Journal of Theoretical and Applied Information Technology
31st January 2014. Vol. 59 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ALL-INCLUSIVE REVIEW ON VARIOUS TECHNIQUES OF WEB USAGE MINING

[1]P. SENTHIL PANDIAN, [2]Dr.S. SRINIVASAN

[1]Associate Professor & Head, Department of Information Technology, Latha Mathavan Engineering College, Madurai, Tamilnadu, India.

[2]Professor & Head, Department of Computer Science and Engineering, Anna University, Regional Office, Madurai, Tamilnadu, India.

E-mail: [1]psenthilpandian@gmail.com, [2]sriniss@yahoo.com

ABSTRACT

Numerous users use World Wide Web (WWW) as their default resource for obtaining knowledge and many organization need to empathize their customer's preference, behavior and future need to improve their business. Web usage mining is a part of web mining and an active research topic. The main goal is to find, model and analyze the behavioral pattern of the users. The captured patterns are represented as a collection of objects, or as pages, which are frequently used or accessed by a set of users having common interest. The primary advantages are the extraction of segmented data from server logs and discover desired usage patterns from the web. The importance of application level data can be highly distinguished from web server data via web usage mining. Enormous outstanding techniques have been developed to improve the extraction process. This paper presents a survey of recent methodologies in the field of web usage mining.

Keywords: *Data Mining, Web Usage Mining, Behavior Pattern, Preprocessing And Extraction Technique*

1. INTRODUCTION

Web mining is an area of data mining that concentrates more on extracting the information from WWW, which are useful and interesting. Web mining can be further classified as three different areas as shown in the Figure 1.

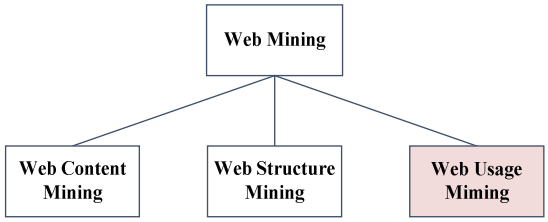

Figure 1: Taxonomy of Web Mining

Web content mining focuses on raw information available on the web pages. Likewise, the web structure mining deals with the structure of the web site. The web usage mining concentrates on the extraction of data that are presented at log files of server. These files primarily incorporate the logs, which are collected when the web servers are accessed by the user, and they are delineated as standard formats. Various techniques are proposed to represent the collected information as the standard format. Typical applications are widely depended on those modeling techniques, such as adaptive web sites, web personalization and user modeling. Web mining, specifically web usage mining has become a most flourishing research area for the past 10 years. Approximately, 150 papers have been published before 2000. Since 2000, enormous papers have been published in this area showing the increased rate of interest. This postulate presents a survey of resent development in the field of web usage mining.

The web usage mining is highly concentrated due to the effective use in various web-oriented applications. User activities are recorded on the log files. Following is the three different types of log files namely client log files, server log files, and proxy log files. In this survey, it is found that the server log files are used widely for usage mining process. Since, it provides the most accurate and required details of the user than the other two log files for the process of usage mining. The server log files are considered incomplete since it does [34] not record the requests that are sent for cashed

Journal of Theoretical and Applied Information Technology
31st January 2014. Vol. 59 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pages. In general, there are three different type of web server log files used to capture the user behavior on the web site [32] as listed below.

- Common log file format (NCSA): It is a standardized but it cannot be customized
- Extended Log Format (W3C): Flexible and can be customized depending on the requirement
- IIS Log Format (Microsoft): It is also not customizable

The aforementioned log files are used as input to the usage mining. The process of usage mining can be classified into following inter-dependent stages. Figure 2 shows the overall flow of web usage mining.

- Pre-processing
- Pattern discovery
- Pattern analysis

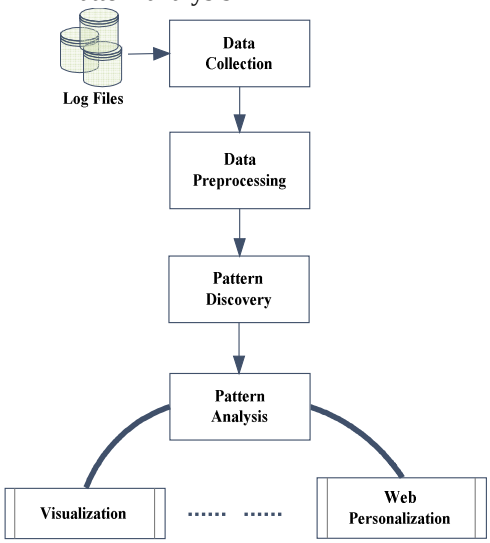

Figure 2: Overall flow of Web Usage Mining

Data collection is the process of obtaining the information from the log files. Collected data size is large as well as the number of irrelevant and raw entries. Such files cannot be used directly for the process of usage mining. Therefore, it becomes mandatory to preprocess those documents, which helps the subsequent steps of the usage mining. These files are given as input to the pattern discovery phase where the techniques like association discovery, classification, clustering, and sequential pattern discovery are applied to the data to find the user information. The patterns detected in the pattern discovery step is analyzed and presented in a form that is easily understandable by the users.

Rest of the paper is structured as follows: section 2, deals with the study of information that is related preprocessing phase. Pattern discovery and analysis discussed in section 3. Finally, section 4 concludes the paper.

2. REVIEW ON PREPROCESSING

This section deals with the available preprocessing technique. The original log file cannot be used by the usage mining process directly, which was due to the huge amount of irrelevant data present in the log files.

Graphs play an important role in [25] to disclose the information about the interest. Here, access information was collected from the search keyword given by user and the client log file. From these data a graph was generated. Page rank algorithm was applied to mine user's interest. Subsequently unimportant and unnecessary nodes were eliminated from the graph. Finally, the graph was subdivided and user behavior was determined from these sub-graphs. With this, it was decided that no special techniques were applied to remove the noise rather than removing the unwanted nodes.

In [38], the log files are classified using decision trees. Before classification they are preprocessed to clean the data. The unsuccessful HTTP status code and request method except GET and POST are treated as irrelevant data. These data are removed and the filtered data are converted to the decision tree. The instances are arranged down the tree from the root to the leaf node. Enhanced version of decision tree is used to classify the log file.

Pre-processing was considered as laborious in [13] since it consumes 80% of total mining time. However, it was essential to implement preprocessing since it makes the web mining process more efficient and effective. A cleaning method was proposed to remove irrelevant and unimportant links from the web log files. As a result, a log file that is filtered and therefore has only the essential links was obtained by comparing the link table and the raw log file. This technique can be applicable only in a particular domain. Here, only data cleaning was carried out through this technique.

Data cleaning and user identification process was carried out in [34], which concentrated on all type of log files. Consider an example to the technique proposed in [34] for a server log file.

- Data cleaning: Pages that contain images are removed
- User identification: Accurate detection of user IP address, Operating system, and user agent parameters are used

Here, session identification, an important phase in preprocessing, which was not applied to the user.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Authors of [29] have focused the server log file and preprocessed it through cleaning, session identification and user identification. Initially, the raw web log files are converted to graph.

Table 1: Summary Of Preprocessing Technique

PAPER	LOG FILE SOURCE	PREPROCESSING TECHNIQUE	ALGORITHM USED	AUTOMATED
[25]	Client Log file	Data cleaning	Page Rank	YES
[13]	Server Log	Advance Data Cleaning Filtering Data Visualization	NA	NA
[34]	Server Log	Data Cleaning User Identification	NA	NA
[29]	Server Log	Data Cleaning User Identification Session Identification	NA	NA
[20]	Server Log	Path Completion Data Cleaning User Identification Session Identification	Proposed	NA
[27]	Server Log	Data Cleaning Data Structuration (User Sessions)	NA	LODAP
[38]	Server Log	Data Filtering Cleaning Classification	Decision Tree	NA
[42]	Server Log	Data Cleaning log identification session identification	Page Rank	NA

The graph is converted into adjacency matrix, which provides the user's interests. Significant pages were computed and the log files are cleaned to determine the information about cookie. The individual users were identified via the user-id, content-related Meta data and requested pages. Along with this session was identified through visited pages of user and 30 minutes timeout. Here session classification was not discussed, which helps to preprocess the log file.

An effective technique was proposed in [20], where the cleaning was carried to remove the irrelevant information. In addition to cleaning, path computation, session and user identification was also employed. A tool named LOADP was developed in [27], which takes raw web log file as its input and generates user session as its output. It is also considered as effective tool for preprocessing the log files of usage mining process. In [28] Fuzzy C-Means algorithm was proposed for clustering the session. It was also carried in [30] using Particle Swarm Optimization algorithm. They have also used the general Euclidean Distance method to measure the distance between the clusters. Following the above approach [42] has focused on the clustering technique. It used server log files and they were preprocessed before the clustering. Data cleaning, log identification, and session identification are the techniques carried out to clean the data and made the log data ready for clustering. The summary of the various preprocessing techniques are enlisted in Table 1. A new technique named FPCM, which was a hybridization of Probabilistic C-Means algorithm and Fuzzy C-Means algorithm. Similarity measure is a major issue in clustering the documents, which was addressed in [26] through the equation (1).

$$Sim(c_1, c_2) = \frac{2*depth(LCA(c_1,c_2))}{depth(c_1)+depth(c_2)} \quad (1)$$

In equation (1) the depth refers the number of pages and the c_1 and c_2 denotes the concept hierarchy.

3. PATTERN DISCOVERY AND ANALYSIS

As explained in [5], web usage data can be discovered and extracted from various sources. Web log file is one of the most important and most widely used resources for mining the user behavior patterns. However, the log files have several complexities and problems that should be tackled. The noisy data and other irrelevant data are removed using the preprocessing and thereafter the web usage mining technique is implemented to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

mine usage patterns. The definition in [5], "Web usage mining (WUM) is the application of data mining techniques to discover usage patterns from Web data, in order to understand the needs of Web-based application". Following are the major pattern discovery process.

3.1 Association Rule Mining

Association rule mining was one of the most widely used techniques to discover the interesting patterns from huge data sets. As seen earlier, mining of usage pattern was a laborious work. In order to assist the discovery process association rules can be used. Various techniques and algorithms have been proposed to discover the association rules from the huge log files. Some of the existing works that were related to the finding the association rules have been discussed in this section.

Two different type of usage pattern namely the association rule and the sequential patterns fining have been discussed in [9]. Apriori and AprioriAll algorithms have been used to learn the association rules and sequential patterns from the session file. The total number discovered association rules are purely depends on the support and confidence threshold values. Likewise, support threshold plays a vital role in determining the number of discovered sequential pattern. The support threshold value was varied and the corresponding number of rules and patterns generated was shown in Table 2. Here, the value of threshold for confidence was set as default as 0.5 for association rules. As the threshold values increases the number of rules generated is decreased.

Table 2: Number Of Patterns Generated

Threshold	0.02	0.008	0.003	0.0028	0.002	0.001
Number of association Rules	2	39	732	4556	4,800,070	>1,000,000,000
Number of sequential Patterns	8	35	357	409	2,834	609,453

To refine the discovered patterns, i.e. to find the interesting pattern from the huge number of discovered pattern, the rules or sequential patterns are ranked according to the interesting measure. This paper has used the seven different interesting measures to detect the interesting patterns. The redundant patterns or rules were pruned based on the structural relationship between patterns.

FP-growth based algorithm to process the web log and generation of association rules has been proposed in the paper [36]. This paper has the following two algorithms.

1. FP-tree construction
2. FP-growth for mining frequent pattern

First algorithm takes a transactional database as its input and generates the FP-tree as its output. The database is scanned to determine the frequent item and the support value of each item. The tree was constructed with the selected frequent item based on the support value and the best access time. The output of algorithm 1, i.e. the FP-tree acts as the input to the second algorithm. Here, interesting measures were defined and based on the values for interestingness measure the frequent pattern was generated.

In [9] and [36] authors have detected the rules only based on the predefined support and confidence threshold. However, they were not more suitable for online based usage pattern mining. To overcome this problem a new algorithm named Adaptive Support Association Rule Mining (ASARM) was proposed in [10]. This algorithm mines the rules for a specific user not for all users. ASARM has the following advantages.

- The user interests were overlapped to determine the interest of the given user
- The interesting measure of one user can be used to determine another users interestingness measure even though they were not correlated

The process of ASARM algorithm has been divided into two different parts as follows.

1. ASARM1
2. ASARM2

The main process of ASARM1 was to choose the minimum support according to the interest ratio of the specific user and calls the ASARM. The ASARM2 mines the rules that are interested. Here number of rules generation can be controlled. Results show that the performance of the proposed algorithm functions better than the existing traditional algorithms.

The techniques that were discussed above consume more bandwidth. Therefore, it was

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

necessary to have an algorithm that consumes less bandwidth. The performance of the pattern mining can be improved by pre-fetching the information. This idea was demonstrated in [11], where the user's future requests for the documents were determined. This paper proposes a scheme termed Prefetch Enhanced Cache (PEC) for (a) Effective pre-fetching and (b) Coordination of caching and pre-fetching. The pre-fetching process is carried through the association rules. This paper also explains how the pre-fetching algorithm exploits the web log mining. The major advantage of the proposed method was that it was not affected by the high-order dependencies among the noise present in the user transaction or the document reference.

In [18], SGI's MineSet was used to mine the association rules from the generated session and URL files. A data file and schema file were given as input to the MineSet. The schema file has the information about the fields that were present in the data file. Whereas, the data file contains the information about the binary session vectors. In order to generate the association rules for the given web log, this algorithm generates the session and URL files in prior to those. Temporal interval based access pattern discovery from web log files was proposed in [41]. Large event set, uniform event set, and relational rules were the steps followed to discover the temporal rules of the given log file. The large event set contains the information of minimum support threshold that satisfies a user. Uniform event set denotes the set of events that has occurred continuously. These set are framed and used to generate the resultant temporal rules.

The tradition techniques for generation of association rules were not highly supportable to the discovery of user behavior pattern from the log files. In [37] a system was implemented to discover the association rules. Through this system, the problem of web usage association rule over-generation by pruning the rules was alleviated. Experiments showed that the discovered rules can be successfully sorted or ranked using the interestingness measure.

Table 3 presents a summary of all the papers and their technique used to discover the rules were presented.

Table 3: Technique used to discover rules

Paper	Technique
[9]	Apriori and AprioriAll
[36]	FP-growth
[10]	ASARM
[11]	PEC
[18]	SGI's MineSet
[37]	Proposed system

3.2 Semantic Web Mining

Semantic web and web mining were the two fast developing research areas, which were combined to generate the semantic web mining. Though various research were carried using association rule to discover the usage patter, their predictions were majorly depends on the prior beliefs. Therefore, they were needed to be completed with the understanding about what a site and its usage patterns were about. Such an approach was referred as semantic approach. While incorporating the semantic concept into the web mining faces a problem of conceptualizations. These problems were addressed and a proposal explaining about how semantic can enhance web usage mining was given in [14]. It also deal with the incorporating the knowledge about the usage behavior into the web hypermedia, which exceeds the semantic of sites. [14] specifies that a web usage information should take not only the details present in the server logs but also considers the meaning that it holds on accessing a sequence of web pages. Following three directions were considered for cooperation of semantic and web usage mining.

1. Ontologies of complex behavior development
2. Deployment of the developed ontology into the semantic web description and mining tools
3. Techniques and tools, which allows the incorporation of both the user's and expert's background knowledge into mining

So far the semantic modeling was applied only to the upper part of the life cycle. In [17], an approach was proposed to integrate the conceptual models in the lower part of the application life cycle, through make use of them in both usage and quality analysis. The web application development based on the model-driven approach consists of various steps as shown in Figure 3. Here, a C-log was created by combining sever log and conceptual schema of web application.

A framework was presented in [19] to enhance the record of web usage with the semantic. Here, authors have mapped all URL request to one or

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

more concepts of the ontology. Sessions were clustered depending on a specific user interests based on the semantically enhanced weblogs and to enhance the weblogs the association rules are mined. Another system was proposed in [16], which uses the usage log as well as the semantic of web page content. The content of the web pages were annotated using the conceptual taxonomy. System architecture is portrayed in Figure 4. Creation of C-logs refers to the extension of web usage logs that enclose context semantics. Conceptual hierarchy was used to annotate the semantic content, which enables to cluster and rank the web documents.

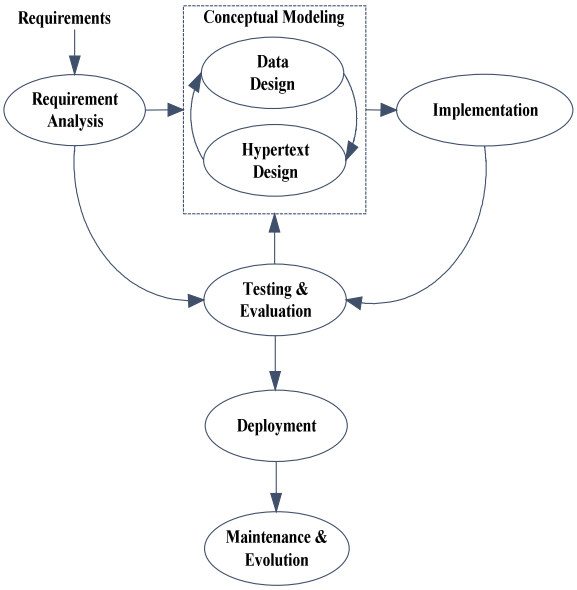

Figure 3: Phases In Development Of Web Application

The C-logs were analyzed using MINE RULE in [24] in order to identify the types of patterns, such as: most frequently visited page content, recurrent navigation paths, and anomalies. The case study in [24] showed that, embedding the conceptual information into log files improves the mining process.

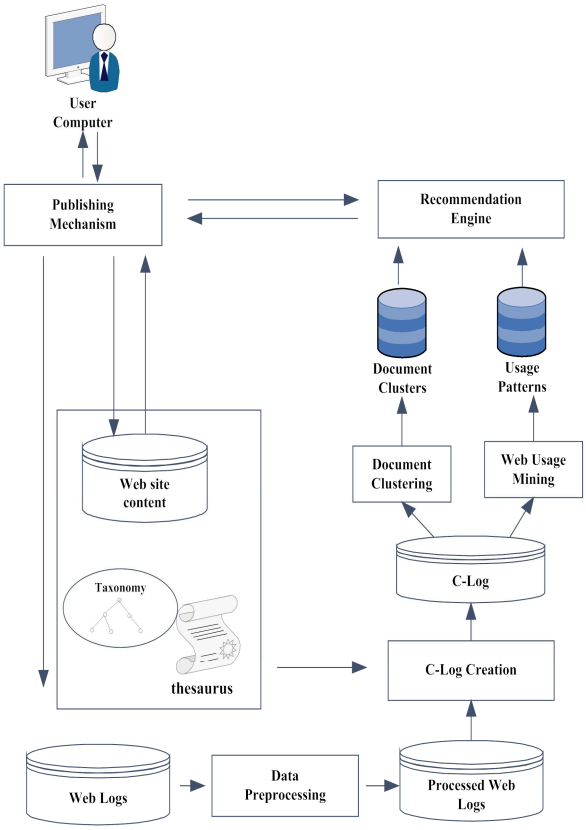

Figure 4: System Architecture Of SEWeP

3.3 Sequential Pattern Mining

The sequential pattern mining technique finds the inter-session patterns of sessions. The inter-session pattern refers to the set of items, which is followed by another time-ordered item set. Using this, it is possible to determine the future visit patterns. Other types of analysis that can be carried using the Sequential Pattern (SP) are as follows:

- Change point detection
- Trend analysis
- Similarity analysis

The sequential patterns can be used in web usage mining to determine the frequent navigational path. SP denotes the sequence of frequently occurring items, which occupies the large proportion of transactions. A sequence $\langle S_1, S_2, S_3, \dots\dots, S_x \rangle$ occurs in a transaction $T = \langle T_1, T_2, T_3, \dots\dots, T_y \rangle$ (where $x \leq y$) only if there exist x positive integer $1 \leq n_1 < n_2 < n_3 < \cdots < n_x \leq y$, and $S_i = T_{ni}\ \forall\ i$. In addition to this, it is said that $\langle CS_1, CS_2, CS_3, \dots\dots, CS_x \rangle$ is a contiguous sequence in t if there exists an integer $0 \leq m \leq y - x$, and $CS_i = T_{m+i}\ \forall\ i = 1 \dots\dots x$. The pair of adjacent items, S_i and S_{i+1}, should come

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

successively in T for the Contiguous Sequential Pattern (CSP).

For T, sequential pattern S's support in T is expressed as the fraction of T that contains S. Likewise, the confidence of the rule $X \rightarrow Y$ can be determined using the equation 2.

$$C(X \rightarrow Y) = Sup(X \circ Y)/Sup(X) \quad (2)$$

In equation 2, ° represents the concatenation operator.

In [1], CPSs were used to find the frequent navigational paths. It also said that the items of SP need not require adjacent when the underlying ordering was preserved. This represents the more general navigational patterns inside a site. The discovery or analysis of navigation patterns becomes easier for the models when the transaction of web were represented as sequence of page view. One such model was Markov model, which represents the user's navigational behavior and their activities in the web site. In this model the page views were denoted as states along with the probability to between two states. This expresses the likelihood of the user from one page to another page. From the state representation, number of useful users and site metrics can be manipulated. Markov model was used as the underlying model to determine the link prediction and web pre-fetching. The web pre-fetching was used to decrease the system latencies [4]. The objective of this model is to determine the user's future action based on their prior browsing behavior. In [2] this model was used to find the high probability user navigation in web site. Following the previous work, [15] uses a mixture of Markov model to cluster the navigational sequences and perform exploratory analysis of user's navigational behavior. Figure 5, expresses how the web transaction can be modeled as a Markov model. In general Markov model has complexity problem, which was addressed in [22] through selective Markov Model. Another way for representing the Markov model is through trie structure [1]. An example of the aforementioned structure was aggregate tree. Figure 6, represents the aggregate tree for the transaction that was given in Figure 6.

3.4 Classification based Pattern discovery

Decision rule is one of the most widely used techniques for web usage mining [5]. The set of rules are the result of it and express the user's interest. These extracted rules were used to personalize the site information for a specific user [3]. Conditional Random Fields (CRF) were used in [7] as an alternative for discovering the behavior pattern. CRFs are used for the classification of sequential data, which is a probabilistic framework. Probable subsequent web pages for web users were used in [31] to discover the user behavior pattern. They were also compared with the other well known probabilistic frameworks. An Error Correcting Output Coding (ECOC) of the CRF was proposed in [33] to predict the subsequent web pages from a large-size web site.

3.5 Clustering based Pattern discovery

Clustering based pattern discovery is the most widely used and implemented techniques in web usage mining. This sub-section gives some of the work that was related to mining usage patterns based on the clustering techniques.

Table 4: Advantages Of Clustering Algorithms Proposed In Discussed Paper

Paper	Application	Advantage
[23, 21]	Session Clustering	Improves the first order Markov model
[35]	Session Clustering	Enhanced the existing clustering algorithm
[40]	Item clustering	Enhanced the K-means algorithm
[6]	Path clustering	Avoids curse of dimensionality problem
[39]	Web data clustering	Improves FCM algorithm
[8]	Personalization	Automatic web personalization

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Transaction	Frequency
A, B, E	10
B, D, B, C	4
B, C, E	10
A, B, E, F	6
A, D, B	12
B, D, B, E	8

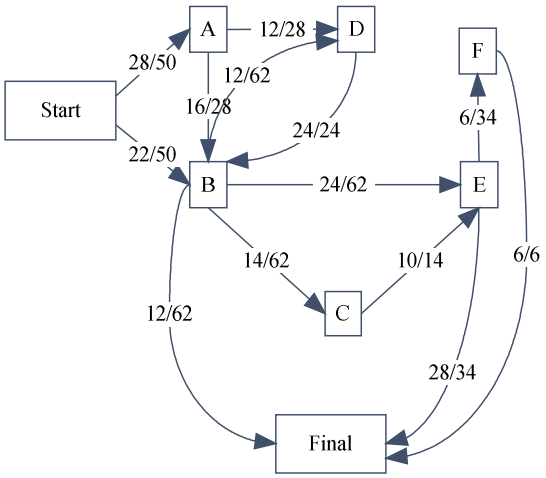

Figure 5: An Example For Markov Model

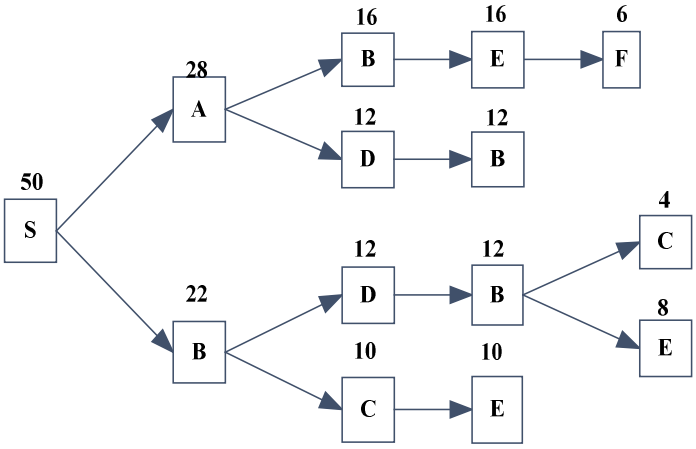

Figure 6: An Example For Aggregate Tree

As seen in sequential pattern mining, the Markov model has some problem in discovering the usage patterns. This leads to discover the patterns with less accuracy. To increase the accuracy of Markov model, a dynamic clustering-based method was proposed in [21, 23]. This method has extended the first-order Markov model to combine probabilities of second-order. Here, they used the state cloning method to fill the separated in-links. It also used a clustering algorithm to find the way to distribute the state's in-link among the state and its clones. A new algorithm was proposed in [6], which was based on the function of the longest common subsequence of their click streams. This algorithm clusters the path, in which the individual paths were combined by considering the time information to form a cluster. This is an efficient, scalable and fast graph-based algorithm was used.

Session clustering is used to enhance the web usage mining process. In [35] a framework was proposed to cluster the sessions. This method converts the web log details into numerical data. From these a session vector is found, so that swarm optimization can be applied for clustering the web log files. This technique enhances the usual session clustering technique. In [40] an algorithm was proposed to enhance the k-means algorithm to cluster the items in the web log transactions. Even though this algorithm enhances the k-means algorithm it suffers from complexity. A modified and improved FCM based algorithm was proposed in [39] to deal with the complex information present in the log files and to cluster those data. Fuzzy based artificial immune system was proposed in [2], which overcomes the existing techniques problems such as uncertainty and fuzziness inherent. A common user profiles were generated through the technique proposed in [8]. The generated profiles can be implemented to personalize web site and to make recommendations. The algorithm depends on the Dempster-Shafer's theory to cluster the task, which has the ability to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

capture the uncertainty among Web user's navigation behavior. Table 4 summarizes the techniques used in different paper for clustering and its advantages.

4. PATTERN ANALYSIS

Pattern analysis is the final step in the of usage mining. The main objective of this process is to eliminate the irrelevant rules and patterns that are extracted at discovery process. Following are the application of pattern analysis.

- Web caching
- User behavior learning and web personalization
- Design of site or page is improved

Pattern analysis is mainly applied at e-commerce website, which helps the designers to develop a site that satisfies a group of visitors. Service organization, newsletter websites are some other sites that benefits from the pattern analysis [43]. Various tools are available to analyze the extracted pattern but there is no single tool to analyze all type of knowledge from the web server log. Apart from accuracy, quality of extracted pattern plays an important role in efficiency of mining process.

On removing the non contributory information and non human log entries from the log, the quality of web log is improved. Session identification and path identification can be applied to enhance the quality of the web.

5. CONCLUSION

Web usage mining is a booming technology that helps in various ways like personalizing the website, visualization, etc. This paper presents a review of various techniques used in preprocessing, pattern discovery and pattern analysis. This survey is presented with an objective to serve the scholars working in web usage mining, predominantly to those systems that are accessible over the web. Each section concentrates on the various techniques used in the different phases of the usage mining along with a summary of the discussed technique. Having idea gained from this survey, we would like focus on generating a new technique for personalizing the website in future, which is more accurate as well as has elevated quality.

REFERENCES

[1] M. Spiliopoulou and L. Faulstich, "WUM: a tool for web utilization analysis," *The World Wide Web and Databases*, pp. 184-203, 1999.

[2] J. Borges and M. Levene, "Data mining of user navigation patterns," *Web usage analysis and user profiling*, pp. 92-112, 2000.

[3] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic personalization based on Web usage mining," *Communications of the ACM*, vol. 43, pp. 142-151, 2000.

[4] R. R. Sarukkai, "Link prediction and path analysis using Markov chains," *Computer Networks*, vol. 33, pp. 377-386, 2000.

[5] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, pp. 12-23, 2000.

[6] A. Banerjee and J. Ghosh, "Clickstream clustering using weighted longest common subsequences," in *Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining*, 2001, pp. 33-40.

[7] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[8] Y. Xie and V. V. Phoha, "Web user clustering from access log using belief function," in *Proceedings of the 1st international conference on Knowledge capture*, 2001, pp. 202-208.

[9] N. Cercone and A. An, "Comparison of interestingness functions for learning web usage patterns," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 617-620.

[10] W. Lin, S. A. Alvarez, and C. Ruiz, "Efficient adaptive-support association rule mining for recommender systems," *Data Mining and Knowledge Discovery*, vol. 6, pp. 83-105, 2002.

[11] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos, "Exploiting web log mining for web cache enhancement," *WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points*, pp. 235-241, 2002.

[12] O. Nasaroui, F. Gonzalez, and D. Dasgupta, "The fuzzy artificial immune system: Motivations, basic concepts, and application to clustering and web profiling," in *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*, 2002, pp. 711-716.

[13] Z. Pabarskaite, "Implementing advanced cleaning and end-user interpretability technologies in web log mining," in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Information Technology Interfaces, 2002. ITI 2002. Proceedings of the 24th International Conference on, 2002, pp. 109-113.

[14] G. Stumme, A. Hotho, and B. Berendt, "Usage Mining for and on the Semantic Web," in *National Science Foundation Workshop on Next Generation Data Mining*, 2002.

[15] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Model-based clustering and visualization of navigation patterns on a web site," *Data Mining and Knowledge Discovery*, vol. 7, pp. 399-424, 2003.

[16] M. Eirinaki, M. Vazirgiannis, and I. Varlamis, "SEWeP: using site semantics and a taxonomy to enhance the Web personalization process," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 99-108.

[17] P. Fraternali, M. Matera, and A. Maurino, "Conceptual-level log analysis for the evaluation of web application quality," in *Web Congress, 2003. Proceedings. First Latin American*, 2003, pp. 46-57.

[18] K. P. Joshi, A. Joshi, and Y. Yesha, "On using a warehouse to analyze web logs," *Distributed and Parallel Databases*, vol. 13, pp. 161-180, 2003.

[19] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez, "Conceptual user tracking," *Advances in Web Intelligence*, pp. 955-955, 2003.

[20] F. Yuan, L.-J. Wang, and G. Yu, "Study on data preprocessing algorithm in web log mining," in *Machine Learning and Cybernetics, 2003 International Conference on*, 2003, pp. 28-32.

[21] J. Borges and M. Levene, "A dynamic clustering-based Markov model for web usage mining," *arXiv preprint cs/0406032*, 2004.

[22] M. Deshpande and G. Karypis, "Selective Markov models for predicting Web page accesses," *ACM Transactions on Internet Technology (TOIT)*, vol. 4, pp. 163-184, 2004.

[23] J. Borges and M. Levene, "A clustering-based approach for modelling user navigation with increased accuracy," in *Proc. of the 2nd Intl. Workshop on Knowl. Discovery from Data Streams (IWKDDS) & PKDD*, 2005.

[24] R. Meo, P. Lanzi, M. Matera, and R. Esposito, "Integrating web conceptual modeling and web usage mining," *Advances in Web Mining and Web Usage Analysis*, pp. 135-148, 2006.

[25] T. Murata and K. Saito, "Extracting Users' Interests from Web Log Data," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006, pp. 343-346.

[26] C. Nichele and K. Becker, "Clustering web sessions by levels of page similarity," *Advances in Knowledge Discovery and Data Mining*, pp. 346-350, 2006.

[27] G. Castellano, A. Fanelli, and M. Torsello, "LODAP: a log data preprocessor for mining web browsing patterns," in *Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*, 2007, pp. 12-17.

[28] G. Castellano, F. Mesto, M. Minunno, and M. Torsello, "Web user profiling using fuzzy clustering," *Applications of Fuzzy Sets Theory*, pp. 94-101, 2007.

[29] G. Stermsek, M. Strembeck, and G. Neumann, "A user profile derivation approach based on log-file analysis," in *Intl. Conf. on Information and Knowledge Engineering (IKE)*, 2007.

[30] S. Alam, G. Dobbie, and P. Riddle, "Particle swarm optimization based clustering of web usage data," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, 2008, pp. 451-454.

[31] Y. Z. Guo, K. Ramamohanarao, and L. A. Park, "Web page prediction based on conditional random fields," in *Proceeding of the 2008 conference on ECAI*, 2008, pp. 251-255.

[32] L. Yun, W. Xun, and G. Huamao, "A Hybrid Information Filtering Algorithm Based on Distributed Web log Mining," in *Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on*, 2008, pp. 1086-1091.

[33] Y. Guo, K. Ramamohanarao, and L. Park, "Grouped ecoc conditional random fields for prediction of web user behavior," *Advances in Knowledge Discovery and Data Mining*, pp. 757-763, 2009.

[34] K. Suneetha and R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File," *IJCSNS International Journal of Computer Science and Network Security*, vol. 9, pp. 327-332, 2009.

[35] T. Hussain, S. Asghar, and S. Fong, "A hierarchical cluster based preprocessing methodology for Web Usage Mining," in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Advanced Information Management and Service (IMS), 2010 6th International Conference on, 2010, pp. 472-477.

[36] H. Peng, "Discovery of interesting association rules based on web usage mining," in *Multimedia Communications (Mediacom), 2010 International Conference on*, 2010, pp. 272-275.

[37] M. Dimitrijević, Z. Bošnjak, and S. Subotica, "Web Usage Association Rule Mining System," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 6, pp. 137-150, 2011.

[38] K. Suneetha and R. Krishnamoorthi, "Classification of web log data to identify interested users using decision trees," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 2, 2011.

[39] K. Suresh, R. MadanaMohana, A. RamaMohan Reddy, and A. Subramanyam, "Improved fcm algorithm for clustering on web usage mining," in *Computer and Management (CAMAN), 2011 International Conference on*, 2011, pp. 1-4.

[40] V.Chitraa and D. A. S. Thanamani, "An Enhanced Clustering Technique for Web Usage Mining," *International Journal of Engineering Research & Technology* vol. 1, pp. 1-5, 2012.

[41] X. Yu, M. Li, I. Paik, and K. Ryu, "Prediction of Web User Behavior by Discovering Temporal Relational Rules from Web Log Data," in *Database and Expert Systems Applications*, 2012, pp. 31-38.

[42] R. Khanchana and D. M. Punithavalli, "A Web Usage Mining Approach Based On New Technique In Web Path Recommendation Systems," *International Journal of Engineering Research & Technology*, vol. 2, pp. 1-6, 2013.

[43] G. Velayathan and S. Yamada, "Behavior-based web page evaluation," in *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*, 2006, pp. 409-412.