A Survey on Discovery of Knowledge about Web users Web Access Behaviour

E. Manohar[1]*, D. Shalini Punithavathani[2]
[1] Research Scholar, Anna University Chennai, India.
[2] Principal, Government College of Engineering, Tirunelveli, India.
***Corresponding author: E-Mail: manohar2k@ymail.com**

ABSTRACT

The advancement of the web technology becomes very rapid. Mining the web gives us a lot of productive information. The web services consist of the development of latest technological development. The count of web users has also increased rapidly. Web users' information is very much essential for the effective utilization of web services. Most of the companies invest a large amount of money to discover the knowledge about the web users' information. Various machine learning and artificial intelligence techniques are available nowadays to discover the knowledge about the web users. Moreover, so many resources are available to discover the knowledge about the web users. They are web log, web review, web rating, web ranking, web survey, browser agent, web user authentication and tracking cookies. Web consists of a huge amount of structured and unstructured information and hence the web users find it difficult to get the required information at the right time. The purpose of knowledge discovery is to organize the website according to the web users' requirement which may also reduce the network traffic. This paper presents the survey and analysis about the web users' usage mining techniques.

KEY WORDS: web usage mining, pattern discovery, association, clustering, classification, sequential pattern mining.

1. INTRODUCTION

Internet consists of a vast collection of information and so a huge number of new websites are created everyday but most of them do not get sufficient response from the web users and so many websites are deleted every day. A large number of the websites are unstructured. The advancement in technology is the one of the major reasons for the increase in marketing, based on the web. Personalization of the web is significant for presenting the website effectively to the web users. The web users' number is doubled periodically (Duhan, 2009). As the quantity of web information increases, web personalization is necessary to minimize the network traffic as internet is common among people nowadays. The personalization of the web may be attained in discovering knowledge about the web users. This knowledge can be obtained by web survey, browser agent, tracking cookies, web log, web rating, web review and web ranking. By using machine learning and artificial intelligence techniques, the knowledge about the web users can be mined. The web usage information mining can be classified as usage, structure and content mining (Facca and Lanzi, 2005). Web log is an important source to discover knowledge about the web users. Web log consists of the access behaviour of the users in the website (Langhnoja, 2013). Discovery of knowledge using web log is of three phases. They are pre-processing, knowledge discovery and knowledge analysis. The various machine learning methods which are used for discovering the knowledge about the web users are statistical analysis, association mining, cluster analysis and classification technique. Web log can be classified as web access logs, web agent logs, web error logs and web referrer logs (Baoyao, 2006). Web server log is classified as common log and extended log format. Common log consists of client IP, user name, bytes transferred, server name, request and status. Extended Log Format consists of bytes that are sent and bytes received, server, request, requested service name, time taken for transaction to complete, version of transfer protocol used, user agent, cookie ID, and referrer.

Web Log Pre-processing: Web log pre-processing increases the accuracy in discovering the knowledge. Pre-processing the web log takes a lot of time in discovering knowledge and it involves data collection or data fusion, integration of data, data cleaning, data reduction, web user session identification and web user identification. Data pre-processing is time consuming but leads to accurate knowledge about the web users. Data pre-processing is also called preparation of data which is a part of knowledge discovery (Jozef, 2012). The pre-processing process improves the quality of the information and accuracy in mining (Pooja Kherwa, 2015).

Extracting the Web Log: Extracting the file of the log is the initial step in pre-processing. The web server uses comma (,) and space (" ") character as separator for each web log field.

Web Log Cleaning: The web log cleaning phase cleans the irrelevant and redundant data; so that the correct information can be stored in the knowledge discovery phase of web log consists of huge irrelevant records, the data cleaning phase deletes irrelevant records.

Data cleaning reduces the log size. Analyzing a quantity of data is a complex and time consuming activity. Web log cleaning phase removes the web robots, images, audio file, video, java script, CSS and status code information.

Web User Identification: The identification of the web user is one of the challenging tasks which is useful for the personalization of web where unique web user is identified. Mostly the individual web user is identified by using IP address. Identifying the user is difficult because of local caches, proxy servers and firewalls. (Ciesielski and Lalani,

2003). Many other techniques to identify the web users are also available. The proactive method uses cookies whereas reactive method uses web log file. By navigating the web log file, we can identify the timing sequence of the users (Sharma, 2008). In reactive strategy, clustering technique is employed. The individual web users may be identified by means of constructing the transaction of web user's web access. Cookies and authentication mechanism may be utilized for identification of the users. The web user may be identified by calculating the time spent in every page and IP address is the key to find out the users of the web. Clustering method is also adapted for the identification of the users.

Web Users Session Identification: A website may contain many web log records every day, depending on its popularity (Mohammad). A web log consists of different information such as IP address, the user access, date and time, and the document or image requested. To enumerate the knowledge about the user, the grouping of day according to individual user's individual session is very much essential. Session is nothing but the navigation of several web pages at the particular access time period. A web user has one session or many sessions during a particular period. Session is termed as a particular period of time between two visits to the same website by the user. The web users' session can be identified by using two methods; they are proactive method such as cookies and reactive method such as navigation-driven method. The navigation-driven method includes two methods and they are maximum forward reference transaction identification and reference length method. Most of the techniques use 30 minutes as a default time out and establish time out of 25.5 minutes for some observed data. The website log has been analyzed and statistics is obtained from the web log file; so that an appropriate time out can be obtained for that particular website.

Web Transaction Identification: Joining or dividing several sessions into a meaningful cluster is called transaction identification. The web user page visit can be categorized as auxiliary page or content page. If the web user uses the page for navigation, then it is called auxiliary page. If the web user uses the page of the web to retrieve the content in that web page, then it is called content page. Many approaches of web transaction such as reference length and identification of transaction identification by maximal forward reference are used commonly.

Web Path Completion: Path completion of the pages of web by web users is a challenging task. Sometimes in the web log, particulars of the web users' access may not be present because of the clicking of back and forward button in the browser by the web users and also the use of proxy server. In the above cases, the web links information will not be there in the web log. In order to discover the missing path information, the web path completion techniques are used trace the missing path. In path completion, not only missing path is identified but also the time taken on the missing pages is also determined. These lost pages are considered as the auxiliary pages and the average reference length of these pages is to be estimated.

There are various techniques available to navigate the web users' access behaviour of which log file is one of the effective techniques to know about the users' access behavior. Web users' access information is stored in web logs. Log file contains the web user page request details, web users' IP address, web users' accessing date and time details, HTTP code and bytes served. User agent details are all stored in web log file. For the effective web log pre-processing, the following steps should be carried out effectively; they are data collection, integration, cleaning, data reduction, session identification and user identification, the page view identification, completion of path and episode identification. To trace the web users' path, the time related and referrer related heuristics for path completion are employed.

The Table.1, shows the analysis of many types of pre-processing algorithms which are used in various steps in pre-processing.

Table.1. Web Log Pre-processing Survey

Algorithm / Title	Author	Description
Data mining of genetic programming run logs	Ciesielski and Lalani, 2003	Idea is to get as much information as possible about the user-IP.
Data Preprocessing: A Milestone of Web Usage Mining	Sharma, 2008	First three and last two pages and list of directories.
Applying Packets Meta data for Web Usage Mining	Mohammad Ala	Classification of Log Files into a number of files; each one represents a class, using Decision Tree Classifier.
Pre-processing of Web Logs for Mining World Wide Web Browsing Patterns	Ismail Toroslu,	New approach to find frequent item sets employing Rough set Theory.
Log Data Preparation for Mining Web Usage Patterns	Castellano, 2007	LODAP-Four modules are involved namely-Data Cleaning, Data structuring, Data Filtering and Data Summarization.
Discovery of Web Robot Sessions based on their Navigational Patterns	Tan and Kumar (2002)	In Data Cleaning, removal of outliers or irrelevant data eliminating web robots generated log entries.

Algorithm / Title	Author	Description
Learning to remove Internet advertisement	Kushmerick, 1999	Proposed a feature based method which identifies internet advertisement and removes them.
An Overview of Data Pre-processing in Data and Web Usage Mining	Suresh and Padmajavalli, 2006	User accessing behavior is to be constructed as transaction.
Data preparation for mining World Wide Web browsing patterns	Robert Cooley, 1997	Users are identified using cookies or authentication mechanism
User Behaviour Analysis Based on Time Spent on Web Pages	Istvan, 2009	Using Page Viewing time.
Analysis of Web User Identification Methods	Renata Ivancsy and Sandor Juhasz, 2007	User Identification using their IP Address.
Pattern oriented hierarchical clustering	Morzy, Wojciechowski, 2000	Users are distinguished based on their navigational pattern using clustering methods.
The Impact of Site Structure and User Environment on Session reconstruction in Web Usage Analysis	Spiliopoulou, 2003	Using Proactive and Reactive Strategies for Differentiating users.
Web User Session Reconstruction Using Integer Programming	Robert Dell, 2008	Using Integer Programming construction of all sessions simultaneously.
A Tool for Web Usage Mining	Jose Domench and Javier LorenZo, 2007	In this, Referrer based method and time oriented heuristics methods are combined.
A Web Usage Lattice Based Mining Approach for Intelligent Web Personalization	Baoyao Zhou, 2006	Time stamp based method. The default time is 30 minutes.
Data Preparation For Mining World Wide Web Browsing Patterns	Cooley, 1999	Time oriented Heuristics 30 minutes.
Characterizing browsing behaviours on the World Wide Web	Catlegde and Pitkow , 1995	Time oriented Heuristics 25.5minutes to 24 Hrs.
A Novel Technique for Sessions Identification in Web Usage Mining Pre-processing	Chitraa, Antony Selvadoss Davamani, 2010	This method based on navigation uses web topology in graph format.
Research on Path Completion Technique in Web Usage Mining	Yan Li, 2008	Referrer-based method using proxy servers and local caching.

Knowledge discovery: Knowledge discovery about the web users depends on different algorithms and methods developed from various fields. The discovery of knowledge includes artificial intelligence, statistics, data mining, machine learning and pattern recognition. Several algorithms such as statistical analysis, association mining, clustering, classification and sequential mining are used during different stages depending on their diverse requirements. The knowledge discovery based on machine learning is the most famous and successful method in discovering knowledge about the web users.

Association Rule Mining: Association rule refers to a group of pages that are associated to one another with minimum support value. If the association has maximum value, then the pages of web may be linked to the existing web page. The association mining algorithm is an application of usage mining where the mining is focused on the next interesting web page of the web user. The association mining algorithms are used in knowledge discovery about the web users; some of the popular association rule algorithms are roughest theory (2005) and Markov chain model. The various knowledge association rules are listed in the Table.2.

Table.2. Survey on Association Rule Based Web Usage Mining

Algorithm	Author	Description
AIS	Agarwal,1994	In this algorithm, only one item consequent association rules are generated.
SETM Algorithm	Houtsma and Swami, 1993	The new users' sets generation is the same as in AIS algorithm. But it is saved in sequential structure. Generation process is separated from counting. Support count is differentiated by the sequential structure.
Apriori	Wang, 2009	In Apriori algorithm, the Candidate item sets are generated using the previous pass without considering the transactions in the database.

Algorithm	Author	Description
Apriori TID	Zc Li, 2005	'C' is generated of which each member has the Transaction ID of each transaction and the large item sets are present in this transaction. This set is used to count the support of each candidate item set.
FP growth Algorithm	Jiawei Han, 2004	Divide and Conquer Technique is used for the association of web log data in FP growth algorithm.
RARM	Woon, 2001	A versatile tree structure known as the Support-Ordered Tree Item set structure to hold pre-processed transactional data for the association of web log data.
Improved Apriori All	Jianlong Gu, 2011	The characteristics of user ID at the time of producing candidate set and scanning database is to determine whether to put it into large set at the time of producing next set. Apriori algorithm helps to minimize the candidate set at the time of its production.
Custom-built Apriori algorithm	Sandeep Sing, 2010	We can customize the algorithm in such a way that pruning operation is performed only on the candidate item sets whose size > 2. To generate frequent itemsets of size k only the transactions whose size >= k are considered.
Association rule hiding algorithm	Natarajan, 2012	The algorithm will keep privacy in data mining. It will keep confidentiality and performance.
Maximal forward references	Chen, 2004	A maximum forward reference is defined as the longest consecutive of forward reference before the first backward reference is made.
Markov Chain	Dempster (1908)	It does not depend on the events but it performs the next state depending on the current state.

Clustering: The popular knowledge discovery technique is the clustering method. Clustering is grouping similar web access for the knowledge discovery. The different types of cluster which are used in here are page cluster, and usage content. This knowledge discovery is useful to personalize the web content of the users. The discovery of knowledge is very useful in web assistance and search engine personalization. Clustering consists of three methods; partitioning, hierarchical and model based methods. Table.3, shows the various clustering based mining.

Table.3. Survey on Cluster Based Knowledge Discovery

1) Partitioning methods: The data are divided into k groups. Various algorithms can be employed for different purposes.		
a) Clustering User Session - Algorithms		
Algorithm	**Author**	**Description**
Expectation-Maximization (EM)	Dempster (1908)	Expectation–maximization algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models.
Fuzzy clustering	Wolfram (1983)	Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster.
Graph partitioning	Bruce Hendrickson, 1997	In graph partition, the graph is partitioned into sub graph, based on different properties.
Self-Organizing Maps	Kohonen and Teuvo, 2001	The self-organising map produces discrete samples which are called map.
Ant-based	Marco and Thomas, 2004	This algorithm identifies the optimal path through graphs.
k-means with genetic algorithm	Krishna and Narasimha Murty, 1999	The optimal partition of a given data into a number of clusters.
b) Clustering index page synthesis – Algorithm		
Algorithm	**Author**	**Description**
Page Gather	Cutting, 1992	The collection of document is clustered in to various groups. The user selects any number of cluster group based on summaries. The information is clustered again until successive iteration.
2) Hierarchical methods: The web data are decomposed to create hierarchical structure of the cluster.		
Algorithm	**Author**	**Description**
BIRCH Algorithm	Tian Zhang, 1997	It performs hierarchal clustering over a large set of data.

3) Model-based methods: Model based methods identify the suitable combination between the given dataset. There are many algorithms used in mathematical model for clustering the web users' session.		
Algorithm	**Author**	**Description**
Autoclass	Johan Stutz and Peter Cheeseman, 1996	Autoclass is a clustering algorithm of mixed data type which determines the optimal classes based on prior distribution.
Self-Organizing	Vesanto (2000)	The low-dimensional regular grid is utilized to express the properties of the data. If the data is huge, then the similar data should be grouped.
COBWEB	Douglas Fisher, 1989	COBWEB takes an object at a time to decide whether it should be accommodated in the existing cluster or added to the hierarchy as new cluster.

Classification: Classification is a knowledge discovery technique which classifies the information into two different classes. To identify the different class, the users' page can be personalized based on the specific class. Classification may be based on different categories. There are different classification algorithms for mining. Table 4 analyses the discovery of knowledge based on various classification techniques.

Table.4. Survey on Classification Based Knowledge Discovery

1) Classification according to web users' interest - Algorithms		
Algorithm	**Author**	**Description**
HCV	Xindong, 1995	HCV discovers a set of rules representing the users' interests
CDL4	Shen, 1996	CDL4 Semi incremental learning method.
2) Classifying the interesting page - Algorithms		
Algorithm	**Author**	**Description**
RIPPER	William Cohen, 1995	Repeated Incremental Pruning to Produce Error Reduction. It is effective in large datasets.
C4.5	Ross Quinlan, 1993	Based on information entropy the C4.5 builds decision trees from a set of training data.
Naive Bayesian	Charles Elkan, 1997	In Bayes theorem, the Independence assumption between features is implemented for classification.
3) Classification of session - Algorithms		
Algorithm	**Author**	**Description**
Rough Set Theory	Zdzislaw Pawlak, 2002	The algorithm identifies the lower and the upper approximation of the set.

Sequential Pattern Mining: Sequential mining identifies the occurrence of sequential events which is to determine whether there is any relevant sequence in that occurrence. Discovering knowledge can be utilized to predict the next page the web users are going to access and also it guides the designer to personalize the advertisement to the web users. The discovery of knowledge using sequential pattern mining can be determined by using two methods; they are deterministic method and stochastic method. Deterministic method holds the navigational movement of the users. Stochastic method uses sequential web page access for predicting the next visit by the web users. Various sequential mining algorithms are shown in Table.5.

Table.5. Survey on Sequential Pattern Mining

Algorithm	**Author**	**Description**
GSP Algorithm	Ramakrishnan Srikant (1996)	GSP Algorithm is used for sequence mining. The algorithm uses Apriori algorithm which discovers the sequential pattern in level wise.
FreeSpan	Jiawei Han, 2000	The algorithm integrates sequence pattern along with frequent sequence to predict the subsequent frame
Sequential Pattern Discovery using Equivalence classes	Philippe Fournier-Viger (2014)	ERMiner search using equivalence classes of rules.
Data Mining of User Navigation Patterns	Jose Borges and Mark Leven, 2002	The algorithm extracts the navigation pattern from web user sessions.
Markov models	Markov (1908)	This model predicts the future state based on current state and not by the occurrence of events.

Knowledge Analysis: After knowledge discovery, the knowledge analysis is the final phase. Knowledge analysis phase identifies the relevant information from the discovery phase. The knowledge analysis phase drops all the knowledge which is not relevant to the particular application. The pattern of the users is identified by navigating the web users' access page. The visual presentation of the knowledge aids easy interpretation. The vital representation of knowledge analysis is the graphical presentation of the knowledge.

2. CONCLUSION

In this survey, the various algorithms and different methods for discovering knowledge about the web users are analyzed. Every algorithm has its own advantages and disadvantages. Various machine learning and artificial intelligence techniques are used to discover knowledge. Machine learning algorithm which includes clustering, association, classification and sequential pattern mining are mainly used in majority of the knowledge discovery methods. In future, various researches are to be done to utilize diverse sources in knowledge discovery and also implement many tools to make the discovered knowledge more accurate.

REFERENCES

Agarwal R and Srikand, Fast algorithm for Mining Association Rule, IBM, 1994.

Amitabha Das, Wee-Keong Ng and Woon, Rapid Association Rule Mining, Proceedings of the Tenth International Conference on Information and Knowledge Management, 2001, 474-481.

Baoyao Z, Siu C and Alvis C, Fong M, An Effective Approach for Periodic Web Personalization, Proceedings of the IEEE/ACM International Conference on Web Intelligence, IEEE, 2006.

Bruce Hendrickson, Robert Leland and Rafael Van Driessche, Skewed Graph Partitioning, Sandia National Labs, 1997, 1-8.

Castellano G, Fanelli A, Torsello M, Log Data Preparation For Mining Web Usage Pattern, IADIS International Conference Applied Computing, 2007, 371-378.

Catlegde L and Pitkow J, Characterizing browsing behaviors in the world wide Web, Computer Networks and ISDN systems, 1995.

Charles Elkan, Boosting and Naive Bayesian Learning, Technical Report No. CS97-557, 1997.

Chen J, Sun L, Zaiane O and Goebel R, Visualizing and Discovering Web Navigational Patterns, Seventh International Workshop on the Web and Databases, Paris, 2004, 17-18.

Chitraa V, Davamani A, A Survey on Pre-processing Methods for Web Usage Data, International Journal of Computer Science and Information Security, 7 (3), 2010.

Ciesielski V and Lalani A, Data mining of web access logs from an academic web site, Proceedings of the Third International Conference on Hybrid Intelligent Systems (HIS'03), 2003.

Cooley R, Mobasher B, Srivastava J, Knowledge and Information System, Springer-Verlag, 1999.

Cooley R, Mobasher B, Srivastava J, Web Mining, Information and Pattern Discovery on the world wide web, International Conference on Tools With Artificial Intelligence, 1997, 558-567.

Cutting D.R, Karger D.R, Pedersen J.O and Tukey J.W, Scatter/Gather, A Cluster-Based Approach To Browsing Large Document Collections, Proc. 15th Annual International ACM SIGIR Conference on R&D in IR, 1992.

Dempster A.P, Laird N.M. Rubin D.B and Markov A.A, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, 39 (1), 1977, 1-38.

Douglas Fisher, Noise-tolerant conceptual clustering, Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, 1, 1989, 825–830.

Duhan N, Sharma A.K and Bhatia K.K, Page Ranking Algorithms: A Survey, in Proc. IEEE International Advance Computing Conference, 2009, 1530-1537.

Facca F.M, Lanzi P.L, Mining interesting knowledge from weblogs, a survey, in Proc. Data and Knowledge Engineering, 53 (3), 2005, 225-241.

Ismail H and Toroslu M, Graph Theoretic Approach for Session Reconstruction Problem, Data & Knowledge Engineering, 73, 2012, 58–72.

Istvan Nagy K and Csaba G, User Behaviour Analysis Based On Time Spent On Web Pages, Web Mining Application in E-Commerce and E-Services, Studies in Computational Intelligence, 172, 2009, 117-136.

Ivancsy R and Juhasz S, Analysis of Web User Identification Methods, World Academy Of Science, Engineering and Technology, 1, 2007, 10-29.

Jianlong Gu, Baojin Wang, Fengyu Zhang, Weiming Wang and Ming Gao, An Improved Apriori Algorithm, Proceedings of the International Conference on Applied Informatics and Communication, 2011, 127 -133.

Jiawei Han, Jian Pei, Behzad, Quiming Cheng, Umeshwar Dayal and Mei-Chung Hsu, Free Span: frequent pattern-projected sequential pattern mining, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 2000, 355-359.

Jiawei Han, Jian Pei, Yiwen Yin and Runying Mao, Mining Frequent Patterns without Candidate Generation, A Frequent-Pattern Tree Approach, Data Mining and Knowledge Discovery, 8, 2004, 53–87.

John Stutz and Peter Cheeseman, Auto class-A Bayesian Approach to Classification, Fundamental Theories of Physics, 70, 1996, 117-126.

Jose Borges and Mark Leven, Data Mining Of User Navigation Patterns, Springer Berlin Heidelberg, 2002, 92-112.

Jose M, Domenech and Javier L, A Tool for Web Usage Mining, 8th International Conference on Intelligent Data Engineering and Automated Learning, 2007.

Jozef K, Michal M, Martin D, User Session Identification Using Reference Length, 9th International Scientific Conference on Distance Learning in Applied Informatics, 2012, 175-184.

Kohonen and Teuvo, Self-Organizing Maps, Springer Series in Information Sciences, 30 (3), 2001.

Krishna K, Narasimha Murty M, Genetic K-means algorithm, IEEE Trans Syst Man Cybern B Cybern, 29 (3), 1999, 433-439.

Kushmeric N, Learning To Remove Internet Advertisements, Third Annual Conf. on Autonomous Agents, ACM Press, NY, 1999.

Langhnoja S.G, Barot M.P and Mehta D.B, Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern, International Journal of Data Mining Technology and Application, 2 (1), 2013, 141-150.

Marco Dorigo and Thomas Stutzle, Ant Colony Optimization, A Bradford Book the MIT Press Cambridge, 2004.

Markov A.A, Extension of limit theorems of the calculus of probabilities to sums of quantities associated into a chain, Fiz.-Mat. Otd. 7th Series, 22 (9), 1908, 363–397.

Maurice Houtsma and Sun Arun Swami, Set-Oriented Mining for Association Rules in Relational Databases, IEEE, 1993, 25-33.

Mohammad A, Adding new level in KDD to make the usage mining more efficient, First National Information Technology Symposium (NITS 2006) Bridging the Digital Divide, Challenge and Solutions, 2006, 5-7.

Morzy T, Wojcie M and Zakrazewicz M, Web Use Clustering, International Symposium on Computer and Information Sciences, 2000.

Natarajan R, Sugumar R, Mahendran M and Anbazhagan K, Design and Implement an Association Rule hiding Algorithm for Privacy Preserving Data Mining, International Journal of Advanced Research in Computer and Communication Engineering, 1 (7), 2012, 486 -492.

Philippe Fournier-Viger, Ted Gueniche, Souleymane Zida, Vincent S, Tseng Jozef K, Michal M, Martin D, ERMiner:, Sequential Rule Mining Using Equivalence Classes, International Symposium on Intelligent Data Analysis, 2014, 108-119.

Pooja Kherwa, Jyotsna Nigam, Data Preprocessing, A Milestone of Web Usage Mining, International Journal of Engineering Science and Innovative Technology (IJESIT), 4 (2), 2015.

Ramakrishnan Srikant and Rakesh Agrawal, Mining Sequential Patterns, Generalizations and Performance Improvements, in EDBT, Advances in Database Technology, 1996, 1-17.

Robert F, Dell P, Roman E, and Juan D, Web User Session Reconstruction Using Integer Programming, IEEE/ACM International Conference on Web Intelligence and Intelligent Agent, 2008.

Ross Quinlan J, C4.5, Programs for Machine Learning, Morgan Kaufmann Publishers Inc, 1993.

Sandeep Singh Rawat and Lakshmi Rajamani, Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm, International Journal Of Computer Science & Information Technology, 2 (4), 2010, 28-37.

Sharma A, NY Web Usage Mining, Data Pre-processing, Pattern Discovery and Pattern Analysis on the RIT Web Data, Rochester Institute of Technology, Rochester, 2008.

Shen W.M, an Efficient Algorithm for Incremental Learning for Decision Lists, Information Science Institute, 1996.

Spilopoulou M, Mobasher B and Berendt B and Nakagawa M, Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis, INFORMS Journal on Computing, 2003.

Suresh R and Padmajavalli R, An overview of Data Pre-processing in Data and Web Usage mining, IEEE, 2006.

Tan P and Kumar, Discovery of Web Robot Sessions Based on their Navigational Patterns, Kluwer Academic Publisher - Data Mining and Knowledge Discovery, 6 (1), 2002, 9-35.

Tian Zhang, Raghu Ramakrishnan and Miron Livny, BIRCH, A New Data Clustering Algorithm and Its Applications, Data Mining and Knowledge Discovery, 1 (2), 1997, 141-182.

Vesanto J and Alhoniemi E, Clustering of the self-organizing map, IEEE Transactions on Neural Networks, 11 (3), 2000, 586-600.

Wang P, Shi L, Bai J and Zhao Y, Mining association rules based on Apriori algorithm and application, IEEE, 2009.

William Cohen W, Fast Effective Rule Induction, Machine Learning, Proceedings of 12th International Conference, 1995.

Xindong Wu, Knowledge Acquisition from Databases, Ablex Publishing Corporation, 1995.

Yan L, Boqin F and Qinjiao M, Research on Path Completion Technique in Web Usage Mining, International Symposium on Computer Science and Computational Technology, IEEE, 2008.

Youquan H, Decentralized Association Rule Mining On Web using Rough Set Theory, Journal of communication and computer, 2 (7), 2005.

ZC Li, A high efficient aprioritid algorithm for mining association, IEEE, 2005.

Zdzislaw Pawlak, Rough set theory and its applications, Journal of Telecommunication and Information Technology, 2002, 7-10.