International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 01, January 2016
www.ijates.com
ijates
ISSN 2348 - 7550

A SURVEY ON WEB LOG MINING AND PATTERN PREDICTION

Nisha Soni[1], Pushpendra Kumar Verma[2]

[1]M.Tech.Scholar, [2]Assistant Professor, Dept.of Computer Science & Engg. CSIT, Durg, (India)

ABSTRACT

Web sites have abundant web usage log which provides great source of knowledge that can be used for discovery and analysis of user accessibility pattern. The web log mining is the process of identifying browsing patterns by analyzing the user's navigational behaviour. The web log files which store the information about the visitors of web sites is used as input for web log mining and pattern prediction process. First these log files are pre-processed and converted into required formats so web usage mining techniques can apply on these web logs for frequent patterns. The obtained results can be used in different applications like modification of web sites, system improvement, business intelligence, and personalization etc.

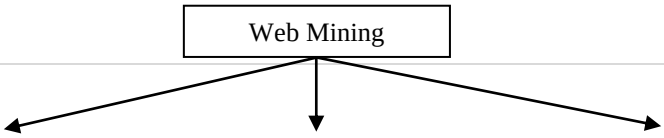
Keywords: Browsing Patterns, Frequent Patterns, Web Log Mining, Web Usage Log.

I. INTRODUCTION

A web is vast, temporary, dynamic, multiform and mostly amorphous data repository, which stores incredible amount of information/data that enhance the complexity of how to retrieve the relevant information from the different user's point of view. Today Web is widely used in every aspects of day to day life. As daily use of web/Internet is increasing mining of web/database is more demanding for the retrieval of more accurate/relevant information. To guess the user's behaviours and personalize information to shrink the traffic load and create the Web site suited for the different set of users, the Web service providers desire to find the technique [1].

Web mining is the process of discovery and analysis of relevant/useful information from the web by implementing different type of data mining techniques. Web mining can be categorized in three ways: web content mining, web usage mining and web structure mining. Web content and structure mining is mainly deal with content and structure of the web, while web usage mining uses the secondary data which is derived from the interactions of the users with the web [2].

Web Usage mining can be defined as discovery and analysis of frequent pattern or web usage pattern from web logs. Web usage mining is a web mining technique hence, also known as web log mining. Personalization, modification of websites, system improvement, business intelligence etc are the various applications of web usage mining.



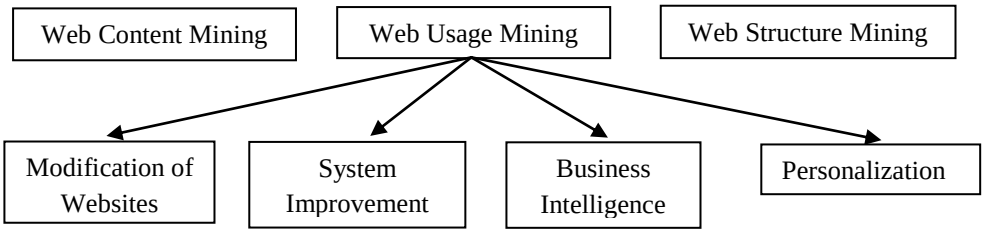


Fig. 1: Web Mining Categories

II. LITERATURE REVIEW

Web log mining technique is apply to improve web services and mining web navigation patterns efficiently. Sagar More [3] proposed a method for mining path traversal patterns. In that paper they first clear the concept of throughout surfing pattern (TSP) which predict the path of website visitor. In next part they apply modified graph traverse algorithm to make mining from TSP in efficient manner. The modified algorithm use the filtering technique which removes duplicate and unwanted data in web browsing session and show the effectiveness as compared to formal algorithms.

Anshul Bhargav and Munish Bhargav [4] proposed a method for pattern discovery and analysis of web usage patterns from web logs that helps the website Administrators to better serve the needs of their websites users. In that paper they proposed a framework which is based on three steps: pre-processing, pattern discovery and users classification. Work mainly focused on doing users classification on three bases: country based, site entry based and access time based classification. This helps in the efficient administration and personalization of the websites and for the increase in profit from the particular websites.

Jia-Ching, Chu-Yu and Vincent S. [5] proposed an efficient incremental data mining algorithm named id-Matrix-Miner for mining web navigation patterns with dynamic threshold. They also proposed a new data structure, id-Matrix, for storing the useful information so as to avoid re-mining the original database. Hence, the web navigation patterns can be discovered efficiently when the navigation sequence database is updated.

III. WEB USAGE MINING

3.1 Overview

Web log mining or web usage mining process involves five steps: [2]

I. Data Collection
II. Data Preparation
III. Pattern Discovery
IV. Analysis of Discovered Patterns
V. Use of Discovered Patterns for Different Applications

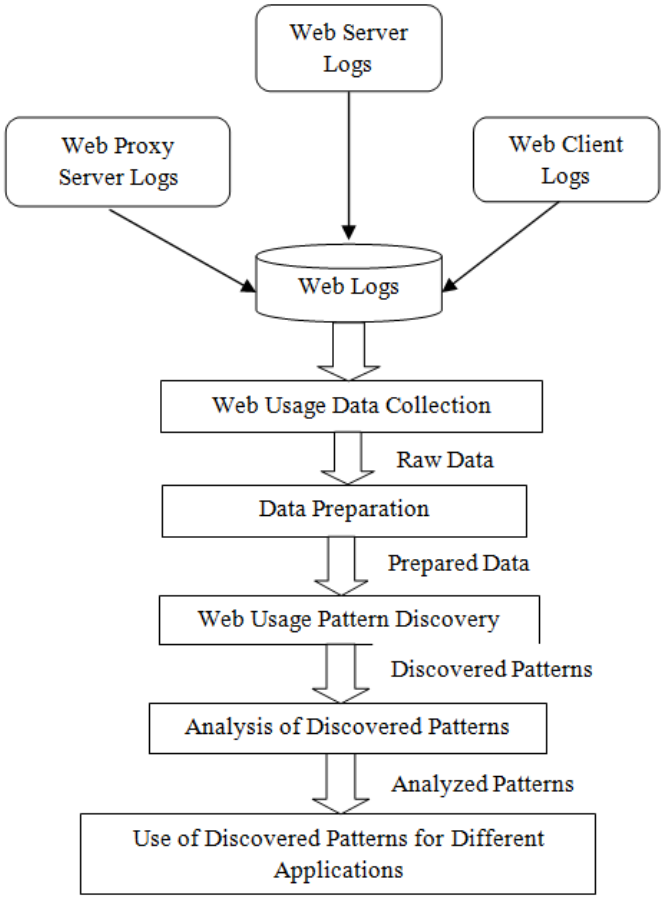


Fig. 2: Web Log Mining Process

3.1.1 Web Usage Data Collection

Usage data collection in form of web logs, which contains record of user activities on Web sites. These records are basically collected from:

i) Web Servers

ii) Web Proxy Servers

iii) Web Client Servers

3.1.2 Data Preparation

The data collected from the web logs are raw data that may contain duplicate, unwanted, partial and conflicting data so the main motive of data pre-processing is to transfer raw log files in particular format which data mining techniques can handle easily. Accuracy of pattern prediction is directly proportional to quality of data hence; the data preparation is an important phase of web log mining process. The main task involved in pre-processing are: data cleaning, user identification, session identification, path completion, data integration and formatting.

3.1.3 Pattern Discovery

In web usage mining process pattern discovery is the third step in which the cleaned/filtered log file generated in the pre-processing step is used to discover web usage patterns.

3.1.4 Pattern Analysis

Pattern analysis is the fourth step in web usage mining process. In this phase extracted patterns are analyzed through OLAP tools, intelligent agents and knowledge management query techniques to sort out the monotonous rules/patterns [2].

3.1.5 Application of Analyzed Patterns

This is the final step of web usage mining process. In this phase different patterns analyzed from analysis phase are used for different applications such as: personalization, customer satisfaction, web site design, system improvement etc.

3.2 Web Usage Mining Technique

Four main Web usage mining techniques are shown here: association rules, sequential patterns, supervised classification and clustering.

3.2.1 Association Rules

Association rules are if/then statements that help to uncover relationship between seemingly unrelated data in a relational database or other information repository. The target is to find pages that are accessed together by majority of the user and hence should be linked in a proper way in order to maximize user satisfaction by providing to the user the access flow they expect.

3.2.2 Sequential Patterns

Frequent subsequence among large amount of sequential data is discovered using Sequential patterns. Sequential patterns are employed to find sequential navigation patterns that appear in users' sessions frequently, during Web usage mining process.

3.2.4 Supervised Classification

Supervised classification can be used in predicting the class to which an object or individual is likely to belong. Supervised classification is suitable to apply if the data is known to have a small number of classes or group, the classes or groups are already known and some training data with their classes known is available.

3.3 Clustering

Clustering techniques find groups of similar items among large amount of data on the basis of a general idea of distance function which computes the similarity between groups. Clustering is similar to classification but, in contrast to supervised classification, clustering is useful when the groups in the data are not already known and the training data is not available.

IV. CONCLUSION AND FUTURE WORK

Web log mining process is used to discover hidden and interesting patterns which are frequently accessed by the most of users of particular web site, which is applicable to find solutions of many real world problems. This paper studies a complete overview of web usage mining and various techniques of data mining that can be employed for web log mining for finding frequent patterns from web logs. The main research challenge of this field is finding association among different user's access patterns.

REFERENCES

[1] Gupta, A., Arora, R., Sikarwar, R., & Saxena, N. 2014. Web Usage Mining Using Improved Frequent Pattern Tree Algorithms. International Conference on Issues and Challenges in Intelligent Computing Techniques (pp. 573-578). IEEE. DOI: 10.1109/ICICICT.2014.6781344

[2] Sisodia, D.S., & Verma, S. 2014. Web Usage Pattern Analysis Through Web Logs: A Review. International Joint Conference on Computer Science and Software Engineering (pp. 49-53). IEEE. DOI: 10.1109/JCSSE.2012.6261924

[3] More, S. 2014. Modified Path Traversal for an Efficient Web Navigation Mining. International Conference on Advanced Communication Control and Computing Technologies (pp. 940-945). IEEE. DOI: 10.1109/ICACCCT.2014.7019232

[4] Bhargav, A., & Bhargav, M. 2014. Pattern Discovery and Users Classification Through Web Usage Mining. International Conference on Control, Instrumentation, Communication and Computational Technologies (pp. 632-636). IEEE. DOI: 10.1109/ICCICCT.2014.6993038

[5] Jia-Ching, Ying, Chu-Yu, Chin & Tseng, Vincent S. 2012. Mining Web Navigation Patterns with Dynamic Thresholds for Navigation Prediction. International Conference on Granular Computing (pp. 614-619). IEEE. DOI: 10.1109/GrC.2012.6468696

[6] Gupta, G. K. (2009). Introduction to Data Mining with Case Studies. Clayton, Australia: Prentice- Hall Of India Pvt. Limited.