Predictive Analysis of Users Behaviour in Web Browsing and Pattern Discovery Networks

P.Saravana kumar/ R.Iswarya
HOD/Assistant Professor/ Project Engineer
PS Muthu college of Arts and Science,Theni,TamilNadu,India/
CDAC, Chennai, TamilNadu, India

R.Vidhya
Assistant Professor
SRM University, Chennai, TamilNadu, India

Abstract- Web-Page prediction is a classification problem 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. To achieve this we depend on the web access log files which are recorded in the server. These web access log files can be mined to extract interesting pattern so that the user behavior can be understood. Predicting user's behavior while serving the Internet can be applied effectively in various critical applications. Such application has traditional tradeoffs between modeling complexity and prediction accuracy. This paper presents an overview of web page prediction and also provides a survey of the Markov model used for predicting the next web page that is to be visited by the user.

Keywords – Web Usage Mining, Survey of the Markov model, Web page prediction

I. INTRODUCTION

The paper deals with an overview of web page prediction and also provides a survey of the Markov model used for predicting the next web pages that is to be visited by the user..

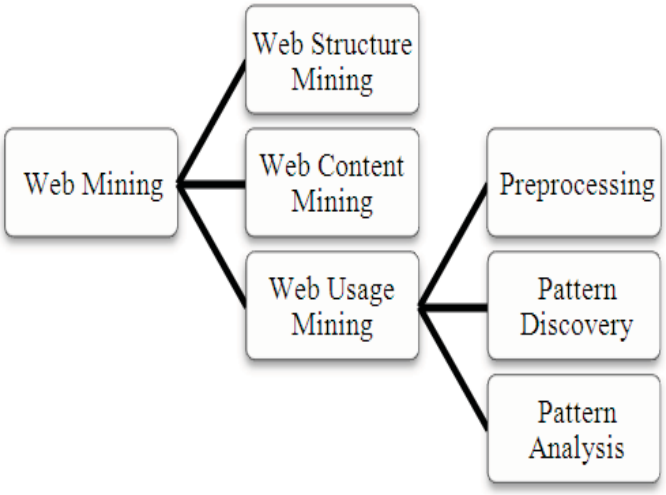
II. WHAT IS WEB MINING?

Web mining can be broadly defined as discovery and analysis useful information from the WWW**.** Based on the different emphasis and different ways to obtain information.

A. Web usage mining –

Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site some users might be looking at only textual data, whereas some others might be interested in multimedia data

Figure 1. Web Mining Classification



B. Web Structure Mining –

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site,according to the type of web structural data.

C. Web Content mining –

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content.

III. SCOPE OF PREDICTIVE ANALYSIS AND WEB MINING

Scope includes implementing the type – Web usage mining is the application of data mining techniques to discover usage pattern from Web data, in order to understand and better serve the needs of Web-based applications.

The application is web based; Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web and extract files from the web server log files its load and the performance can be monitored using the open source technology Data mining tool Rapid Miner.

IV.CLASSIFICATION ON WEB USAGE MINING IN PREDICTIVE

Basically the web usage mining classifies the process into three categories.The first is preprocessing state in which user sessions are inferred from log data. The second searches for patterns in the data by making use of standard data mining techniques, such as association rules or mining for sequential patterns. In the third stage an information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns. Links between pages and the similarity between contents of pages provide evidence that pages are related.

1. DATA PREPROCESSING PHASE

The data pre-processing phase was performed using a reduced log file, which was “cleaned” by removing all useless, irregular, and missing data from the original common log file. After the initial pre-processing, a session filter was applied to the reduced log file for feature extractions. The purpose of the filter was to aggregate all user requests within a session into a single set of variables.

i) Data Collecting
ii) Data Cleaning
    a) Missing Data
    b) Noisy Data
iii) Data Integration

a) Aggregation
b) Normalization

iv) Data Reduction
a) Data Cube
b) Numerosity
c) Data Discretization

2. PATTERN DISCOVERY PHASE

Statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns. The data mining phase included two sub-phases: (a) descriptive analysis, and (b) artificial intelligence analysis. Descriptive analysis was used with summarizing, clustering, and association rules techniques to generate an overview on the dataset, to gain an insight into Users' characteristics, and to depict Users' browsing patterns. Artificial intelligence analysis was used for predictive purposes.

i) Statistical & Path Analysis
ii) Association Rules
iii) Sequential Patterns
iv) Clustering & Classification Rules
v) Boosting
vi) Bagging

3. PATTERN ANALYSIS PHASE

The pattern analysis phase included data interpretation and evaluation of the results. This phase was needed to identify meaningful results from outcomes of the data mining phase

i) Pattern Filtering
ii) Aggregation
iii) OLAP Analysis
iv) Rules, Patterns,& Statistics
v) Visualization Techniques
vi) Evaluation & Report Generation (Meta Data View, Data View, Plot View, 3D…Etc)

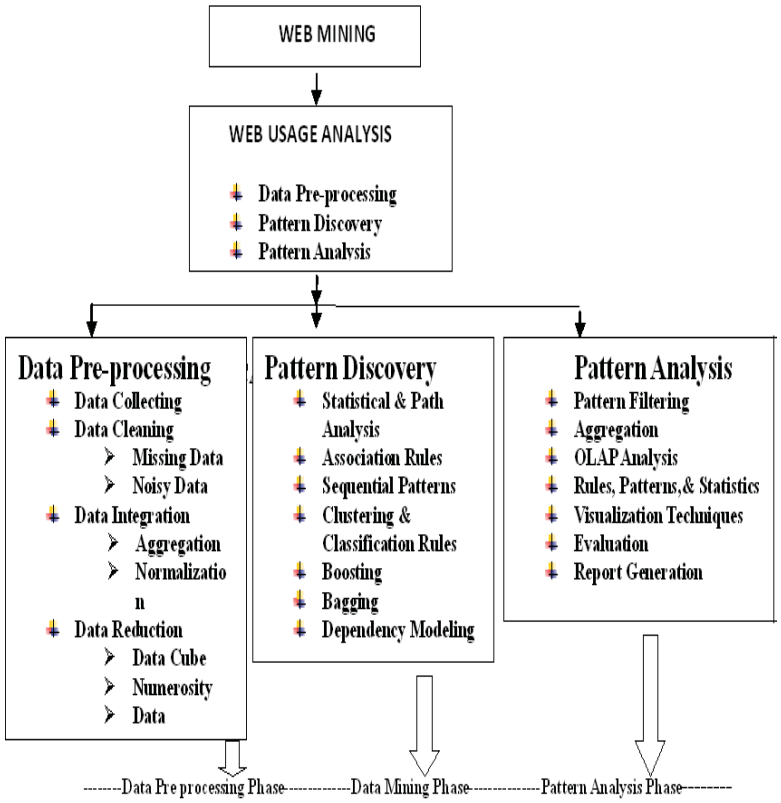


Figure 2. Web Usage Mining Phase

V.WUM WORK FLOW

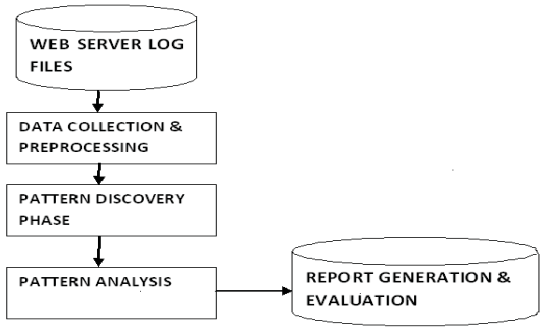


Figure 3. WUM Work Flow

WUM Work flow specifies the step by step execution of the rapid Miner tool. Once the tool got extracted from web server log files , The first is preprocessing state in which user sessions are inferred from log data. The second searches for patterns in the data by making use of standard data mining techniques, such as association rules or mining for sequential patterns. In the third stage an information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns. Links between pages and the similarity between contents of pages provide evidence that pages are related

VI. Roles and Responsibilities of Mining Techniques

TABLE I. Event Prediction

PREDICTION OF NEXT EVENT	➢ MARKOV CHAINS ➢ SEQUENCE MINING
DISCOVERY OF ASSOCIATED EVENTS OR APPLICATION OBJECTS	➢ SEQUENCE MINING ➢ ASSOCIATION RULES
DISCOVERY OF VISITOR GROUPS WITH COMMOMN PROPERTIES AND INTERESTS	➢ CLUSTERING
DISCOVERY OF VISITOR GROUPS WITH COMMON BEHAVIOUR	➢ CLUSTERING ➢ SESSION CLUSTERING
CHARACTERIZATION OF VISITORS WITH RESPECT TO A SET OF PREDEFINED CLASSES	➢ CLASSIFICATION
CARD FRAUD DETECTION	➢ CLASSIFICATION
CLASSIFICATION WITH K-NN BASED ON AN EXPLICIT SIMILARITY MEASURE	➢ K- NEAREST NEIGHBOR ALGORITHM
RETURNS CLASSIFICATION MODEL USING ESTIMATED NORMAL DISTRIBUTIONS	➢ NAIVE BAYES CLASSIFICATION

a. Bagging

The concept of bagging (voting for classification, averaging for regression-type problems with continuous dependent variables of interest) applies to the area of predictive data mining, to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data. It is also used to address the inherent instability of results when applying complex models to relatively small data sets.

b.Boosting

The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification

c.Key features of Rapid Miner Tool

1) Freely available open-source data mining and analysis system
2) Runs on every major platform and operating system
3) Most intuitive process design
4) Multi-layered data view concept ensures efficient data handling
5) GUI mode, server mode (command line), or access via Java API
6) Simple extension mechanism
7) Powerful high-dimensional plotting facilities
8) Most comprehensive solution available: more than 500 operators for data integration and transformation, data mining, evaluation, and visualization
9) Automatic meta optimization schemes
10) Definition of re-usable building blocks
11) Standardized XML interchange format for processes
12) Graphical process design for standard tasks, scripting language for arbitrary operations
13) Machine learning library WEKA fully integrated

14) Access to data sources like Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, Text files and more

15) Most comprehensive data mining solution with respect to data integration, transformation, and modelling methods

16) High prediction accuracy.

17) Handling a large number of databases.

18) Enhanced decisioning in a less time.

VII.WEB USAGE MINING ARCHITECTURE

i) While extracting simple information from web logs is easy, mining complex structural information is very challenging.

ii) Data cleaning and preparation constitute a very significant effort before mining can even be applied. The relevant data challenges include: elimination of irrelevant information such as image files and cgi scripts, user identification, user session formation, and incorporating temporal windows in the user modelling. After all this pre-processing, one is ready to mine the resulting database. We have developed a general architecture for Web usage mining.

iii) The WEBMINER is a system that implements parts of this general architecture. The architecture divides the Web usage mining process into two main parts.

iv) The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes pre-processing, transaction identification, and data integration components.

v) The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine.

vi) Parallel to these layers lays the testing database. It will have the details of all the transactions that are taking place during the test run.

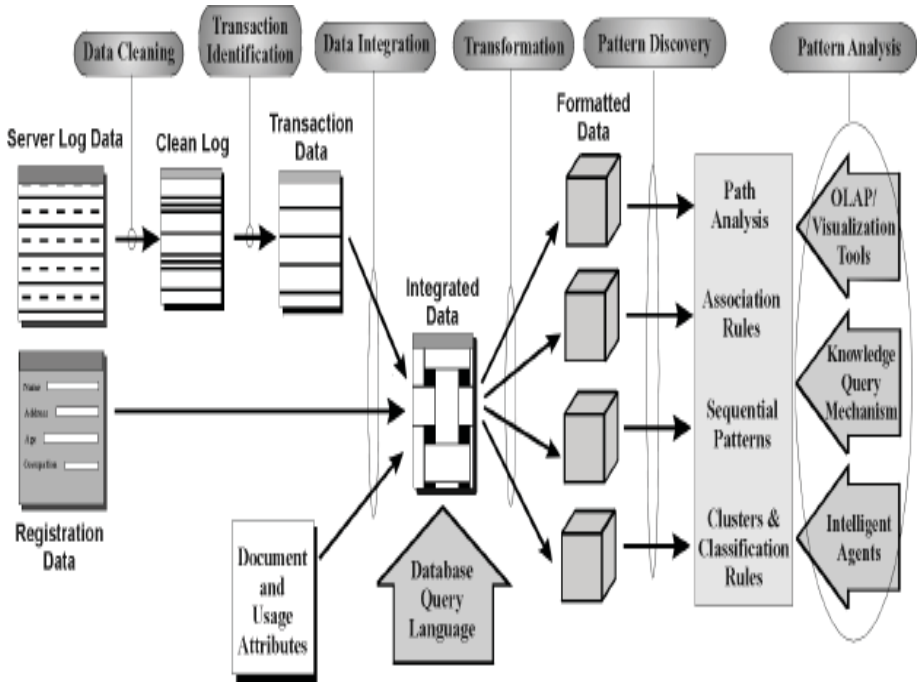


Figure 4. Overall system Architecture

VIII.CONCLUSION

In the near future, plan to extend this paper by conducting in-depth analysis and study of our proposed pattern discovery. Additionally, plan to explore other features in the session's logs by extracting statistical features from data sets to improve prediction accuracy.

REFERENCES

[1] Prediction of User's Web-Browsing Behavior: Application of Markov Model-Mamoun A. Awad and Issa Khalil
[2] M. Awad and L. Khan, "Web navigation prediction using multiple evidence combination and domain knowledge," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans
[3] Cybern. A, Syst., Humans, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.
[4] Internet Traffic Archive.: http://ita.ee.lbl.gov/html/traces.html
[5] http://rapidminer.com/learning/getting-started/
[6] http://1xltkxylmzx3z8gd647akcdvov.wpengine.cdn.com/wpcontent/uploads/2013/10/DataMiningForTheMasses.pdf
[7] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," In *Proc. 20th Int. Conf. VLDB*, Santiago, Chile, 1994.
[8] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, China Machine Press, Beijing. 2007.
[9] S. Jespersen, T. B. Pedersen and J. Thorhauge, "Evaluating the Markov Assumption for Web Usage Mining," *WIDM'03*, pp.82-89, 2003.
[10] D. Dhyani, S. S. Bhowmick and W. K. Ng, "Modeling and Predicting Web Page Accesses Using Markov Processes," *DEXA'03 IEEE*, 2003.
[11] Cooley, R., Mobasher, B. and Srivastava, J. "Web mining: information and pattern discovery on the World Wide Web," in International Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, 1997.
[12] Jalali, M., et al. "A new clustering approach based on graph partitioning for navigation patterns mining," in International Conference on Pattern Recognition, 2008, 1-4.
[13] Berry, M., J., A., & Linoff, G., S., (2000). Mastering data mining. New York: Wiley.
[14] Edelstein, H., A. (1999). Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp.
[15] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery & data mining. Cambridge, MA: MIT Press.
[16] Han, J., Kamber, M. (2000). Data mining: Concepts and Techniques. New York: Morgan-Kaufman.
[17] Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.
[18] Pregibon, D. (1997). Data Mining. Statistical Computing and Graphics, 7, 8.
[19] Weiss, S. M., & Indurkhya, N. (1997). Predictive data mining: A practical guide. New York: Morgan-Kaufman.
[20] Westphal, C., Blaxton, T. (1998). Data mining solutions. New York: Wiley.
[21] Witten, I. H., & Frank, E. Data mining. New York: Morgan-Kaufmann.