Web Log Mining and Caching-A Comprehensive Survey of Research

[1]Arshi Khan, [2]Pushpraj Singh Chauhan

Computer Science and Engineering Department
Bansal Institute Of Research and Technology
Bhopal, India

***Abstract*: Due to accessibility of web data in a access amount web traffic is rapidly increasing over a past few years.As www has a huge imact on internet due to sharing of data by users.Web prefetching and caching is process to prefetch frequent pages which are to be retrieved furher and caching is to store these pages.In this paper prefetch pages stored in a proxy server cache is to be managed through cache replacement policies are proposed due to which hit ratio is likely to be improved when user make a same request in a near future.**

***Keywords*: World wide web, Web log file, Web usage mining, Cache replacement policies, Least recently used, Least frequently used, Optimal Replacement, Web prefetching, Web caching, Hit Ratio.**

I. INTRODUCTION

Web becomes a dominant platform for storage, sharing,retrieving large amount of data, so that huge number of users are increased rapidly. World wide web is playing an importantl role in today's era, as many request are fulfilled and so many information is collected and analyze by theuser. To study web user behaviour, web mining is important where preprocessing of data is done with patterngeneration algorithm and pattern analysis approach.

In preprocessing [4] all irrelevant data is removed and then various daa mining techniques is applied to discover the patterns. Patterns which are generated are that patterns which is used most frequently. In pattern discovey various data minng algorihms is applied like apriori and FP growth[10].

After pattern discovery, analysis is done by charts,graph,SQL query mechanism.

In web mining preprocessing , pattern generation,prefetching and caching concept is applied to web log files. Web log files consist of- Web server logs, Proxy server logs, Browser logs.

Field Name	Value
IP address	172.17.1.17
Remote log name	Anonymous
Authenticated user name	AJLOUN-ISA
Timestamp	2010-03-11 11:15:50
Access request	http GET
Result status code	711
Bytes transferred	34738
Referrer URL	http://www.google.jo/search ?hl=en&source=hp&q=%D8 %A7%D9%84%D8%BA% D8%AF&btnG=Google+ Search&meta=
User Agent	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)

Fig.1 Sample of web log record

In web server log all the information related to server like IP address, timestamp,method,requested URL and protocol is stored as a log file, while in proxy server log information related to proxy server is also stored. In browser log or a client side, web data log are used to store the client information .

In preprocessing process a web data log is taken and various constraints are applied as .gif file and other file extensions were removed,and unresponse server code is removed from that log file to filter useful information or a data from an unwanted or irrelevant data.In web log file user identification and server identification process is also there where user access the pages are kept

frequently and session information is also stord. After preprocessing process filter data is applied for pattern generationalgorithm like apriori algorithm and fp growth algorithms, then useful pattern is mine in web usage mining. Then pattern analysis is done representing graph and performance of pattern generation algorithms by occurrence of frequent data .

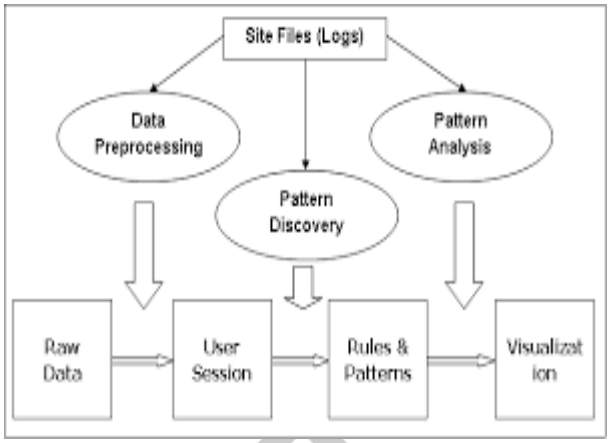

Fig.2 Preprocessing and web mining

Types of sources of data:
When user interact with the web server web data log are automatically generated. These files are stored in textual format and contain all the details of the user ad session.Web log data are available from various sources as:

1.Proxy server web log which contains proxy server information.
2.Web server logs which contain server information.
3.Browser logs which stores client information.

WEB CACHING
Web caching plays an important role in increasing the performance of web based applications.It is used to store those object in the cache which is likely to be requested in the near future.To manage the cache properly various cache replacement policies are used as size of cache is minimum to store all objects accessed by the user.
Mechanism to apply caching is of three levels- Main server level, proxy server level and client level.Implementing proxy server decrease the response time of the user with less network bandwidth. So cache replacement policies are applied on a proxy cache to improve the better hit ratio. In many caching scheme hit ratio varies in a limited range from 30-40% in respect to other replacement policies. In conventional policies web caching causes cache pollution problem where large number of objects are stored in a cache which are never requested by the user.In order to store hose objects which are likely to be requested by the user is first fetched to gain maximum hit ratio,decrease network traffic and loads on the server.Integration of web prefetching and caching is a necessary step.

WEB PREFETCHING
Web prefetching is a process to prefetch the pages in advance to fulfill the user request. To reduce latency web pages are prefetched by proxy server before a client make a request.
Web prefetching is implemented between proxy server and web server, proxy server and client[11], client and web server. Better implementation is between client and proxy server as web pages are stored in advance in a proxy cache so to reduce internet traffic.

II. LITERATURE SURVEY

In today's era access of internet is increased rapidly, there is a need to improve the user satisfaction and requirement for business, for this proper study of web user behaviour is needed. So there is a concept of web usage mining and caching which is studied and analysed by researchers.

Vijayashri Losarwar and Dr.Madhuri Joshi et al. [5], have describe about web usage mining where preprocessing of log data is done by data cleaning, by user identification and by session identification.Data cleaning is process where unwanted data is removed like .gif, .jpeg,.java script,.css,.txt file extensions. In session identification process the combination of no of pages used by the user is defined and in user identification process IP address,agent and OS act as single user.

Sheetal A. Raiyani, Shailendra jain et al. [6], described in their paper about the preprocessing, data cleaning, session and user identification process.In data cleaning process irrelevant data is removed like .jpeg,.txt,.css file extension and also with failed HTTP status. They also present the rules for identification of user and session, with path completion approach for missing page references.

Ms. Dipa Dixit, Ms. M Kiruthika et al. [7], explain two approaches for web log preprocessing through XML & text file. In XML file preprocessing is done which gives better structures to understand the web log file.In text file they describe about the attribute, record which are separated by a delimiter space for better understanding of web log files.

Surbhi Anand and Rinkle Rani et al. [9], describes about extraction of data field and data storage where data extraction of data is done through extraction algorithm implemented in java platform, after that data is stored through log table created by SQL query.

Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara et al. [10], represent about web personalization which is a combination of content based, collaborative and rule based filtering. They proposed a framework to increase web personalization and also describes the drawback of apriori and FP growth algorithm.

Nanhay Singh, Arvind Panwar and Ram Shrinagar Rao et al. [4], represents about the improvement of proxy server through web user requirement and pre fetching process. Researchers have proposed the web usage mining framework implemented in C using MATLAB and compare the results of LRU and LFU algorithm after prefetching and caching through hit ratio and byte ratio.

Mr. Rahul Mishra, Ms. Abha Choubey et al. [20], researchers have describe about the pattern generation and pattern analysis of web log data by using apriori and FP growth algorithm to calculate the frequent items. They found FP growth algorithm works better for mining frequent access patterns.

Nanhay Singh, Achin Jain, Ram Shringar Raw et al. [21], researchers have describe about the comparison analysis of different pattern recognition techniques , filtering, IP address to domain name and by recognizing the bandwidth/Hit comparison for files uploaded as image to improve the website performance by structure content delivery and presentation.

Waleed Ali , Siti Mariyam Shamsuddin, and Abdul Samad Ismail et al.[23],proposed about prefetching and caching approaches for improving the performance of web proxy server. Through this web objects which are likely to be visited in the future is prefetched and cached. Web caching and prefetching complement each other by working independently or integrated. This paper also presents the web prefetching as spatial locality to predict next objects of requested objects while web caching represents temporal locality for revisited objects .Researchers have also discussed conventional and investigated approach to integrate web caching and prefetching.

Greeshma G. Vijayan1 and Jayasudha J. S. et al. [24], describes about the issues in web traffic and network bottlenecks. As users grow rapidly day by day to access the internet so internet causes severe overloading of many sites and network links because many users didn't have patience to wait for downloading a web page. This cause web traffic and network bottlenecks so web prefetching and caching reduces the web latency. For accessing the web sites efficiently reduction of web traffic is necessary. This paper presents the web prefetching and web caching techniques to reduce the web latency. Web pre-fetching techniques and web caching reduces the web latency to predict web objects requested and also include challenges applied to a mobile environment.

Arun Pasrija et al.[1], proposed the integrated framework for web prefetching and caching. They have designed a system in such a manner so that both works in a seamless manner and to reserve certain amount of caching space reserved for prefetching. In prefetching engine objects are mined to predict frequent objects. This paper represents the integrated framework which improves web caching alone to reduce the latency and increase the network load.

Sonia Setia, Dr. Jyoti, Dr. Neelam Duhan et al.[2], describes about Web caching and web prefetching techniques to improve user perceived latency.

.

K R Baskaran, Dr. C.Kalarasan, A Sasi Nachimuthu et al. [3], . proposed prefetching using clustering technique with machine learning concept amd it is combine with support vecor machine to give better result for caching. This paer presents SVM is better than clustering technique and LFU used improves bandwidth utilization and access latency. Main aim is to gain high bandwidth utilization, by reducing load on the origin server with high access speed are possible by combining Web caching and prefetching techniques.

III. WEB USAGE MINING

To mine user access information in terms of web log data web usage mining is used where log files are analyzed to discover frequent patterns. Web usage mining is a popular research field where user interact with the web or a server to predict the user's behaviour. Number of pages accessed by the user frequently is generated and analysis is done through graph, charts ,files etc. Web usage mining consis of 3 basic steps to mine the most frequent data.

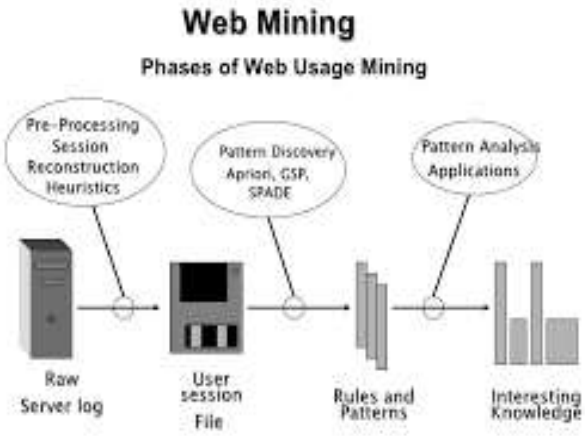

Fig.3 Web Usage Mining

1.Data Pre-processinng

Data pre-processing in web usage mining is a basic step where irrelevant data and information is removed by data cleaning and by applying other methods. Web log files are first collected and then filtered so as to analyze the web data accurately by removing various file format extensions and by user,session identification. Various preprocessing techniques[8] are data cleaning [16], user identification, session identification, data transformation and path completion is used.

2. Web Pattern Discovery

After preprocessing step interesting web patterns are discovered by data mining algorithms like apriori and FP growth, by applying association rules,sequential patterns, classification[22] and clustering. All these operations are performed to study web user behaviour.

3.Web patterns analysis

It is a process where uninteresting patterns are discovered from previous patterns discovery.Patterns are analyzed by making use of OLAP tools and by SQL query mechanism.

Web usage mining can be used in various application shown below:

a.Prefetching and Caching: To improve the performance of web server and web applications,concept of web usage mining came into existence.Prefetching and caching helps immensely for better server performance.

IV. PREFETCHING AND CACHING

Prefetching is a process of fetching most frequent data which is accessed by the user. When user request made a request from the web server, then to study the user behaviour session identification is necessary .In this pages[12] which is again and again accessed by the user are prefetched and put up in the cache. If user again make a request for the same data then it would be accesses from the cache not from he server, so the access time is going to be decrease.Prefetching concept makes the application to run faster,as already access pages or data are present in user cache. It is the most important process after pre processing, web mining,so that useful data generated is already prefetched and cached to enhance the performance of web mining[13].Through this we can easily understand the behaviour of user by generating frequent patterns by applying data mining algorithms.

To access frequent pages and to reduce the access time prefetching is used and pages are kept in the cache so that web application[14] is improved.Accessibilty of most frequent data predicts the user past behaviour and shows that new user has advantage of this analysis for page access.

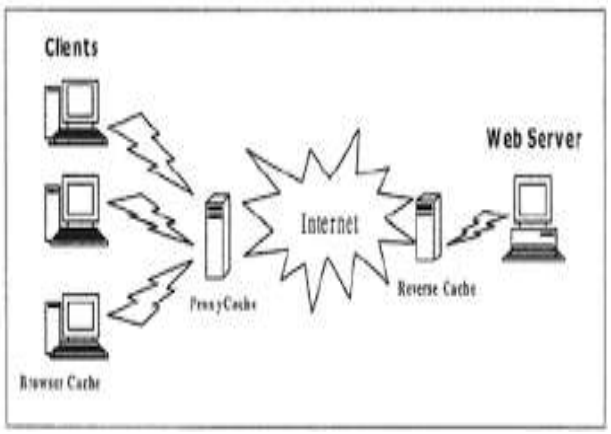

Fig.4 Prefetching and Caching

Caching is done to store frequent data or pages, so that user request is fulfill fastly. When user serves the request from the server ,then response is made from server or proxy server, but if again and again same request is generated by the user then that data is sored in a cache[12], so that faster accessibility is done. It is provided by the concept of caching where various cach replacement policies are applied to calculate the hit ratio and byte ratio.

REPLACEMENT POLICIES

Cache replacement policy is used to manage the cache according to user request .Pages are discarded according to users need,to improve the performance of cache hit ratio.Various cache replacement policies tries to remove those pages which are not likely used by user in the near future.Prefetched pages are cached so that user made the same request further. We have studied three web caching algorithms LRU , LFU with their comparison.

V.CONCLUSION

Web log files are used to depict user behavior while accessing www.In this paper we have discussed about preprocessing, mining through association rules, generating patterns of frequent data and improving web application by prefetching and caching web data with the help of replacement policies LRU, LFU..

REFERENCES

[1]) Survey on Improving the Performance of Web by Evaluation of Web Prefetching and Caching Algorithms" (Arun Pasrija) (2013)

[2] Survey of Recent Web Prefetching Techniques" (Sonia Setia, Dr. Jyoti, Dr. Neelam Duhan) (2013)

[3] Study of Web Pre-Fetching With Web Caching Based On Machine Learning Technique " (K R Baskaran, Dr. C.Kalarasan, A Sasi Nachimuthu) (2013)

[4] Nanhay Singh, Arvind Panwar and Ram Shrinagar Rao, Ambedkar Institute of Advanced Communications Technologies and Research, Delhi, India: "Enhancing the Performance of Web Proxy Server through Cluster Based Pre-fetching Techniques", International Conference on Advances in Computing, Communications and Informatics (ICACCA), 2013

[5] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

[6] Sheetal A. Raiyani, Shailendra jain, "Efficient Preprocessing technique using Web log mining", 1Department of CSE(Software System),Technocrats Institute of Technology, Bhopal,India; 2Department of CSE, Technocrats Institute of Technology, Bhopal,India, International Journal of Advancements in Research & Technology, Volume 1, Issue6, November-2012 1 ISSN 2278-7763

[7] Ms. Dipa Dixit Lecturer, Ms. M Kiruthika Assistant Professor, Fr.CRIT, Vashi, "Preprocessing Of Web Logs", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010, 2447-2452

[8] Ankit R Kharwar1, Chandni A Naik2, Niyanta K Desai3, 1Assistant Professor, Department of Computer, Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli , 2,3Student of M.Tech Computer Engineering in Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli, "A Complete Pre-Processing Method for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250- 2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 10, October 2013)

[9] Surbhi Anand, Surbhi Anand, Department of Computer Science & Engineering, Thapar University, Patiala-147004 (India), "An Efficient Algorithm for Data Cleaning of Log File using File Extensions", International Journal of Computer Applications (0975 – 888) Volume 48– No.8, June 2012

[10] Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara, "A Process Oriented Perception of Personalization Techniques in Web Mining", International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-2, January 2013

[11] V. Shanmuga Priya1, S. Sakthivel, Department of computer science, Periyar University, TamilNadu, India, "An Implementation Of Web Personalization Using Web Mining Techniques", International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, IJCSMC, Vol. 2, Issue. 6, June 2013, pg.145 – 150.

[12] V. Sathiyamoorthi, Department of CSE, Sona College of Technology, Salemi-5, and Dr.Murali Bhaskaran, Principal,Paavai College of Engineering, Paachal, 637018, Tamil Nadu, India, "Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data through Proxy Server", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.11, November 2011

[13] Ramya C, Dr. Shreedhara K S and Kavitha G, M.Tech (Final Year), Professor & Chairman and Lecturer, Dept. of Studies in CS&E, U.B.D.T College of Engineering, Davangere Davangere University, Karnataka, INDIA cramyac@gmail.com and ks_shreedhara@yahoo.com, "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process", International Conference on Communication and Electronics Information (ICCEI 2011)

[14] Abdul Rahaman Wahab Sait, and Dr.T.Meyappan, "Data Preprocessing and Transformation Technique to Generate Pattern from the Web Log", International conference on Computer Science and Information Systems (ICSIS'2014) Oct 17-18, 2014 Dubai (UAE) [15] Wasvand Chandrama, Prof. P.R.Devale, Prof. Ravindra Murumkar, Department of Information technology, Research scholar of Bharati Vidyapeeth University College of Engineering, Pune, Maharashtra 411046, India., ISSN 2348 – 7968, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.

[16] Sheetal A. Raiyani, Shailendra Jain, Dept. of CSE(SS),TIT,Bhopal], "Enhance Preprocessing Technique Distinct User Identification using Web Log Usage data", ISSN:2249-5789, International Journal of Computer Science & Communication Networks,Vol 2(4), 526-530

[17] Michal Munk, Jozef Kapusta, Peter Švec, Constantine the Philosopher University in Nitra, Department of Informatics, Tr. A.Hlinku 1, 949 74 Nitra, Slovakia, "Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor", International Conference on Computational Science, ICCS 2010

[18] Mr. Shivkumar Khosla, Mrs. Varunakshi Bhojane, Department of Computer Engineering, Mumbai University, India, "Capturing Web Log and Performing Preprocessing of the User's Accessing Distance Education System", International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.2, Issue.5, Sep.-Oct. 2012 pp- 3128-3130 ISSN: 2249-6645 [19] Doru Tanasa and Brigitte Trousse, AxIS Project Team, INRIA Sophia Antipolis, "Advanced Data Preprocessing for Intersites Web Usage Mining", 1094-7167/04/$20.00 © 2004 IEEE 59, Published by the IEEE Computer Society

[20] Mr. Rahul Mishra, Ms. Abha Choubey, Computer Science & CSVTU India, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 9, September 2012

[21] Nanhay Singh, Achin Jain, Ram Shringar Raw , nsingh1973@gmail.com , achin_jain25@yahoo.com , rsrao08@yahoo.in, Ambedkar Institute of Advanced communication Technologies & Research Delhi, India, "COMPARISON ANALYSIS OF WEB USAGE MINING USING PATTERN RECOGNITION TECHNIQUES", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

[22] Chaitra L Mugali and Asst. Prof. Padma Dandannavar, "WEB LOGS PRE-PROCESSING AND ANALYSIS: A Survey", International Journal of Emerging Technology In Computer Science and Electronics, Volume 14, Issue 2, ISSN: 0976-1353

[23] A Survey of Web Caching and Prefetching"(Waleed Ali , Siti Mariyam Shamsuddin, and Abdul Samad Ismail)(2011)

[24] A Survey On Web Pre-Fetching and Web Caching Techniques in a Mobile Environment"(Greeshma G. Vijayan1 and Jayasudha J. S.) (2012)