Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

A Comparative Study of Recommendation System Using Web Usage Mining

Mr. D. C. Karthick Kumar[1], Mr. R. Lokesh Kumar[2], Dr. P. Sengottuvelan[3]
PG Scholar[1], Assistant Professor[2], Associate Professor[3]
Department of IT[1,2,3]
Bannari Amman Institute of Technology[1,2,3]

Abstract—Web Mining is one of the Developing field in research. Exact custom of the Web is to get the beneficial material in the sites. To reduce the work time of user the Web Usage Mining (WUM) technique is introduced. In this Technique use Web Page recommendation for the Web request from the user. For the recommendation system in Web Usage Mining (WUM) variousauthor has introduce different Algorithm and technique to improve the user interest in surfing the Web. Web log files are used todefine the user interest and there next recommend page to view.The data stored in the web log file consist of large amount oferoded, incomplete, and unnecessary information. So, the Web log files have to pre-process, customize, and to clean the data. In this paper we will survey different recommendation technique to identify the issues in web surfing and to improve web usagemining (WUM) pre-processing for pattern mining and analysis.

Keywords-Web Mining, Web Usage Mining, Pre-processing, Recommendation technique, Pattern Mining

I. INTRODUCTION

With the rapid growth of the Web, it becomes more and more difficult for Web users to find useful information. In particular, a Web user often wanders aimless on the Web without visiting pages of his/her interests, or spends a long time to find the expected information. Web page recommendation is thus proposed to address this problem. It aims to understand the users' behaviors, and guide users to visit pages of their interests at a specific time. An essential task of Web page recommendation is to understand users' navigation behaviors from their Web usage data, and devise a model to predict what pages the users are more likely to visit at the next step.

Web Usage Mining is the submission of data mining techniques to determine motivating usage shapes from Web data in order to know and well aid the needs of Web-based tenders. Usage data seizures the self or cause of web users sideways with their browsing performance at a web site. Web usage mining that one can be confidential extra dependent on the kind of tradition data measured:

Web Server Data: The user logs are collected by the web server. It contains various unwanted data like IP address, page reference and access time.

Application Server Data: Profitable bid aides have substantial landscapes to qualify e-commerce tenders to be built on top of them with slight power. A key article is the talent to roadway countless breeds of commerce dealings and records them in submission attendant woods.

Application Level Data: Novel kinds of actions can be well-defined in a claim, and sorting can be bowed on for them thus producing antiquities of these especially definite events. It must be famous, but, that many end claims require a blend of one or more of the procedures useful in the groupings overhead. Studies correlated to work are afraid with two expanses: constraint-based data mining systems applied in Web Usage Mining and settled software riggings.

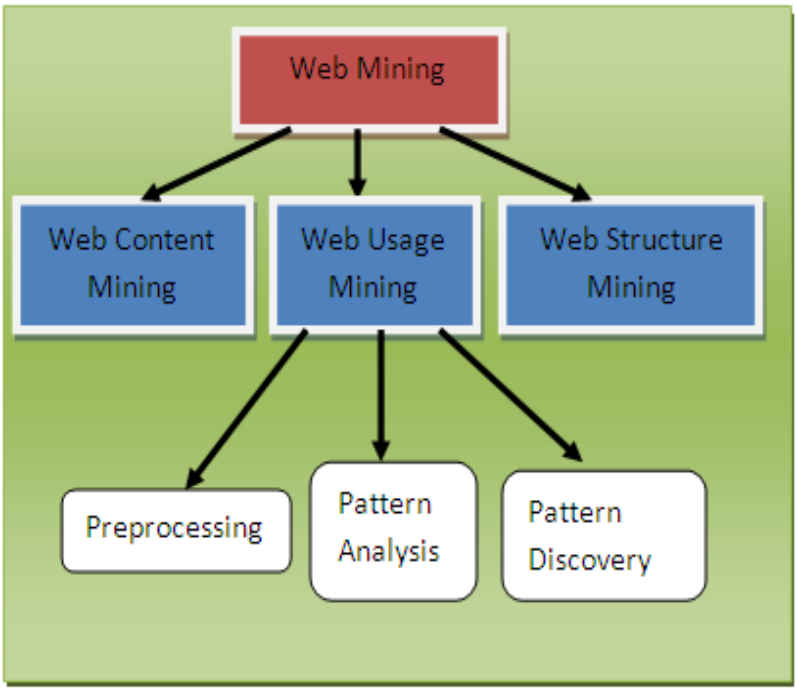

Figure 1. Classification of Web Usage Mining

Agent log file is used to record the information about user's browser, browsers version and operating system. Different versions of different users browsing history are very helpful for designer and web site changes are made accordingly.

Access log file is one of the major web log server, it will record each click, hits and access of the users for capturing information about user, can use number of attributes. Table 1.describes different attributes of access log file along with their explanation.

Table 1. Attributes of Log File and Description

Attributes	Description
Client IP	Client machine IP Address
Client name	Client Name if required by server otherwise hyphen(-)
Date	Date is recorded when User Made access and transfer.
Time	Time of transformation is recorded

Server site name	Internet service name as appeared on client machine
Server Computer name	Server name
Server IP	Server IP provided by internet service provider
Server port	Server port configured for data transformation
Client server URL stream	Targeted default web page of web site
Client server URL Query	Client query which starts after "?"
Server client status	Status code returned by server link
Server client Bytes	Number of bytes sent by server to client
Client server Bytes	Number of bytes received by client
Client server methods	Client method or model of request can be Get, POST or HEAD
Time taken	How much spent by client to perform an action
Client server version	Protocol version like HTTP
Client server host	Host header name
User agent	Browser type that client used
Cookies	Contents of cookies
Referrer	Link from where client jump to this site
Server client win32 status	Windows status code

II. MOTIVATION

Preprocessing is the essential step for web usage mining. Preprocessing phase will occur after creating a web log file. A web log file is an input of preprocessing phase of web usage mining. Web log file is large in large, contains number of raw datas, images, audio/ video files. The main aim of preprocessing is to make an unstructures/semistructured web log file data into a structured form by

eliminating the unwanted datas from log. We came to a conclusion that preprocessing is very important in web usage mining. Preprocessing steps increases the quality of data and improves the effictiveness and efficiency of the other steps in web usage mining like pattern discovery and pattern analyais.

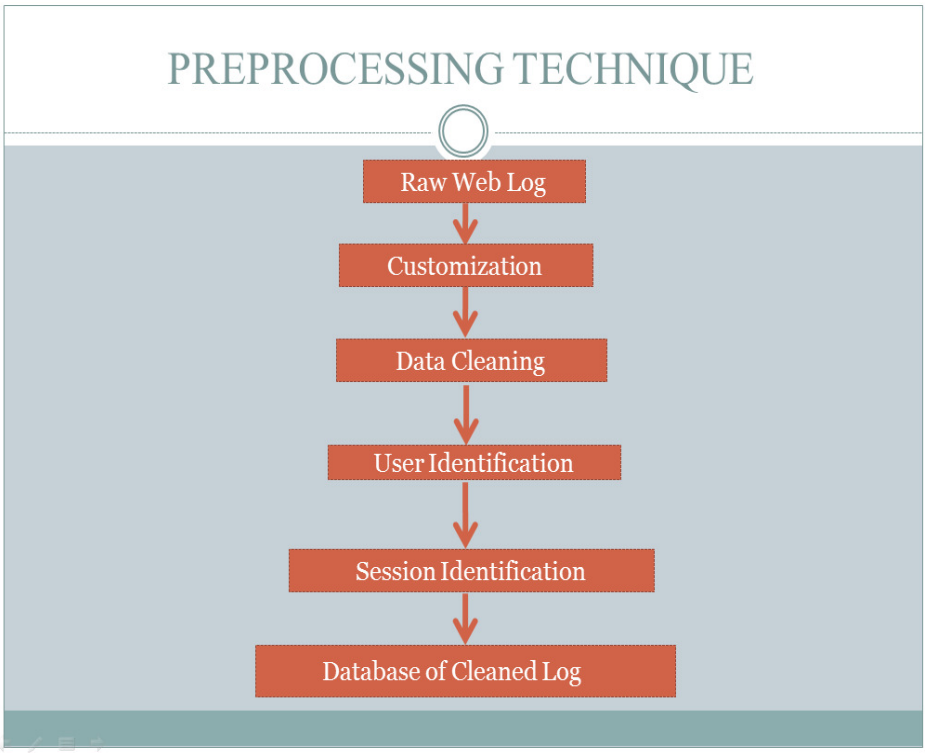

Figure 2. Pre-processing Technique Process

III. LITERATURE STUDY

The focus of literature review is to study, compare and contrast the available Recommendation system techniques. Due to large amount of irrelevant entries in the web log file, the original log file cannot be directly used in WUM process. Therefore, the pre-processing of web log file becomes more essential.

Authors Xiaogang ,Wang Yan Bai and Yue Li from Wuhan University of Science and Engineering in china has made a work fashionable An Material Recapture Way Based On Serial Admittance Patterns they has proposed an intelligent web recommender organization known as WAPPS based on successive web admission decorations. In that the successive shape mining algorithm CS-mine is used to mine everyday sequential web access patterns. The mined shapes are stored in the Pattern-tree, which is then used for similar and creating web links for operational approvals. The projected system has completed upright act with high pleasure and applicability.

In every Sequential Pattern Mining the web log should containdate-timestamp, client IP address, user ID, requested URL, and HTTP status code. With a various sample database of web access sequences the support threshold MinSup = 3 in sample database. In pattern-tree construction algorithm is based on the set of sequential web access patterns mined by the sequential pattern mining component using cs-mine. Recommendation Rules Generation Algorithm is also used for shorter matching paths usually have lower accuracy in the Web log files. With various formula they have produce the complexity Analysis.

Authors Ralf Krestel and Peter Fankhauser from Germany made work in topic Language Models and Topic Models for Personalizing Tag Recommendation in thisthey introduce an approach to

personalized tag recommendation that combines a probabilistic model of tags from the resource with tags from the user. As models we investigate simple language models as well as Latent Dirichlet Allocation. Extensive experiments on a real world dataset crawled from a big tagging system show that personalization improves tag recommendation, and our approach significantly outperforms state-of-the-art approaches.

In their work they have travelled user-centered and source- centered methods for modified tag endorsement. We likened and active a linguistic presentation tactic and a method based on Latent Dirichlet Distribution. We additional- more methodically examined the use of linguistic replicas and LDA for tag endorsement presentation that simple linguistic replicas constructed from users and possessions harvest modest presentation while overwhelming only a portion of the computational costs likened to more cultured procedures. We presented that the grouping of together approaches (LDA and LM) custom-made to manipulators and assets outstrip state-of-the-art tag recommendation procedures with admiration to a comprehensive diversity of performance metrics. It would also be stimulating to see whether the performance of the procedures variations after functional to a snap or video classification structure in its place of a labelling scheme for web folios. Finally, we also design to explore how added circumstantial information such as time, position, and existing mission can be charity to additional advance tag endorsement.

Authors Shiva Nadi, Mohamad Saraee, Mohamad Davarpanah-Jazi from Islamic Azad University of Najafabad in Iran has made a research on a topic A Fuzzy Recommender System for Dynamic Prediction of User's Behavior with various algorithm and fuzzy clustering techniques. They make the preprocessing in the raw log files whichhave useful information about access of all users to a specific website. They have integrated web content mining to web usage mining for finding users interesting rules with Fuzzy clustering techniques.

Knowledge discovery Process is created with various steps has follow they are as preprocessing for remove the unwanted images and data in log files by data cleaning process. In Web content mining they use the document clustering to group the pages in content based clusters. Integrate the log files with website pages to give recommendation for on-line user's for the first time in sites.Recommendation Process is provide if a new user starts a transaction, our model matches the new user with the most similar user clusters and provides suitable recommendations to him as Support Identification. In Match Score Identification has match score calculation defines highest match user cluster for active user also it give us a list of corresponding user clusters from the highest match score down to lowest match score.

For the Experiment they have use the log files from Information and Communication Technology Center of Isfahan municipality in Iran (F A VA) for IP address 80.191.136.6 for a period of one week. They have produced an effective result with the simulation experiment.

Authors Mozhgan Azimpour-Kivi form Sharif University of Technology in Iran and Reza Azmi from Alzahra University in Iran has made work with A Webpage Similarity Measure for Web Sessions Clustering Using Sequence Alignment. Web sessions clustering is a process of web usage mining task that aims to group web sessions with similar trends and usage patterns into clusters. This process is used to improve the web management and to provide the efficient web recommender system for the user. The proposed techniques similarity measure for comparing web sessions with sequential Alignment method and use two web sessions based on the time a user spends on a webpage, and also the frequency of visit of each page within the session.

In their work they have compare URLs for web pages similarity and also the importance of the user results in web similarity sessions. In URL they not consider the content of web page only path leading to a

web page by making the tree structure to determine the length of the longest token string among two websites. In Web Page Similarity Based on the Importance to the User is based on the interesting of the user by visiting the pages within the time session. It is calculated and compare with two pages to provide recommendation for the next users. With the similarity of web session they apply the Sequence Alignment method with the various formulas they have calculate similarity between two URLs in the webpages used by interesting user. In the feature we have a better estimation of the similarity of two web pages, we can use other methods proposed for web content mining such as Information Retrieval or semantic web approaches. Furthermore, for having a more general evaluation, we can use a larger collection of web sessions data and apply different clustering algorithms on these data.

Table 2: Various Recommendation Systems and Techniques

Author(s) Name	Name of Paper	Year of Publication	Technique Used
Xiaogang Wang, Yan Bai, Yue Li	An Information Retrieval Method Based On Sequential Access Patterns	2010	Sequential Pattern Mining
Ralf Krestel, Peter Frankhauser	Language Models & Topic Models for Personalizing Tag Recommendation	2010	Language Models and Latent Dirichlet Allocation
Shiva Nadi, Mohamad Saraee, Mohamad Davarpanah-Jazi	A Fuzzy Recommender System for Dynamic Prediction of User's Behavior	2010	Fuzzy Clustering Technique
MozhganAzimpour-Kivi, Reza Azmi	A Webpage Similarity Measure for Web Sessions Clustering Using Sequence Alignment	2011	Sequence Alignment Method

IV. CONCLUSION AND FUTURE WORK

Preprocessing of web log file is first necessary and important process for web usage mining. Cleaned data after preprocessing is solid base for pattern mining and pattern analysis. Quality of pattern mining and pattern analysis is fully dependent on preprocessing process. In this survey, we summarized the existing web log preprocessing techniques and concluded some results. Most authentic source for web usage mining considered Server log file. So it must be standardized and needs to be updated to capture user access data. Some of preprocessing techniques are applied but we can use less or even ignored preprocessing techniques to improve the quality of preprocessed data. For future work we should explore preprocessing techniques and use them with the combination of existing techniques to make the whole process more robust. New practices can offer the user to investigate the log file at changed level of thought such as user assemblies to advantage improved considerate of log file we need ordered huddling by using wished-for bundling system.

REFERENCES

[1] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization", ACM Transactions on Internet Technology, Vol. 3,No. 1, 2003, pp. 1-27.

[2] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl,"GroupLens: applying collaborative filtering to usenet news", Communications of the ACM, 40(3), 1997, pp. 77-87.

[3] T. Joachims, D. Freitag, and T. Mitchel, "WebWatcher: a tour guide for the World Wide Web", Proc. of the 5th International Joint Conference on AI, Japan, 1997, pp. 770-775.

[4] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, "Effective personalization based on association rule discovery from web usagedata", Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01), 2001.

[5] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns", Journal of Knowledge and Information Systems, Vol. 1, No. 1. 1999.

[6] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in Proceedings of the 16th International Conference on World Wide Web (WWW 2007). ACM, 2007, pp. 211–220.

[7] S. A. Golder and B. A. Huberman, "The structure of collaborative tagging systems," CoRR, vol. abs/cs/0508082, 2005.

[8] D. Bollen and H. Halpin, "An experimental analysis of suggestions in collaborative tagging," in 2009 IEEE/WIC/ACM International Conferenceon Web Intelligence, WI 2009, Milan, Italy, 15-18 September 2009,Main Conference Proceedings. IEEE, 2009, pp. 108–115.

[9] M. P. Singh, "Web Usage Mining and Personalization" , a chapter in Practical Handbook of Internet Computing, Munindar P. Singh (ed.), CRC Press, 2005.

[10] G. Castellano, A M. Fanelli, C. Mencar and M. Alessandra Torsello, "Similarity-based Fuzzy clustering for user profiling", IEEEIWICIACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, 2007.

[11] G. Castellano, AM. Fanelli, P. Plantamura, M.A Torsello, A Neuro-Fuzzy Strategy for Web Personalization, Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, 2008

[12] VA Koutsonikola and AI. akali, 2009. A fuzzy bi clustering approach to correlate web users nad web pages, IntI.Knowledge and Web Intelligence, VoU, Nos.1/2, pp. 3- 23

[13] D.H.Kraft, J. Chen, M. Bautista, MJ., and MA Vila, "Textual Information Retrieval with User Profiles Using Fuzzy Clustering and," Intelligent Exploration of the Web,Heidelberg, Germany: Physica-Verlag, 2002.

[14] S. Taherizadeh and N. Moghadam, 2009. Integrating web content mining into web usage mining for finding patterns and predicting users behaviors, International Journal of Information Science and Management, Vol, 7, No.1, pp.51-66

[15] T. Hussain, S. Asghar, and S. Fong, "A hierarchical cluster based preprocessing methodology for Web Usage Mining," in 6th International Conference on Advanced Information Management and Service (IMS), 2010, pp. 472-477.

[16] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53-65, Nov. 1987.

