Elysium Journal
Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

WEB USAGE MINING: IMPROVE THE USER NAVIGATION PATTERN USING FP-GROWTH ALGORITHM

K.Dharmarajan[1], Dr.M.A.Dorairangaswamy[2]

[1]Research and Development Centre, Bharathiar University, Department of Information Technology, Vels University, Chennai, India. E-Mail id: dharmak07@gmail.com

[2]Professor, St. Peter's University, Chennai, India .E-Mail id: drdorairs@yahoo.co.in

Abstract: In this paper main goal of web usage mining is to understand the behavior of web site users through the process of data mining of web access data. Knowledge obtained from web usage mining can be used to enhance web design, introduce personalization service and facilitate more effective browsing. Web access log analysis is to analyze the patterns of web site usage and the features of user's behavior. Weblog data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link and this information can be used in several applications like adaptive web sites, modified services, customer summary, pre-fetching, generate attractive web sites etc. In this paper we are using the FP-growth algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user's interest.

Keywords: Web Usage Mining; Fp-Growth Algotithm; Website; Assocation rule.

1. INTRODUCTION

Internet is an enormous repository of web pages and links. Web pages provides huge amount of information for Internet users. Today, data mining techniques are used by many companies to focus the customer retention[1]. Financial, artificial intelligence, communication and marketing organization are the companies using the data mining techniques. Web usage mining is one of main application of mining techniques in logs. There is tremendous growth and growth in internet. Users' accesses are documented in web logs. So on to the web data and forecast the user's visiting behaviors and obtains their interests by investigating the samples. The log files are files that contain information about website visitor activity[3]. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site information on his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text. Weblog mining is a special case of usage mining, which mines Weblog entries to discover user traversal patterns of Web pages.

2. WEB USAGE MINING PROCESS

The main aim of the innovation system is to find web user clusters from web server log files [2]. These discovered clusters show the characteristics of the underlying data distribution. Clustering is useful in characterizing user groups based on patterns, categorizing web documents that have similar functionalities [4].

This method allows for the collected works of Web log information for Web pages. This usage data provides the paths leading to accessed Web pages [5]. This information is often gathered automatically into access logs via the Web server

Web Usage Mining is a four-step process. The first step is data collection, the second step is data pre-processing, the third step is pattern discovery and the last step is pattern analysis.

2.1 PREPROCESSING

The pre-processing stage involves cleaning of the click stream data and the data is partitioned into a set of user transactions with their respective visits to the web site [6]. "Consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery". This step can break into at least four sub steps.

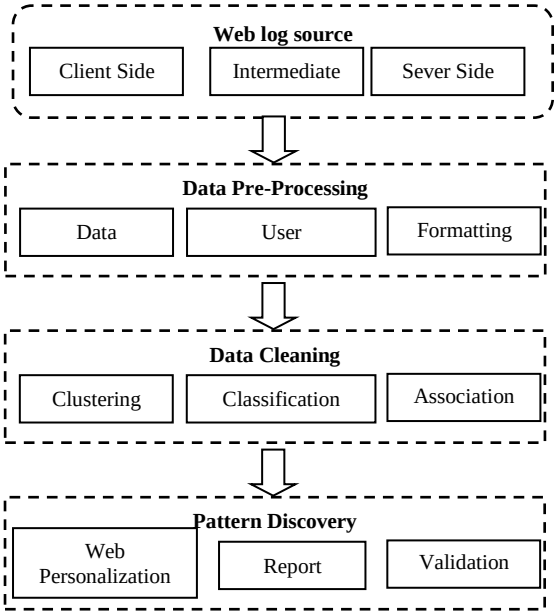

Fig. 1 Shows the Web Usage Mining Process

2.2 DATA CLEANING

It is the process of removing irrelevant items such as jpeg, gif, sound files and references due to spider navigation to improve the quality of analysis[7]. User Identification is the process of identifying users by using IPaddress and user agent fields of log entries. A user session is considered to be all of the page accesses that occur during a single visit to a Web site.

2.3 Pattern Discovery

Draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

2.4 PATTERN ANALYSIS

Pattern Analysis is the final stage of WUM (Web Usage Mining), which involves the validation and interpretation of the mined pattern.

3. WEB LOG FILES

Web Log Files are files that contain information about website visitor activity. Log files are created by web servers automatically [8]. Each time a visitor requests any file (page, image, etc.) from the site, information of his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text.

3.1 TYPE OF WEB LOG FILE

Access log file: Data of all incoming request and information about client of server. Access log records all requests that are processed by server.

Error log file: list of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log [2].Access and error logs are mostly used, but agent and referrer log may or may not enable at server.

3.2 WEB LOG FILE FORMAT

- W3C Extended log file format
- NCSA common log file format.
- IIS log file format

4. FP-GROWTH ALGORITHM

The FP-growth algorithm produces frequent data sets from FP Tree by navigating in a bottom up approach [6]. This method decreases the total number of user data sets by generating a compacted type of the database in terms of an FP-tree[7]. It is frequent information and allows for the effective discovery of frequent data sets. It is a two-step approach and faster than other association mining algorithms.

Step 1: Create a compact user navigation called the FP-tree.
It is built using 2 passes over the data-set.
Step 2: Extracts frequent set item from FP-tree .Traversal through FP-Tree.

Algorithm:
Input: A web log file for the www.srivaarielectricals.com dataset in .csv file format, represented by FP-tree constructed and a rare support threshold.
Output: The dataset of frequent patterns.
Method: call FP-growth (FP-tree, null).
Procedure FP-growth (Tree, x)
{
If Tree contains a single prefix path then
Prefix-path FP-tree
{
Let S be the single prefix-path part of Tree;
Let M be the multipath part with the top branching node replaced by a null root;
for each combination (denoted as ß) of the nodes in the path S do
Generate pattern ß ∪ x with support = rare support of nodes in ß;
let freq pattern set(S) be the set of patterns so generated;
}
else let M be Tree;
for each item xi in M do
{
Generate pattern ß = xi ∪ x with support = xi .support;
Construct ß's conditional pattern-base and then ß's conditional FP-tree Tree ß;
if Tree ß ≠ Ø then
call FP-growth(Tree ß , ß);
let freq pattern set(M) be the set of patterns so generated;
}
return(freq pattern set(S) ∪ freq pattern set(M) ∪ (freq pattern set(S) × freq patternset(M)))
}

4.1 WEB MINING TOOLS –RAPID MINER

This article, we implement how the weblog data which can used in the methodology of web usage mining and user pattern analysis in the Rapid Miner environment[6] .It is software developed using Java programming language. Rapid Miner is the user friendly data mining Tool. Which is used to analyze the webpage visitor details.

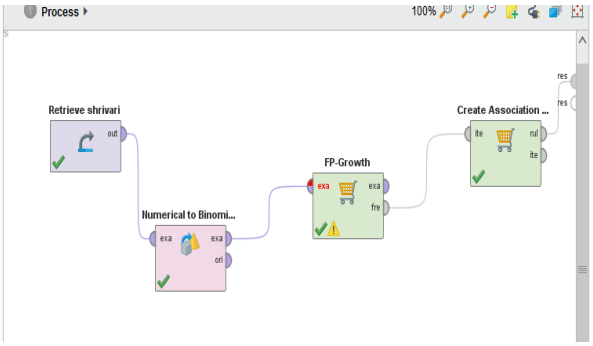

Fig. 2. Flow of FP-Growth Algorithm

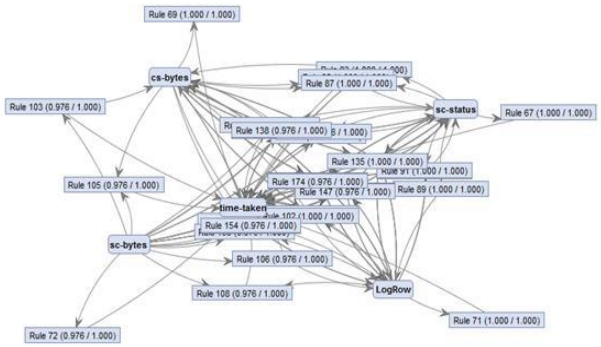

Fig.3. Result of FP-Growth Algorithm association rule mining

P-ISSN: 2347-4408
E-ISSN: 2347-4734

4.2 MINING FRIQUENT PATTERNS FROM WEB LOGS

The problem we will solve is to mine frequent patterns in web logs. As web logs are often very large in size but sparse in density, the efficiency of frequent pattern mining algorithm is important. To investigate the performance of FP-growth in on web logs.[1][6]. The FP-growth implementation we used was Rapid Miner Software All experiments were conducted on a Intel system in the School of Computing Science and Information Technology Vels University.

The system has a main memory of 4.0 GB. CPU Speed is i3-4005U CPU @1.70GHz 1.70 GHz. The operating system is Windows 7 Ultimate 64bit operating system. Timing includes reading, computation and outputting all FPs to disk. The time performance of FP-growth are shown in Figure 2. We can see that FP-growth is always better than other algorithm, which is a dense datasets and the FPs from it are relative longer .When minimal support threshold is lower than 0.02%, the FPs are longer with length more than 5 and can be 11 and FP-growth has better performance.

4.3 APPLICATIONS OF WEB USAGE MINING

Each of the applications can benefit from patterns that are ranked by subjective interesting. Web usage mining is used in the following areas:

1) Web usage mining offers users the ability to analyze massive volumes of clickstream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for web personalization, e-CRM and other interactive marketing programs.

2) Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages

3) By determining frequent access behavior for users, needed links can be identified to improve the overall performance of future accesses .

4) Information concerning frequently accessed pages can be used for caching .

5) In addition to modifications to the linkage structure, identifying common access behaviors can be used to improve the actual design of Web pages and to make other modifications to the site.

6) Web usage patterns can be used to gather business intelligence to improve Customer attraction, Customer retention, sales, marketing and advertisement, cross sales.

7) Mining of web usage patterns can help in the study of how browsers are used and the user's interaction with a browser interface.

8) Usage characterization can also look into navigational strategy when browsing a particular site.

9) Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web.It helps in improving the attractiveness of a Web site, in terms of content and structure.

10) Performance and other service quality attributes are crucial to user satisfaction and high quality performance of a web application is expected[2].it provides a key to understanding Web traffic behavior, which can be used to deal with policies on web caching, network transmission, load balancing, or data distribution.

11) It is used in e-Learning, e-Business, e-Commerce, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, and Digital Libraries.

12) Web usage minin used in determination of common behaviors or traits of users who perform certain actions, such as purchasing merchandisee and it can be used in usability studies to determine the interface quality.

13) Web usage mining can be used in Counter Terrorism and Fraud Detection, and detection of unusual accesses to secure data.

14) Web usage mining can be used in network traffic Analysis for determining equipment requirements and data distribution in order to efficiently handle site traffic.

4.4 USER PATTERN DISCOVERY

In this case, episodes are either all of the page views in a server sessions that the user spent a significant amount of time viewing, or all of the navigation page views leading up to each content page view. The details of how a cutoff time is determined for classifying a page view as content or navigation are also contained in [4]. The click-stream or click-flow for each user is divided into sessions based on a simple thirty-minute timeout. The notion of what makes discovered knowledge interesting has been addressed in [6].Pattern-form defines what type of patterns a measure is applicable to, such as association rules or classification rules. The representation dimension defines the nature of the framework, such as probabilistic or logical. Scope is a binary dimension that indicates whether the measure applies to single pattern, or to the entire discovered set. The final dimension, class is also a binary dimension that can be labeled as subjective or objective.

The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes [5]

5. CONCLUSIONS

In order to make a website popular among its visitors, System administrator and web designer should try to increase its effectiveness because web pages are one of the most important advertisement tools in international market for business. The obtained results of the study can be used by system administrator or web designer and can arrange their system by determining occurred system errors, corrupted and broken links. In this study, analysis of web server log files FP-growth algorithm is the explosive quantity of lacks a good candidate generation method. Future research can combine FP-Tree. The work can also be extended to extract information from image files. The new approach requires minimum repeated database scan for frequent pattern mining in web usage mining. It will reduce the time and space execution. In this paper a new technique is proposed to discover the web usage patterns of websites from the server log files with the foundation of clustering and improved Symantec.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

REFERENCES

[1] Neelima, G., and Sireesha Rodda. "Predicting user behavior through sessions using the web log mining." 2016 International Conference on Advances in Human Machine Interaction (HMI). IEEE, 2016

[2] J Neha Goel, C.K. Jha,,. Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool, International Journal of Computer Applications 180, 250.

[3] Abercrombie, N., Hill, S., & Turner, B. S. (1980). The dominant ideology thesis. London: Allen & Unwin.

[4] York Manuscripts Conference, & York Centre for Medieval Studies. (1989). Latin and vernacular: studies in late-medieval texts and manuscripts;[proceedings of the 1987 York Manuscripts Conference]. A. J. Minnis (Ed.). Brewer.

[5] Srivastava, Jaideep, et al. "Web usage mining: Discovery and applications of usage patterns from web data." Acm Sigkdd Explorations Newsletter 1.2 (2000): 12-23.

[6] Victor, S. P., and Mr M. Xavier Rex. "Analytical Implementation of Web Structure Mining Using Data Analysis in Educational Domain." International Journal of Applied Engineering Research 11.4 (2016): 2552-2556..

[7] Malarvizhi, S. P., and B. Sathiyabhama. "Frequent pagesets from web log by enhanced weighted association rule mining." Cluster Computing 19.1 (2016): 269-277.