converted the formula of information gain, so as to find a new attribute select criterion. This new standard choosing attributes can not only overcome the ID3 algorithm shortcomings, which easily tend to choose more different values attribute as the test attributes ,but also reduce generation time of the decision tree and calculated cost greatly, so that accelerate construction speed of decision tree, improve the efficiency of decision tree classifier.

According (1) (2), we can get

$$E(A) = \sum_{i}^{v} \frac{1}{(p+n)ln2}\left(-pi\ ln\frac{pi}{pi+ni}, -ni\ ln\frac{ni}{pi+ni}\right) \quad (4)$$

Because (p+n) ln2 is a constant in training set, so, we can assume that the function e(A) satisfies the following equation:

$$e(A) = \sum_{i}^{v} \left(-pi\ ln\frac{pi}{pi+ni}, -ni\ ln\frac{ni}{pi+ni}\right) \quad (5)$$

Using McLaughlin formula:

$$f(x) \cong f(0) + f^{1}(0)x + \frac{f^{11}(0)}{2!}x^2 + \cdots\frac{f^{n}(0)}{n!}x^n \quad (6)$$

When f(x)=ln(1+x) and x is very small . We can get:

ln(+x)≈x

so, we simplify e(A) as:

$$ln\frac{pi}{pi+ni} = \ln\ (1 - \frac{ni}{pi+ni}) \approx \frac{ni}{pi+ni}$$

$$ln\frac{ni}{pi+ni} = \ln\ (1 - \frac{pi}{pi+ni}) \approx \frac{pi}{pi+ni}$$

Put top two equations into e(A). We Can get:

$$e(A) = \sum_{i}^{v} \left(pi\ \frac{ni}{pi+ni} + ni\ \frac{pi}{pi+ni}\right) = \sum_{i}^{v} \left(\frac{2pini}{pi+ni}\right) \quad (7)$$

Assuming that each of the number of attributes is N, so the improved attribute information entropy formula is:

$$e(A) = \sum_{i}^{v} \left(\frac{2pini}{pi+ni}\right) N \quad (8)$$

At last we show the flowchart of information entropy calculation as the following fig.1, which is the key part of whole improved ID3.

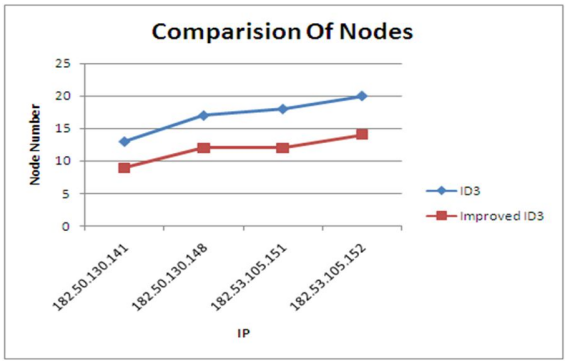
VI.RESULTS

We select 4 datasets to test the traditional ID3 algorithm and improved ID3 algorithm. Comparative analysis of the improved ID3 algorithm and the ID3 differences from the nodes-number, regular-number, accuracy and cost time. Each dataset is conducted 20 experiments, and then calculate the average value, so the experimental data with more generality.

(a) Based on Comparison of Nodes:

IP	ID3	Improve ID3
182.50.130.141	13	9
182.50.130.148	17	12
182.53.105.151	18	12
182.53.105.152	20	14

Table 1: for dataset size for ID3 and Improve ID3



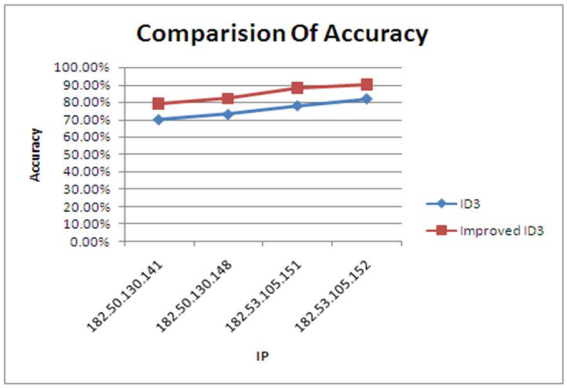
Graph 1: for dataset size for ID3 and Improve ID3

(b) Based On Comparison of Accuracy

IP	ID3	Improve ID3
182.50.130.141	70%	79%
182.50.130.148	73%	82%
182.53.105.151	78%	88%
182.53.105.152	82%	90%

Table 2: Dataset size for ID3 and Improve ID3 for Accuracy

From Table 2 can be seen that the improved ID3 algorithm accuracy is higher than the original ID3 algorithm. And time difference increasing linearly with the increase of data quantity, but the improved ID3 algorithm with the increase tendency of data quantity decline a little bit compared with traditional ID3 algorithm.



Graph 2: Dataset size for ID3 and Improve ID3 for Accuracy.

VII.CONCLUSION

The improved ID3 algorithm can construct more concise decision tree classification model, it's time complexity and cost time in creating decision tree is superior to traditional ID3.The improved ID3 algorithm overcome the traditional ID3 algorithm's shortcomings which easily tend to choose more different values attribute as the test attributes, that can make the structure of decision tree more compact, get a good classification effect and performance.

VIII.REFERENCES

[1] Extraction of Business Rules from Web logs to Improve Web Usage Mining. International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 8, August 2012)

[2] Web Log Data Cleaning For Enhancing Mining Process, International Journal of Communication and Computer Technologies Volume 01 – No.11, Issue: 03 December 2012ISSN NUMBER: 2278-9723

[3] Web Usage Mining: A Survey on Pattern Extraction from Web Logs, International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011

[4] Using Data Fusion and Web Mining to Support Feature Location in Software, M Revelle, B Dit, D Poshyvanyk - (ICPC), 2010 IEEE 18th, 2010 - ieeexplore.ieee.org

[5] Clustering WSDL Documents to Bootstrap the Discovery of Web Services,K Elgazzar, AE Hassan, P Martin - Services (ICWS), 2010 IEEE, 2010 - ieeexplore.ieee.org

[6] A Novel Approach for clustering web user sessions using RST,Ms.Jyoti et al / International Journal on Computer Science and Engineering Vol.2(1), 2009, 56-61

[7] Jiawei Han,Micheline Kamber . Data mining: conceptsandtechniques. MorganKaufmann. 2006. 58-61

[8] Web Usage Mining on Proxy Servers: A Case Study

[9] An Effective System for Mining Web Log