Preparation of Papers in Two-Column Format for Adaptable Web Log Mining from Web Server Logs using Data Preprocessing

Krunal J Joshi
Department of Computer Engineering
Merchant Engineering College

Abstract— With the abundant use of Internet and constant growth of users, the World Wide Web has a huge storage of data and these data serves as an important medium for the getting information of the users access to web sites which are data stored in Web server Logs. Today people are interested in analyzing logs file as they show actual usage of web site. But the data is not accurate so preprocessing of Web log files are essential then after that data are suitable for knowledge discovery or mining tasks. Web Usage or Log Mining, a part of Web mining and application of data mining is used for automatic discovery of patterns in clickstreams and associated data collected or generated as a result of user interactions with one or more Web Sites. We present a comparison study of using enhanced version of decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying interested users. Experimental results conducted shows that the performance metric i.e., time taken and memory to classify the web log files are more efficient when compared to existing C4.5 algorithm.

Key words: Adaptable Web Log Mining, Web Server Logs

I. INTRODUCTION

Web mining is a technique to analyze the online Web contents, navigate between various Web sites and perform transaction of data across the Web. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Figure 1 shows the taxonomy. Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data.

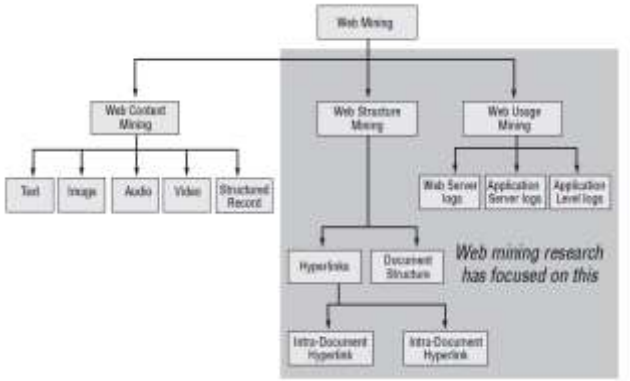

Fig. 1: Objective of web mining

The attention paid to web mining, in research, software industry, and web based organization, has led to the accumulation of significant experience. It is our goal in this chapter to capture them in a systematic manner, and identify directions for future research. Web Mining is based on the knowledge discovery from web. It will extract the knowledge framework and represents it in a proper way. Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. Before the dawn of web mining it was difficult to extract information in proper way from web. But with the advent of web mining it became easy to extract all the features and information about multimedia.

Web mining is the application of data mining techniques to discover patterns from the Web.

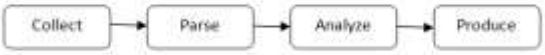

Fig. 1: Steps of web mining

- Parse - extract data from formats
- Analyze - tokenize, rate, classify, cluster
- Produce - useful data

A. Current Trends in Web Mining Applications

WWW can be accepted as a huge digital library. By the same analogy, Web Mining can be viewed as a digital library's librarian. There are several application areas of web mining; the important ones are listed in Figure-3.

The most popular application area of web mining is e-commerce (business-to-customer) and web based customer relationship management. Web usage mining is most dominant application in this context. With the web mining, it is possible to record customer behaviour for web-based business. It is also feasible to adapt web sites based on interesting patterns as a result of analysis on user navigation patterns [Iyer 02]. Web site topology can be customized to provide better facilities for the site user [Mulvenna 00].

Using textual information on the web for predicting customer needs is another research area. [Wuthrich 98, Fung 03] It is possible to associate common phrases with specialized market movements. Knowledge Management Applications can be categorised as "connections" or "collections." Finding common interest communities like Amazon infers a common interest by identifying those who have bought the same book, and recommending other books that the target community prefers. Membership in forums or news groups is a more direct method of identifying a common interest community.

To sum up, the trend in applications of web mining is on the areas that can profit from business-to-consumer e-commerce, or on the areas where Web-based portals are growing rapidly and providing overlapping, and sometimes conflicting information. In all of these areas, the demand comes from the customer. In e-commerce perspective, Internet shoppers play the most important role. Compared to data mining, web Mining applications depend on the end user for the interpretation of the data presented. For example, fraud detection is a typical data mining application, the end-user expect information from software about potential fraud targets. However, web mining applications tend to fall short of this level offunctionalities. The applications might typically identify linked clusters of textual information from which the end user would identify the important patterns. In dynamic systems such as the

Internet, it is a common practice to periodically record samples of activity [7]. Those samples are then used to characterize the activity in the system and to evaluate new mechanisms to be used in this system. This is certainly true of HTTP traffic. On the World Wide Web (WWW), logs of HTTP traffic are recorded continuously as a function of most origin web servers as well as intermediate proxies. The primary function of these logs is to chronicle the operation of these systems. However, as mock-up of HTTP activity, logs generated by these systems are also used for characterization, evaluation and usage reporting. Occasionally, researchers will capture HTTP traffic via other means, such as from augmented client browser. Web Server logs are plain text (ASCII) files, that is independent from the server platform. There are some distinctions between server software, but traditionally there are four types of server logs:

Fig. 2: Application areas of web mining

- Transfer Log
- Agent Log
- Error Log
- Referrer Log

The first two types of log files are standard. The referrer and agent logs may or may not be "turned on" at the server or may be added to the transfer log file to create an "extended" log file format. Each HTTP protocol transaction, weather completed or not, is recorded in the logs and some transactions are recorded in more than one log.

B. Problem definition

Preprocessing stage is important phase in web usage mining to extract and discover user patterns from web server log files.After completion of pre-processing,the cleaned data is stored in databases to be used for fitting in the Generalized Association Rules for rules generations. Conversely, the raw data before analyze is about 1377738 records. Nevertheless, the progression of preprocessing data are prepared discretely due to the system is not currently incorporated. The data preprocessing are prepared separately suitable to the massive amount of data for each log files.

Extraction is a process of removing out uninteresting data or attributes. Ex, The web server logs contains 18 attributes, however removing process has taken out 17 attributes considered uninteresting and only 1 attribute known as "URL" are left in the databases.

Data filtering perform by removing unwanted patterns from each record in the database. Since the pre-processing techniques performed is to mine the interesting patterns, the data end with *.jpg, *.gif, *.bmp be removed. The final data after all process completed is about 38,890 records.

In earlier system it was working only for one log file format and final records has to be stored only into the database (MS SQL-SERVER 2005). And it decreases the performance of Web Server.

II. RELATED WORK

Web server log files store user click streams while navigating a web site. Some of these data are unnecessary for the analysis process and could affect the detection of web attacks. Therefore, preprocessing step comes before applying mining algorithms. Unfortunately, most of the researches in this topic give no details about preprocessing steps. They just mention implicitly that these log files should be converted into suitable format. Figure 11 illustrates steps involved in preprocessing process. It involves integrating data from multiple log files into one single file with one format. The following subsections will contain details about these steps

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. In this section, we provide taxonomy regarding web mining, classification rule mining methodology based on decision trees, the algorithm C4.5 that have been used in the existing work to identify the interested users.

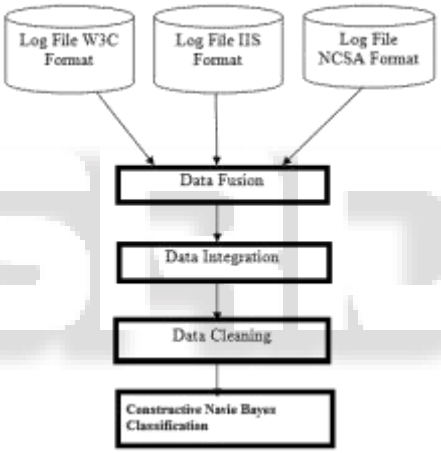

Fig. 4: Proposed preprocessing steps

III. NAIVE BAYESIAN CLASSIFICATION MODEL

The need and requirements of the admin user's of the websites to analyze the user preference become essential, due to massive internet usage. Retrieving the decisive information about the user preferences is achieved, using Naive Bayesian Classification algorithm with quicker time and lesser memory, by means of constructive naïve bayes function. The Naive Bayesian Classification technique as shown in Fig 2, is applied on the web log data to evolve the classification of user page preferences and time spent on the pages of the respective web site (URL).

The web log data (training set) comprises of labels which indicates the class of the observations. New data is classified based on the training set. It preprocesses data in order to remove the irrelevant or redundant attributes and form the normalize data. The decision tree is drawn in a top-down manner. The training sets are at the root. They are partitioned on selected attributes. Partitioning of the training set is processed till there is no more leaf for classifying or there are no samples left and the resultant data is fed to the database.

A. Algorithm

1) Initialization

1) Let T be a training set of samples with k attributes as $A_1, A_2, \ldots A_k$ given by n dimensional vector $Q = \{x1, x_2, \ldots x_n\}$
2) Let P denotes the probability
3) Let G be the Gaussian distribution value Process
4) Given a sample Q, the classifier performs the prediction to determine the attributes having the highest posteriori probability such that
   $P(A_i \mid Q) > P(A_j \mid Q)$ where i,j = 1,2,…k
5) Maximum posteriori hypothesis is calculated using

$$P(A_i \mid Q) = \frac{P(Q \mid A_i)\, P(A_i)}{P(Q)}$$

6) Maximize P(Q | Ai) P(Ai) if both P(Q | Ai) P(Ai) are known or P(Q | Ai) if only P(Q | Ai) is known.
7) If the web log data set contain many attributes it results in maximum of computation time which can be reduced using the following equation
   $P(Q \mid A_i) \equiv \lambda\, P(x_n \mid A_i)$.
8) Calculation of Gaussian distribution with mean μ and standard deviation σ is calculated by

$$G(x, \mu, \sigma) = 1 \qquad \frac{\exp\ (x - \mu)^3}{\sqrt{3} \prod \qquad (3\sigma)^3}$$

9) The above equation can be simplified as
   $P(xn \mid A_i) = G(xn, \mu A_i, \sigma A_i)$

Where μA_i refers to the mean and σA_i refers to the standard deviation value of attribute S_k.

Maximum Likelihood Estimation (MLE) is used for estimating the parameters for a given training data set. If a clear result cannot be achieved due to time or cost constraint, by using the mean and standard deviation the maximum likelihood estimation can be accomplished. The MLE is discussed in the section given below.

B. Maximum likelihood Evaluation (MLE)

Let S be the training set with attributes (s1,s2,…,sn) with a vector as Q. To evaluate the maximum likelihood we have to form the density function that is given as :

$$F(s_1, s_2, \ldots s_n \mid Q) = f(s_1 \mid Q) * f(s_2 \mid Q) * \ldots .f(s_n \mid Q)$$
$$= \Omega\, f(S_i \mid Q) \text{ where } i = 1,2,\ldots n$$

Where s1,s2,…sn specifies the parameters and Q being the vector is a random variable. Maximum likelihood for S can be evaluated as

$$\text{Max}(\mu A_i, \sigma A_i) \mid \Omega\, f(S_i \mid Q)$$

In our work Maximum likelihood Evaluation is calculated for a given training data set – web log data which produces a distribution function with the observed data having the greatest probability.

C. Decision Tree model

Decision tree model is a method most frequently used in data mining. The purpose is to create a model that predicts the resultant of a target variable based on several input variables given by the user as training data set. An example is shown in fig 3. The interior node corresponds to one of the input variables. The input variable consists of the children as edges. Each leaf shows a value of the target variable given with the values of the input variables which are shown by the path from the root to the leaf. The process is repeated in a recursive manner till a resultant value is derived. The parameters used in web log data to classify the user as interested or not interested is given in table 1.

Parameters Used	Explanation
P_1	No of pages view >10
P_2	Time taken >5
P_3	Hyperlink >5
P_4	Personal Information given by user = "yes"
Table	1 Parameter consideration

Using the parameters given in the table 1 a decision tree is formed as in fig 3 using the naïve Bayesian classifier algorithm which helps to determine whether a user who logs into the system is an "interested user" or "not interested user". By means of naïve Bayesian algorithm the memory utilized and time taken can be reduced and maximum likelihood of the parameter is also increased.

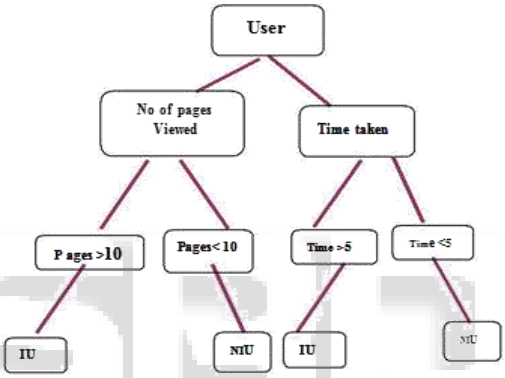

Fig. 5: A Decision Tree generation for Interested User and Not Interested User

IV. IMPLEMENTATION EXPERIMENTAL RESULTS

Web log data was tested on log files stored by the server. We took into account some portions of log files during the same time interval of five hours of five different days;

Fig. 6: Formatted NCSA log file

The data cleaning model evaluate the web log file to determine the log file that are redundant or irrelevant. As described before, the results of data cleaning is derived in 3.1 which removes all group of irrelevant requests.

The C4.5 algorithm is applied on this datasets. The experimental result shows that the time taken by this algorithm is 13.02 Secs and the memory utilization of this algorithm for the same data set is 5.33 KB.

The Naïve Bayesian classification algorithm is applied with the same data sets and experimental results shows that the time taken by this algorithm is 7.68 Secs and memory utilization of this algorithm is 4.05 KB. Which is comparatively efficient than C4.5 algorithm

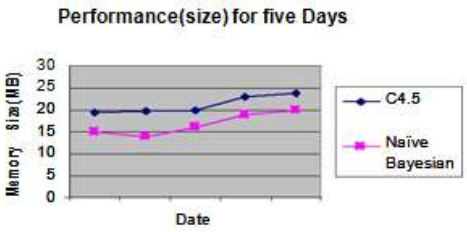

Fig 7: Comparative analysis of average performance (size) for C4.5 and Naïve Bayesian Classification

The figure shows the performance of memory size utilized by the two models. From the graph it is clear that our proposed model Naïve Bayesian Classification outperforms the existing model C4.5 algorithm in terms of minimal memory consumption.

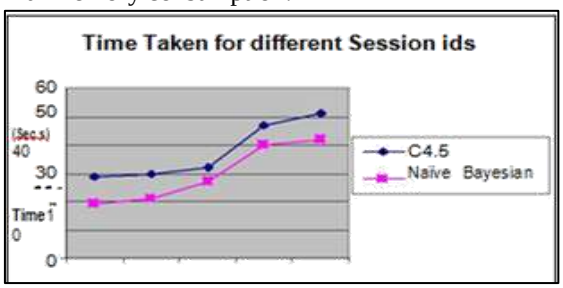

Fig. 8: Comparative study for different session ids

Figure 10 shows on the x axis the session id and time taken measured on seconds on y axis for web log data file observed on five different days for different session ids. It shows that the average time taken to compute the maximum likelihood of user preference evaluation in Naïve Bayesian Classification for different session ids is better, when compared with the existing model of enhanced C4.5 decision tree algorithm.

V. CONCLUSION

This Research work includes preprocessing phase of web usage mining which can be utilized in industry and application oriented system. Here customized web log preprocessing rather than traditional approach which reduces size of raw web log file. Improvement in execution time and memory complexity.

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 recommendation, and Web personalization. So this tool is decrease transaction time and increase performance of web server.

REFERENCES

[1] Srivastava J, Desikan P and V Kumar , "Web Mining-Concepts,Applications & Research Direction" in 2002 Conference
[2] Srivastava J, Desikan P and V Kumar , "Web Mining-Accomplishment & Future Directuins" in 2004 Conference
[3] Rekha Jain and Dr G. N Purohit,"Page Ranking Algorithms for Web Mining" International Journal of Computer Applications (0975 – 8887 Volume 13– No.5, January 2011
[4] Srivastava, J., Cooley, R., Deshpande, M., And Tan, P-N. (2000)." Web usage mining: Discovery and applications of usage patterns from web data", SIGKDD Explorations, 1(2), 12-23.H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch.4.
[5] Maier T. (2004). A Formal Model of the ETL Process for OLAP-Based Web Usage Analysis. In Proc. of "WebKDD- 2004 workshop on "Web Mining and Web Usage Analysis", part of the ACM KDD: Knowledge Discovery and Data Mining
[6] Meo R., Lanzi P., Matera M., Esposito R. (2004). Integrating Web Conceptual Modeling and Web Usage Mining. In Proc. of "Web KDD- 2004 workshop on Web Mining and Web Usage Analysis", part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA.
[7] Desikan P. and Srivastava J. (2004), Mining Temporally Evolving Graphs. In Proceedings of "Web KDD- 2004 workshop on Web Mining and Web Usage Analysis", B. Mobasher, B. Liu, B. Masand, O. Nasraoui, Eds. part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA.
[8] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J. (2001). "Measuring the accuracy of sessionizers for web usage analysis", In Workshop on Web Mining, at the First SIAM International Conference on Data Mining, 7-14.
[9] Srivastava, J., Cooley, R., Deshpande, M., And Tan, P-N. (2000). "Web usage mining: Discovery and applications of usage patterns from web data", SIGKDD Explorations, 1(2), 12-23.
[10] J. Hou and Y. Zhang, "Effectively Finding Relevant Web Pages from Linkage Information", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.
[11] R. Kosala, and H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
[12] Raju G.T. and Sathyanarayana P. "Knowledge discovery from Web Usage Data : Complete Preprocessing Methodology, ", IJCSNS 2008

