© 2015 IJSRSET | Volume 1 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN: 2394-4099
Themed Section: Engineering and Technology

A Survey on User Behaviour Prediction Using Web Server Log

Komal Anadkat[1], Prof.Bakul Panchal[2]
[1]P.G. Student, Information Technology, L. D. Engineering College, Ahmedabad, Gujarat, India
[2]Assistant Professor, Information Technology Department, L. D. Engineering College, Ahmedabad, Gujarat, India

ABSTRACT

Web prediction is a classification problem in which we attempt to predict the next set of request that a user may make based on the knowledge of the previously visited pages.User behaviour prediction deals with collecting the web server logs, prepossess the data and analyze the pattern. The main goal of this process is to extract the useful pattern from raw data collection. Prepossessing contains cleaning the data, user identification and session identification which saves 80% of processing data. Web server log contains much useful information like server ip, time, date, error code, browser etc. there are various prediction models like n-gram session identification, association rule mining, markov model and support vector machine are available. Here comparison is done on the basis of accuracy, processing time and scalability.

Keywords: Prepossessing, Web Server Log, Association Rule Mining, Markov Model

I. INTRODUCTION

As one of the three categories of web mining, web log mining [1] is aimed at how to extract useful knowledge model such as association rules, sequential patterns and clustering analysis from web data, whose results can be applied to optimize the website structure, webpage prefetching, adaptive website and many other aspects [2]. There are mainly three steps in web log mining: data preprocessing, pattern recognition and pattern analysis, among which data.

Preprocessing is the critical and primary task [3]. The data preprocessing includes data cleaning, user identification, session identification and path completion, etc., whose results will directly affect the efficiency and accuracy of web log mining [2].

In order to attract new customers and retain current customers, web sites' administrators want to know their customers' profits. But web servers record and accumulate data about user interactions whenever requests for resources are received. The mass amounts of log data make it impossible to find useful information directly. What can help administrators get useful information from these log files? It is web log file analysis. Web log file analysis began with the purpose to offer a way to Web site administrators to ensure adequate bandwidth and server capacity to their organization. [1] Web prediction is a technique in which we attempt to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. When a prediction model for a certain Web site is available, the search engine can utilize it to cache the next set of pages that the users might visit.

II. LOG FILE PREPOSSESING

The Procedure of data preprocessing : 1) Data cleaning 2) user identification 3) Session identification

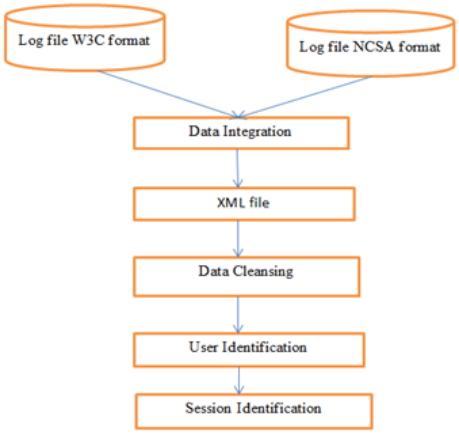


Figure 1: Log file Pre-processing

1) Data Cleaning [4]

Data cleaning means deleting needless data, which can't reflect the characters of user's accessing behaviour. It covers with some aspects as follows 1) Irrespective attribute: The attributes paid attention to include users IP address, URL pages requested, and accessing time and so. And other attributes should be got rid of.2) Content of pages as picture, video and audio resources and the logical units as script CSS files.3) the actions that not request for pages and fail requested pages. The actions that not request for pages failed requested pages. The former can be judged by the code of POST or GET and the later can be judged by the state code.

2) User Identification [4]

Traditional user identification is carried out according these rules: 1) Different IP address refer to different users.2) The same IP with different operating system or different browser should be consider as different user.3) While the IP, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before according to the topology of the site.

3) Session Identification [4]

Web log mining covers a long time periods, therefore users may access the site more than once. Session identification is in order to divide the access records into several accessing sequences, in which the pages are requested at the same time. Traditional session identification algorithm is based on a uniform and fixed timeout. While the interval between two sequential requests exceeds the timeout, new session is determined.

III. LITERATURE SURVEY

There are various models available for prediction of user behaviour from web server logs. each model has their own merits and demerits. To prepare an efficient prediction model which can handle the large amount of data effective the proper analysis of each one should be done.

1. N-Gram for Session Representation [5-6]

In user behaviour prediction the well known representation of session is n-gram. N-gram depicts sequences of page clicks by a population of users surfing aWeb site. Each component of the N-gram takes specific page ID value that identifies a Web page. For example, the N-gram <A10, A16, A4, A2> describes the fact that the user has visited pages in the following order: page 10, page 16, page 4 and finally page 2. Many models further process these N-gram sessions by applying a sliding window to make training examples have the same length. As an example, consider a log file L consisting of the following request paths:

A,B,C,D
A,B,C,F
A,B,C,F
B,C,D,G
B,C,D,G
B,C,D,F

If we were to construct a 3-gram model, we have two 3- grams to build our prediction model on. These are

A,B,C;
B,C,D

Our application of the algorithm returns the following hash table H3():

N-Gram Prediction

A,B,C	F
B,C,D	G

However, if we were to build a 2-gram model, then we have the following 2-grams to contend with:

A,B;
B,C;
C,D

Based on the log data, we can build the following 2-gram prediction model H2():

N-Gram Prediction

A,B	C
B,C	D
C,D	G

Zhong Su*, Qiang Yang* and Hongjiang Zhang [6] proposed n-gram(+) ,which is more accurate than previous n-gram.

As an example, assume that we have built up 3-gram and 2-gram models as H3 and H2 . Suppose that we observe that the current clicking sequence consists of

only one click "DBC". In this case, the prediction algorithm checks H3 first to see if an index "DBC" exists. It finds out that the index does not exist. Therefore, it checks the 2-gram model H2 for the index "BC", which exists, thus the predicted next click is "D", according to H2.

2. Association Rule Mining [5]

ARM is a data mining technique that has been applied successfully to discover related transactions. In ARM, relationships among item sets are discovered based on their co occurrence in the transactions. Specifically, ARM focuses on associations among frequent item sets. For example, in a supermarket store, ARM helps uncover items purchased together which can be utilized for shelving and ordering processes. In the following, we briefly present how we apply ARM in WPP. In WPP, prediction is conducted according to the association rules that satisfy certain support and confidence as follows.

For each rule, $R = X \rightarrow Y$, of the implication, X is the user session and Y denotes the target destination page. Prediction is resolved as follows:

$$\text{Prediction}(X \rightarrow Y) = \arg\max \sup \frac{(X \cup Y)\ ,\ X \cap Y}{\text{Supp}(X)}$$

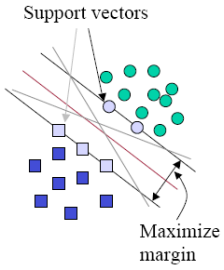
Here the cardinality of Y can be greater than one, i.e., prediction can resolve to more than one page. Moreover, setting the minimum support plays an important role in deciding a prediction. In order to mitigate the problem of no support for X ∪ Y, we can compute prediction (X → Y), where X is the item set of the original session after trimming the first page in the session.

3. Markov Model [5]:-

The basic concept of Markov model is to predict the next action depending on the result of previous actions. In Web prediction, the next action corresponds to predicting the next page to be visited. The previous actions correspond to the previous pages that have already been visited. In Web prediction, the *K*th-order Markov model is the probability that a user will visit the *k*th page provided that she has visited the ordered $k - 1$ pages . For example, in the second-order Markov model, prediction of the nextWeb page is computed based only on the two Web pages previously visited. It can be easily shown that building the *k*th order of Markov model is linear with the size of the training set. The key idea is to use an efficient data structure such as hash tables to build and keep track of each pattern along its probability. Note that a specific order of Markov model cannot predict for a session that was not observed in the training set since such session will have zero probability.

4. Support Vector Machine

Kiran M, Amresh Kumar, Saikat Mukkherjee, and Ravi Prakash has proposed SVM which performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. In the reference of SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyperplane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the support vectors. An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors is maximized.



SVMs maximize the margin around the separating hyperplane.

- The decision function is fully specified by a (usually very small) subset of training samples, the support vectors.
- This becomes a Quadratic programming problem that is easy to solve by standard methods.

Comparison

Sr No	Prediction Model	Advantages	Disadvantages
1	N-Gram	When n>3, a precision gain on the order of 10% or more	As increase in sequence length, there is an increase in precision and decrease in applicability.
2	ASSOCIATION RULE MINING	ARM do not generate several models for each separate *N*-gram like Kth markov model.	ARM endures efficiency and scalability problems by generating item sets and it require exponential time with the number of item sets.
3	MARKOV MODEL	Efficiency and performance, prediction time.	A specific order of Markov model can't predict for a session that was not observed in the training set since such session will have zero probability.
4	SUPPORT VECTOR MACHINE	Accuracy in predicting seen and unseen data compare to Markov model. Robust classification and proven effectiveness	It's suffered from scalability problem in both memory requirement and computation time when the input dataset is too large.

IV. CONCLUSION

In this paper, we have reviewed many prediction model like n-gram session identification, markov model, association rule mining and support vector machine. Each model has its own advantages and disadvantages. For the future work, we will try to modify any of the above model and apply it on hadoop framework for parallel processing to reduce the training time and will compare the results with commercial software such as Google analytics etc.

V. REFERENCES

[1] Bing, L., Web Data Mining, 2nd ed., Berlin: Springer-Verlag, 2011.

[2] Hao, C., Yu-bo, J., Cheng-wei, H., Zhi-qiang, H., "Improved method for session identification in web log mining," Journal of Computer Engineering and Design, vol. 30, 2009, pp. 1321-1323.

[3] Zi-jun, C., Xin-yu, W., Wei, L., "Method of web log session reconstruction," Journal of Computer Engineering, vol.33, 2007, pp.95-97.

[4] He Xinhua,Wang Qiong," Dynamic Timeout-Based a Session Identification Algorithm" Electric Information and Control. Engineering (ICEICE), 2011 International Conference on ISBN 978-1-4244-8036-4,15-17 April 2011

[5] Mamoun A. Awad and Issa Khalil, "Prediction of User's Web- Browsing Behavior: Application of Markov Model", IEEE Transactions on Systems, Man, And Cybernetics—Part B: Cybernetics, vol. 42, no. 4, pp. 1131-1142, August 2012.

[6] Zhong Su*, Qiang Yang* and Hongjiang Zhang," WhatNext: A Prediction System forWeb Requests using N-gram Sequence Models", IEEE