ISSN (Print) : 0974-6846
**Indian Journal of Science and Technology,** *Vol 8(29), DOI: 10.17485/ijst/2015/v8i29/83327, November 2015* ISSN (Online) : 0974-5645

# A Framework for an Efficient Knowledge Mining Technique of Web Page Re-organisation using Splay Tree

**Ananthi Sheshasaayee and V. Vidyapriya***

PG and Research Department of Computer Science, Quaid-E-Millath Government College for Women (A), Chennai- 600002, Tamil Nadu, India; ananthi.research@gmail.com, vidyapriya.research@gmail.com

## Abstract

**Background/Objectives:** Web Usage Mining (WUM) is one of the categories of web mining that identifies user patterns of web data, with the help of knowledge acquires from web logs. **Methods/Statistical Analysis:** The structure of the web site has to be reorganised to suit the user requirements to facilitate the user for the required pages with less page access delay. The Splay trees are efficient balanced trees when total running time is the measure of interest. **Findings:** The motive of mining is to find users' access models automatically and quickly from the vast Web log data, like frequently accessed pages and time spent on those pages. Web usage mining consist of three phases namely Data pre-processing, Pattern discovery and Pattern analysis. Pre-processing tasks are used to translate unprocessed log files which are composed from web server into structured log file data. Pre-processed log file data are used for further process of web usage mining. This paper present the pre-processing technique and an approach for re-organisation of website based on the access frequency of web pages using splay tree structure. **Application/Improvements:** The nodes of the splay tree can be added with the information about priority of recently accessed web pages to reduce the page access delay.

**Keywords:** Pre Processing, Splay Tree, Web Log, Web Site Reorganisation, Web usage Mining

## 1. Introduction

Web data mining can be defined as the discovery and analysis of useful information from the WWW data. There are several works around the survey of data mining on the Web. The Web mining puts down the roots deeply in data mining. The formless feature of Web data creates more complexity in the process of Web mining. A drastic growth in on-line information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services[1]. Web mining can be classified into three areas of interest based on which part of the Web to mine: Web content mining, Web structure mining and Web usage mining as shown in Figure 1.

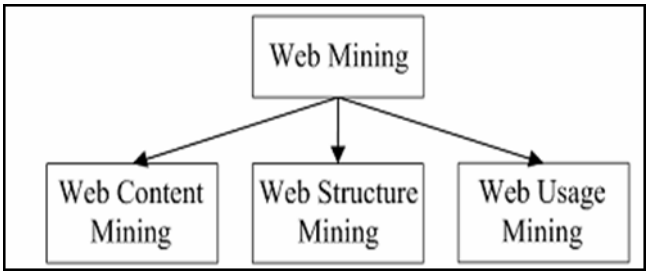


**Figure 1.** Classification of Web mining.

In practice, the three Web mining tasks above could be used in isolation or combined in an application. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is defined as the process of inferring knowledge from the World Wide Web organization and links between references and source from where the

* *Author for correspondence*

request is made in the Web. The web usage mining is the process of extracting interesting patterns in web access logs[2,3]. It is also known as Web Log Mining.

## 2. Phases of Web usage Mining

Web usage mining is the application of data mining techniques on large web log data sets in order to extract useful knowledge about users behavioural[4,5]. Web log server is the primary source for Web log files. Whenever users do some task or request on the web pages, then the log file entries are created in the web server.

### 2.1 Web usage Mining

Analysis and find out meaningful patterns from data generated by client-server transactions on one or more Web servers includes following sources of data[6]:

- Automatically generated data stored in server access logs, referrer logs, agent logs, and Client-side cookies.
- E-commerce and product-oriented user events
- User profiles and/or user ratings.
- Meta-data, page attribues page content, site structure.

Web usage mining focuses on techniques that could predict user behaviour while the user interacts with the Web[7]. The mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally we could classify them into the usage data that reside in the Web clients, proxy servers and servers[8]. The Web usage mining process can be regarded as a three-phase process as shown in Figure 2, consisting of the data preparation or pre-processing, pattern discovery and pattern analysis phases.

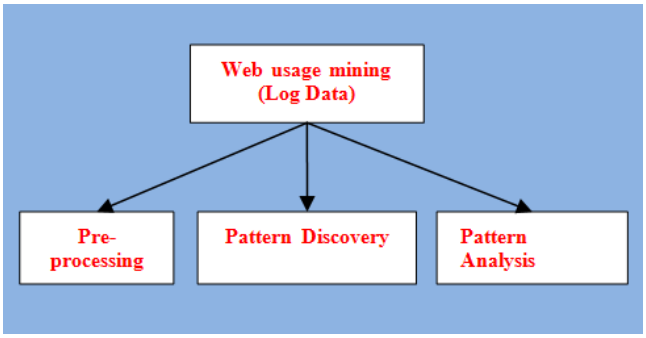


**Figure 2.** Phases of Web usage mining.

In the first phase, Web log data are pre-processed in order to identify users, sessions, page views, and so on. In the second phase, statistical methods, as well as data mining methods such as association rules, sequential pattern discovery, clustering, and classification are applied in order to detect interesting patterns. These patterns are stored so that they can be further analysed in the third phase of the Web usage mining process.

### 2.2 Web Log Format

**Table 1.** Web Log Format

| Log File Fields |
|---|
| date |
| time |
| s-site name |
| s-ip(Server IP) |
| cs-method |
| cs-uri-stem(URI stem) |
| cs-uri-query |
| s-port |
| cs-username |
| c-ip |
| cs (User-Agent) |
| sc-status |
| sc-sub status |
| sc-win32-status |

A web server log file contains requests made to the web server, recorded in chronological order[9]. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. An extended common log format file is created by the web server to keep track of the requests that occur on a web site. The attributes of the log file are listed in the in Table 1.

## 3. Pre-processing

The web usage mining generally includes the following several steps: data collection, data pre-processing, and knowledge discovery and pattern analysis[10].

### 3.1 Data Collection

Data collection is the first step of web usage mining, the data authenticity and internality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality.

The collected log files have to be merged together to get a single file and required attributes are to be extracted from the merged file.

### 3.1.1 Merging of Log Files

Before cleaning server log files we need to merge all the log files which were down loaded from the web server. Following are the steps to merge server log files.

Step 1: Let $L_1$, $L_2$, $L_3$, …… $L_n$ are the server log files which are down loaded from web server and L is the merged file.

Step 2: Append each log files $L_n$ into L.

Step 3: Sort the target log file entries (L) based on date and time wise.

### 3.1.2 Extraction of Attributes from Log Files

Step 1: Remove the fields such as Software, Version and Date fields from Log file L.

Step 2: Remove the characters such as space, comma (,) and slash (/) etc. which are working as field separator in above target data log file and maintain space character as a separator to separate the fields of the target log file.

Step 3: Find out the required attributes fields from above target log file. Following are the required attributes fields Date, Time, cs-method, cs-uri-stem, c-ip and sc-status.

Step 4: Create a data base table with above mentioned required attribute fields.

Step 5: Fetch the respective attribute field's values and store it into above created data base table.

## 3.2 Data pre-processing

Some databases are insufficient, inconsistent and including noise. The data pre-treatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine[11]. The data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion as shown in Figure 3.

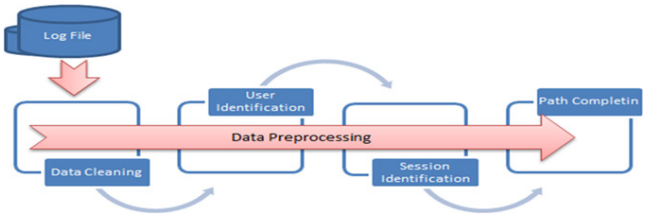


**Figure 3.** Data pre-processing topology.

### 3.2.1 Data Cleaning

The purpose of data cleaning is to eliminate irrelevant items; these kinds of techniques are of importance for any type of web log analysis not only data mining. Follow the below steps to eliminate unwanted entries from target log file data[12].

Step 1: Remove target log file entries which are having file extension .gif, .jpg and .css.

Step 2: Delete entries cs-method field value other than GET and POST.

Step 3: Remove log file entries which were having sc-status code value other than 200 to 299.

### 3.2.2 User Identification

User identification is one of the important tasks in the pre-processing techniques. Table 2 explains the various methods used to find out the unique user from log file entries.

**Table 2.** User identification methods

| Methods |
|---|
| User Identification by IP address |
| User Identification by authentication data |
| User Identification by cookies |
| User Identification by client information |
| User Identification by site topology |

### 3.2.3 Session Identification

Once user is identified, it is very important to identify the sessions. Session Identification is nothing but set of requests done by identified user for some defined period of time to the particular web site[13]. Following are the two techniques used for find out session for the single user.

## 3.3 Session Identification by user Authentication Data

By using user authentication data like cookies mechanism and embedded session id we can identify the session of the single user.

## 3.4 Heuristic Techniques

Following are some of Heuristic method used for Identify Session of particular user.

### *3.4.1 By the Time gap*

When the time gap between two consecutive requests by the same user is greater than certain threshold then a new session is created.

---

**New Session Creation = S. $t_{n+1}$ - S. $t_n$**

---

Where S.$t_{n+1}$ and S.$t_n$ are the time stamp of two consecutive requests.

### *3.4.2 By the Referrer Attributes*

Session can be identified by referrer attributes in log file formats. Suppose p and q is two requests from consecutive pages by same user. If referrer of q was invoked previously in that session S otherwise a new session is created with q as a first requested page.

### *3.4.3 By the Time Spend on Observing Page*

User spent time on Information page and navigation page. Information Page is user interested page and Navigation page is page through which user reaches the interested page.

## 3.5 Knowledge Discovery

Use statistical method to carry on the analysis and mine the pre-treated data. We may discover the user or the user community's interests then construct interest model. At present, the existing machine learning methods includes clustering, classifying, the relation discovery and the order model discovery[14]. Each method has its own advantages and disadvantages.

## 3.6 Pattern Analysis

The purpose of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user[15]. First delete the less significance rules or models from the interested model. After that, use technology of OLAP to carry on the comprehensive mining and analysis[16]. Once more, let discovered data or knowledge be visible. Finally, provide the characteristic results for the knowledge mining.

# 4. Frame Work of Web Page Reorganization

Splay trees are Self-balancing or self adjusting binary search tree. Whenever the accessing of a node is done on the tree for following operations such as retrieval, rotation, insertion and deletion, it has an update rule, with O (log n) amortized time. The rule is to get recently accessed node becoming the root of the modified tree[17].

*Framework*:

Consider a target Website with m number of page id's namely $Pg_1$, $Pg_2$, $Pg_3$ . . . . $Pg_m$

Let L: $L_{f1}$, $L_{f2}$, $L_{f3}$ . . . . . $L_{fm}$. be the pre-processed log files of the target Website, where each $L_{fi}$ consists of the following fields Page Name, User IP address, and Time Spent.

Let the time spent on each page be: $t_{p1}$, $t_{p2}$, $t_{p3}$ … $t_{pm}$.

Let Min (t) is the Minimum time threshold.

Then, the find frequency of access of each page $F_{pg1}$, $F_{pg2}$… Fpgm.

$F_{pgi}$ is computed by adding the number of times the page request appears in the log file, that has time spent value $t_{pi}$ >= Min (t).

Let $MF_{pgi}$ be the Minimum Frequency Threshold.

We define the most frequently accessed page as the page with $F_{pgm}$ > = $MF_{pgi}$.

For the above Frame Work of the web pages organization, the initial Binary Search Tree (BST R) is constructed by the pages with the Minimum Frequency Threshold based on the initial preferences of the Web pages with Page ID's.

To Re-organize the website, the data structure technique namely Splay Tree is proposed.

The following Algorithm explains the splay tree approach in detail.

| *Algorithm for Splay Tree Technique* |
|---|
| Step 1: Compute the frequency of access for each page in the pre-processed log file. |
| Step 2: Consider the pages that have at least Minimum Frequency threshold. |
| Step 3: Rearrange the list in such way that the recently accessed pages are at the end of the list. |
| Step4: Construct the Splay tree where the most recently Accessed page is at the root |

The initial BST is splayed in the order on recently accessed Pages.

# 5. Conclusion

An efficient Web page Mining technique is essential for Reorganize Web Pages in such a way that users can able to access the recently accessed page of the web site with less time delay. In this paper an initial Binary Search Tree has been developed based on the pre-processed log file data

of the web site. An algorithm for the splaying the initial BST is developed. Using this algorithm, the initial BST is splayed in the order based on recently accessed page so that recently accessed page will become the top of the root. Future work would focus on optimizing the tree reordering time, inherent to the splaying operation of the proposed data structure.

## 6. References


1. Rana C. A study of web usage mining research tools. Int J Advanced Networking and Applications. 2012; 03(6):1422–9. ISSN: 0975-0290.
2. Pamnani R, Chawan P. Web usage mining: A research area in web mining. 2010. p. 1–5.
3. Facca FM, Lanzi PL. Mining interesting knowledge from weblogs: A survey. Data and Knowledge Engineering. 2005; 53(3):225–41.
4. Han Q, Gao X, Wu W. Study on web mining algorithm based on usage mining. Kunming: 9th International Conference on Computer-aided Industrial Design and Conceptual Design. CAID/CD. 2008 Nov 22-25. p. 1121–4.
5. Dong Y, Zhang H, Jiao L. Research on application of user navigation pattern mining recommendation. Intelligent Control and Automation, WCICA. Dalian: The Sixth World Congress. 2006; 2:6106–10.
6. Paik HY, Benatallah B, Hamadi R. Dynamic restructuring of e-catalog communities based on user interaction patterns. World Wide Web. 2002; 5(4):325–66.
7. Maier T, Reinartz T. Evaluation of web usage analysis tools. Kunstliche Intelligenz. 2004; 18(1):65–7.
8. Han Q, Gao X. Research of distributed algorithm based on usage mining. Moscow: 2nd International Workshop on Knowledge Discovery and Data Mining. WKDD. 2009 Jan 23-25. p. 211–4.
9. Eirinaki M, Vazirgiannis M,Varlamis I. SEWeP: Using site semantics and a taxonomy to enhance the Web personalization process. SIGKDD '03. 2003 Aug 24-27. p. 99–108.
10. Mobasher B, Cooley R, Srivastava J. Automatic personalization based on Web usage mining. Communications of the. ACM. 2000; 43(8):142–51.
11. Masseglia F, Poncelet P, Teisseire M, Marascu A. Web usage mining: Extracting unexpected periods from Web logs. Data Mining and Knowledge Discovery. 2008; 16(1):39–65.
12. Kumar BS, Rukmani KV. Implementation of Web usage mining using APRIORI and FP growth algorithms. International Journal of Advanced Networking and applications. 2010; 1(06):400–4.
13. Srivastava M, Garg R, Mishra PK. Preprocessing techniques in web usage mining: A survey. Int J Comput Appl. 2014; 97(18):1–9.
14. Rao M, Kumari M, Raju K. Understanding user behavior using web usage mining. Int J Comput Appl. 2010; 1(7):55–61.
15. Berendt B. Web usage mining, site semantics, and the support of navigation. KDD workshop on web mining for ecommerce challenges and opportunities. 2000. p. 83–93.
16. Nasraoui O, Soliman M, Saka E, Badia A, Germain R. A Web usage mining framework for mining evolving user profiles in dynamic web sites. IEEE Trans Knowl Data Eng. 2008; 20(2):202–15.
17. Sleator DD, Tarjan RE. Self-adjusting binary trees. Journal of ACM. 1985; 32(3):652–86.