IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-4 E-ISSN: 2347-2693

An Approach For Web Log Pre-Processing And Evidence Preservation For Web Mining

Richa Chourasia[1*] and Preeti Choudhary[2]

[1*,2]*Department of CSE, Infinity Management & Engineering Institute, Sagar, M.P., India*

www.ijcaonline.org

Received: 09/03/2014 Revised: 24/03/2014 Accepted: 17/04/201x4 Published: 30/04/2014

Abstract— The time needed to scrape out any true information is for the most part used on information preprocessing. The information preprocessing stage lays the foundation for information mining with which, the client extricate and distinguish pertinent data from the World Wide Web. In this paper, we examine information preprocessing systems and different steps included in getting the obliged substance adequately. A powerful web log preprocessing technique is constantly proposed for web log preprocessing to concentrate the client designs. The information cleaning method uproots the unessential passages from web log and sifting calculation disposes of the uninterested characteristics from log record.

Index Term— *Preprocessing, Web usage, Web log*

I. INTRODUCTION

With the explosive growth of data sources accessible on the World Wide Web and also rapid increasing pace of adoption to web commerce in global business, the world wide web has evolved into a gold mine that contains or dynamically generates data that's beneficial to E-businesses. Most of the organizations accentuation on learning visitor's activities through net analysis, and establish the patterns in the visitor's behavior. The results obtained from the online examination, once amalgamate with organization data warehouses supply nice opportunities for the close to future. The online usage mining method involves the discovery of patterns from one or more net servers. It also helps organizations to predict the value of any specific client, cross selling methods for numerous products and also the effectiveness of promotional campaigns etc.

During the past few years the world Wide web has become the biggest and hottest means of communication and information proliferation and promulgation. It provides a platform for exchanging varied information. The quantity of information accessible on the net is increasing chop-chop with the explosive growth of the World Wide Web and the advent of E-Commerce. While users area unit given additional service options and information, it's become tougher for them to find the relevant information of their interest, the problem unremarkably known as information overload.

Web mining may be generally authentic as assay and assay of advantageous recommendation from the World Wide Web. Hence, supported the altered model and altered suggests that to admission information, the action of web mining can be represented into 2 parts: web Usage Mining and web Contents Mining. Web Contents Mining may be declared as the machine-driven obtain and retrieval of information and assets accessible from many sites and on-line databases admitting search engines / web spiders. Web Usage Mining may be authentic as the action of anecdotic and assay of user admission patterns non-inheritable from mining of log files and associated knowledge from a specific computing device.

Corresponding Author: *Richa Chourasia*

Dept. of CSE, Infinity Management & Engineering Institute, Sagar, M.P., India

II. LOG FILES

Log files are significant sources for deciding the health standing of a system and is used to capture the events happened at intervals a in a computer and networks. Logs are assortment of log entries and every entry contains data associated with a particular event that has taken place within a network or in computer system. several logs in the association will accommodate records related to pc security that are generated by several sources, as well as operative systems on servers, digital computer computers, networking equipments and alternative security software's, like firewalls, antivirus, detection of intrusion and its hindrance systems and plenty of alternative applications. Systematic Routine log assay is benign for communicative security incidents, fallacious activity, policy violations and different operational issues. Logs are helpful for the auditing and cyber forensic analysis, it additionally supports internal investigations, distinctive operational trends and semipermanent issues [1].

Initially, logs were used for troubleshooting issues, however these days they're used for several functions among most organizations and associations, like optimizing system and network performance, recording users actions, and providing information helpful for examining vindictive movement. Logs have developed to carry information known with varied numerous kinds of occasions happening within the systems and frameworks. In a company, several logs contain records associated with pc security; common samples of these pc

security logs are audit logs that track user authentication tries and security device logs that record attainable attacks.

Log file study a users question behavior whereas user navigates a look web site. Understanding the users direction preferences helps to enhance question behavior. In fact, the information of the foremost doubtless user access patterns permits service provides to customize and adapt their sites interface for individual users also on improve the sites static structure at intervals the broader hypertext system.

The web log records in addition helps digital forensics in seizing and testing digital computer, obtaining electronic confirmation for digital wrongdoing examinations and searching once machine records for the electoral standards of proof.

III. RELATED WORK

In the research work shown in [16], net Personalization is defined, that is the process of customizing the content and structure of an internet web site to the precise and individual desires of every user taking advantage of the user's steering behavior. sites belonging to a specific class have some similarity in their structure. This general structure of sites can be deduced from the position of links, text pictures|and pictures} (including images and graphs). This data can be easily extracted from a hypertext markup language document. [17] the most knowledge supply in the net usage mining and personalization process is the data residing on the net sites logs. net logs record every visit to a page of the net server hosting it. The entries of an internet log file consists of several fields that represent the date and therefore the time of the request, the ip variety of the visitor's computer(client), the URI request , the http standing code came to the consumer, and so on. The log knowledge collected at net access or application servers reflects steering behaviour data of users in terms of access patterns.

Physically, a page is a collection of web things, generated statically or dynamically, causative to the display of the ends up in response to a user action. A page set is a collection of whole pages among a web site. User session is a sequence of sites clicked by one user during a particular amount. A user session is sometimes dominated by one specific guidance task, which is exhibited through a set of visited relevant pages that contribute greatly to the task conceptually. The guidance interest/preference on one explicit page is diagrammatical by its vital weight value, which is dependent on user visiting duration or click range. The user sessions (or referred to as usage data), which square measure primarily collected within the server logs, is transformed into a method format for the purpose of research analysis via an information cleansing and preparation method. In one word, usage information is a collection of user sessions, which is within the sort of weight distribution over the page area.

An implementation of knowledge preprocessing system for web usage mining and the details of rule for path completion square measure conferred in Yan Li's paper. when user session identification, the missing pages in user access ways square measure appended by exploitation the referrer-based technique which is an effective answer to the issues introduced by exploitation the local caching and proxy servers. The reference length of pages in complete path is modified by considering the typical reference length of auxiliary pages which is estimated prior to through the greatest forward references. As verified by practical web access log, the rule path completion, planned by Yan LI, efficiently appends the lost information and improves the responsibleness of access information for additional calculations of web usage mining.

In internet Usage Mining (WUM), internet session cluster plays a key role to classify internet guests on the premise of user click history and similarity live. Swarm based mostly internet session cluster helps in many ways to manage internet resources effectively like web personalization, schema modification, web site modification and internet server performance. Tasawar Hussain, Dr. Sohail Asghar[3] projected a framework for internet session cluster at preprocessing level of internet usage mining. The framework covers information preprocessing steps to arrange the online log information and converts the specific journal information into numerical data. A session vector was obtained, so applicable similarity and swarm improvement may well be applied to cluster the online log information. Author says that the graded cluster based mostly approach enhances the prevailing internet session techniques for additional structured data regarding the user sessions.

Doru Tanasa[4], in his analysis brought 2 vital contributions for a WUM method. They projected an entire methodology for preprocessing the online logs and a discordant general methodology with 3 approaches (as well as associated concrete methods) for the invention of successive patterns with an occasional support.

Huiping Peng[5] used FP-growth rule for process the online log records and obtained a collection of frequent access patterns. Then victimization the mixture of browse power and web site topology power of association rules for internet mining they found a brand new pattern to produce valuable information for the positioning construction. In order to unravel some existing issues in ancient information preprocessing technology for journal mining, associate degree improved information preprocessing technology is employed by the author ling Zheng[6].

The identification strategy supported the referred website is adopted at the stage of user identification, that is simpler than the normal one supported site topology. At stage of Session Identification, the strategy supported fastened priori threshold combined with session reconstruction is introduced. First, the initial session set is developed by the

tactic of fastened priori threshold, so the initial session set is optimized by victimization session reconstruction. Experiments have proven that advanced information preprocessing technology will enhance the standard of information preprocessing results

JIANG Chang-bin and Chen Li[7] brought about a Web log data preprocessing algorithm based on collaborative filtering. It can perform user session identification fast and flexibly even though statistic data are not enough and user history-visiting records are absence.

Data compression may be a matured domain, and variety of generic and special purpose compression rules and utilities are offered which provides sensible compression rations and timings. General purpose compression utilities like bzip, bzip2[8], gzip [9] use generalized compression algorithms like Burrows-Wheeler remodel [10], Lmlpel Ziv rule [11] to call many. These utilities may offer sensible compression schemes for big scale cluster event logs. However, the performance of log compression is additional ameliorated, by investing concrete attributes ordinarily determined inside these astronomically huge scale cluster logs. 7zip [12] compression utility, accessible on windows and UNIX system platforms, implements several compression rules together with one PPM(Prediction by Partial matching)[13] that is one among the most effective algorithm on English text, and LZMA that is usually offers sensible compression ratios than bzip2.

Apart from these generic compression utilities, Bal´azs, Andr´as[14] discusses the log compression for internet servers. Sahoo et al [15] discusses the filtering of failure logs of enormous scale clusters tested on Blue Gene/L information which might be used as a lossy (Non - lossless) compression technique. Pzip compression proposes an improved compression schema for tabular information with fastened length records with a specific column widths. To the most effective of our information, no work is completed specifically to manage great amount of event logs in a very lossless manner for giant scale clusters whereas rising the compression magnitude relation and timings.

IV. PROPOSED METHODOLOGY

In this paper, we'll going to} accentuate on web utilization mining and therefore the reasons are very simple: With the popularity of E-commerce, the method organizations do businesses has been transmuted. The term e-commerce, primarily characterised by performing electronic transactions through web, has provided a efficacious and cost-effective method of doing business transactions.

Web utilization mining is achieved initially by reporting guests traffic data predicated on internet server log files and alternative source of traffic data. The web server log files were used at first by the system administrators and webmasters for the needs of analyzing the traffic. Besides server logs are additionally accustomed record and trace the visitors' on-line comportment's.

The proposed system for log preprocessing provides parts of web usage knowledge exist in sources as numerous as web logs, referral logs, registration files and conjointly index server logs. Such data has to be integrated to make a whole knowledge set for data processing. yet, before the combination of the information, log files got to be filtered/cleaned, using techniques like filtering the information to eliminate outliers and/or impertinent things, grouping individual page accesses into linguistics units.

Filtering the information to eliminate impertinent things is very important for the analysis of net traffic. Elimination of impertinent enteries will be accomplished by checking the suffix of the uniform resource locator name, that informs the system; what format these kind of files are. taking an example, the embedded graphics will be filtered out from the web log file, whose suffix is typically in the form of "gif", "jpeg", "jpg", "GIF", "JPG", "JPEG", will be removed.

The proposed approach when performing all pre-processing steps ensures the integrity, genuineness, acceptableness and forensically sound proof and therefore will be used to trace out the criminal that commits the cyber crime. The planned mechanism additionally provides the safety to the log files and makes the log repository to the digital forensics. Web log Pre-processing is the method of customizing the content and structure of an internet website to the precise and individual desires of every user taking advantage of the user's direction behaviour.

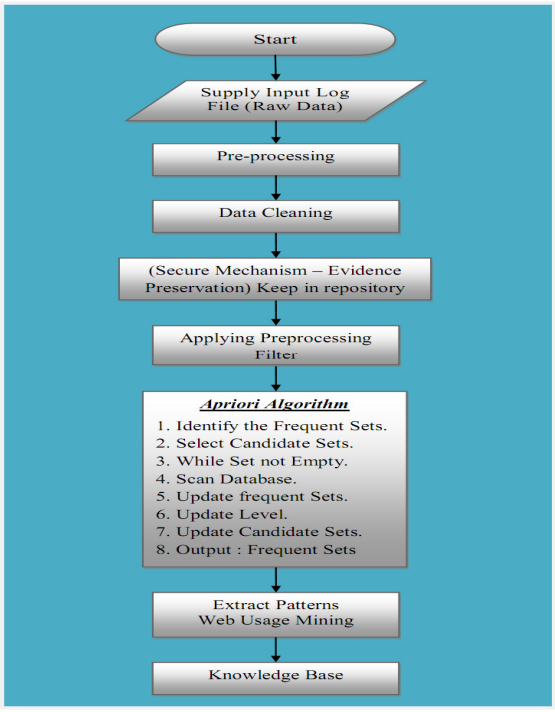

Figure 1: Proposed Algorithm

The steps of internet (web) personalization method include:

• The assortment of internet knowledge.
• The modeling and categorization of these knowledge (preprocessing phase)
• The analysis of the collected knowledge.
• The determination of the actions that ought to be performed.

Association Rule Mining algorithms generally solely determine patterns that occur within the original kind throughout the information. One limitation of the many association rule mining formulas like the Apriori algorithm is that solely the information entries of the precise match for the candidate patterns could contribute to the support of the candidate pattern. This creates a haul for databases containing several little variations between otherwise similar patterns.

Association Rule Mining may be a nontrivial method of identifying valid, probably helpful, and ultimately fascinating pattern (rule) in information. An Apriori rule might lead to an oversized range of rules which can prove to be useless or impractical for analysis. Additional some rules may have not any impact on firewall action like continual rules or rules that doesn't yield any action on the correct hand aspect. As a result, these rules are filtered for additional analysis. The main objective is to generalize specific and distinctive rules to additional general rules.

V. IMPLEMANTATION & RESULTS

This section has highlighted thematic areas where a novel digital technology may bring improvement to the forensic method. The proposed scheme implemented by developing a user application, using Microsoft .Net Framework 4.0, Visual Studio 2008. Which is tested on windows environment with Intel Core i3 2nd gen processor, 2 GB RAM. Since both efforts are cooperative in nature, there is perpetually lots of support from documentation and mailing lists. Bugs are fixed quickly, and requests for options are continuously detected, evaluated, and if possible, implemented. The implemented system works according to the proposed approach [18], designed in 2 modules:

1- Evidence Preservation Module

This process of cyber forensic investigation isn't quite so easy with computer system data on magnetic media. Magnetic media is subject to physical stresses like magnetism, extremes of warmth and cold, and, even, physical or shock. additionally, it's an easy matter to alter logical proof if one will gain access thereto. Knowing and having the ability to prove that the proof has been altered may be terribly tough as a result of the alteration might leave no indication that it ever occurred.

In order to confirm that the evidence investigators use in court is the same evidence they collected, we'd like to be able to mark it logically and seal it in such a way that it merely isn't accessible to anyone, except administrators and investigators. in addition, so nobody will produce their own, slightly completely different (presumably, to their benefit) version of the evidence, and present it as a twin of actual, Therefore, there's a necessity of methodology of building that each one evidence meets the best evidence rule which it's all identical to the original.

The Evidence Preservation Module allows the system administrator to take the log files from all the existing and active nodes of a network. These log files are then used by the implemented software to apply the protection scheme. This scheme will make the log file secure from any alteration. So, that if someone may alter the actual log file. It does not affect the secure logs, which is stored and maintained at the dedicated forensic evidence preservation and extraction system.

This module helps the administrator to make a safe and secure repository of log files, which helps the forensic investigators during the investigation of any cyber crime scene. Although the concept of evidence gathering server is simple but very effective, as there is a secured copy of log files are stored in an organized manner. Hence, the forensic investigator does not need to check the individual nodes of the network because, all it gets from the proposed evidence gathering server.

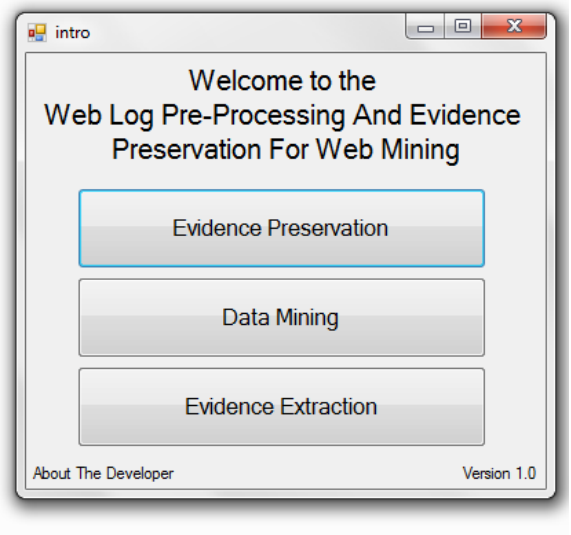

Figure 2: GUI of proposed software
- Reading of Secure file

2- Evidence Extraction Module

Log examination is probably the single most productive part of your investigation if the logs are kept properly. It is also very tedious, especially when the logs are from multiple machines and are thousands of lines long.

The Evidence Extraction module allows the administrator to make available the resources to the forensic investigators to their investigation by ensuring that the evidences stored at evidence gathering server are safe, secure and unaltered. So, that the investigators can use it for investigation and can be use as prime evidence for the judicial system against the accused.

```
Rule   (Support, Confidence)
18 -> 12   (30.1282%, 97.9167%)
12 -> 37   (30.1282%, 97.9167%)
37 -> 12   (30.1282%, 97.9167%)
37 -> 30   (30.1282%, 97.9167%)
39 -> 30   (30.1282%, 97.9167%)
37 -> 39   (30.1282%, 97.9167%)
39 -> 37   (30.1282%, 97.9167%)
37 -> 30,39   (30.1282%, 97.9167%)
39 -> 30,37   (30.1282%, 97.9167%)
```

Figure 3: Rules identified from the web log file

Table 1: Results Obtained from the proposed approach

Log Files	Evidence P reservation Time	Size	Evidence Extraction Time	Size
rc_pc131216.txt	2 ms	314 kb	3 ms	252 kb
rc_pc140116.txt	4 ms	467 kb	7 ms	418 kb
rc_pc140216.txt	4 ms	596 kb	9 ms	531 kb

Table 2: Comparative Study with other approaches

Features	Proposed Technique	Log Dispersion Method	Sudheer Reddy	Tamper Resistance
Log Cleaning	✓	-	✓	✓
Security	✓	✓	-	✓
Security With hash Code	✓	✓	-	✓
Zero Redundancy	✓	-	-	-

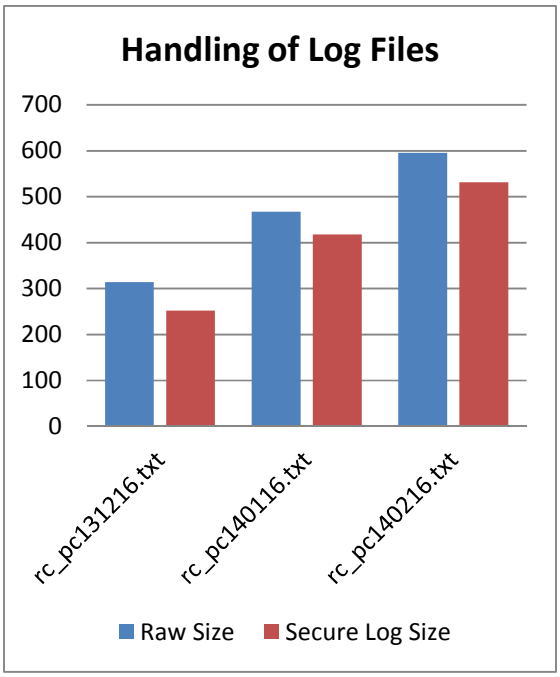

Figure 4: handling of log files at evidence gathering server

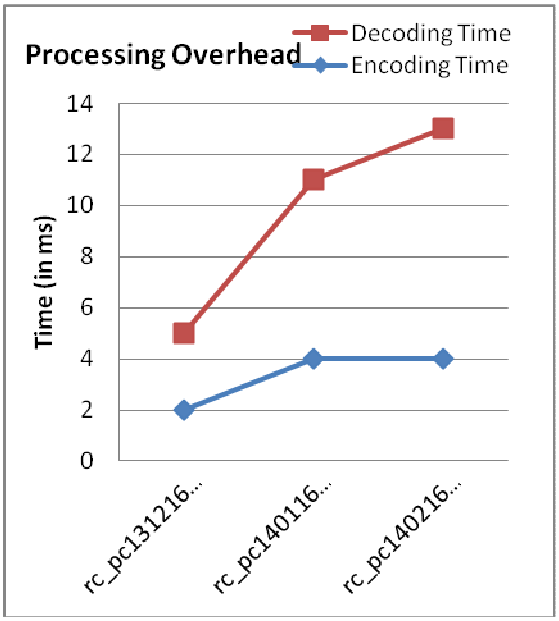

Figure 5: Time overhead in processing of log files

Experimental results are carried out on log files taken from web server of different time-periods. The table 4, shows the actual experimental scenario, where one can easily find out the performance of proposed system. While in table 5, it is clearly seen the novelty of the proposed approach, as it provides security with very less processing overhead and zero redundancy. In this dissertation, we also presented the use of business intelligene by web usage mining by taking

the same log files which are preserved by the proposed technique.

VI. CONCLUSION

In this paper we have described a fully reversible log file repository scheme capable of significantly reducing the amount of space required to store the compressed and preprocessed log, the obtained test results show it manages to improve compression of different types of log files. It is lossless, fully automatic (it requires no human assistance before or during the compression process), and it does not impose any constraints on the log file size.

REFERENCES

[1] Muhammad Kamran Ahmed, Mukhtar Hussain and Asad Raza "An Automated User Transparent Approach to log Web URLs for Forensic Analysis" Fifth International Conference on IT Security Incident Management and IT Forensics 2009.

[2] Yan LI, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique In Web Usage Mining", IEEE International Symposium On Computer Science and Computational Technology, pp. 554-559, 2008.

[3] Tasawar Hussain, Dr. Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence ", 6th International Conference on Emerging Technologies (ICET) IEEE, pp. 21-26, 2010.

[4] Doru Tanasa and Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining ", Published by the IEEE Computer Society, pp. 59-65, March/April 2004.

[5] Huiping Peng, "Discovery of Interesting Association Rules Based On Web Usage Mining", IEEE Coference, pp.272-275, 2010.

[6] Ling Zheng, Hui Gui and Feng Li, " Optimized Data Preprocessing Technology For Web Log Mining", IEEE International Conference On Computer Design and Applications(ICCDA), pp. VI-19-VI-21,2010.

[7] JING Chang-bin and Chen Li, " Web Log Data Preprocessing Based On Collaborative Filtering ", IEEE 2nd International Workshop On Education Technology and Computer Science, pp.118-121, 2010.

[8] Bzip2 and libbzip2 project official home page, http://www.bzip.org/ .

[9] gzip official home page, algorithm description, http://www.gzip.org/algorithm.txt.

[10] M. Nelson. Data Compression with the Burrows-Wheeler Transform. In Dr. Dobbs Journal September 1996.

[11] J. Ziv, A. Lamapel. A Universal Algorithm for Sequential Data Compression. In IEEE Transactions on Information Theory, May 1977.

[12] 7 zip project official home page, http://www.7-zip.org.

[13] M. Drini´c, D. Kirovski et al. PPMexe: PPM for Compressing Software. In Proceedings of the Data Compression Conference, IEEE, 2002.

[14] Bal´azs R´ACZ, A. Luk ´acs. High density compression of log files. In Proceedings of Data Compression Conference (DCC'04), IEEE Page 557, 2004.

[15] Y. Liang, Y. Y. Zhang et al. Filtering Failure Logs for a Blue Gene/L Prototype. In Proceedings of IEEE International Conference on Dependable Systems and Networks , 2005.

[16] 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.

[17] D.Vasumathi, D.Vasumathi and K.Suresh, "Effective Web Personalization Using Clustering", IEEE IAMA, 2009.

[18] Richa Chourasia, Prof. Preeti Choudhary, "A Survey On Web Log Pre-Processing And Evidence Preservation For Web Mining", International Journal Of Innovative Research In Technology & Science, Volume1, Issue 4, Issn:2321-1156.

AUTHORS PROFILE

Ms. Richa CHourasia is a M.Tech Scholar at Infinity Management & Engineering Institute, Sagar, M.P and her area of interest is web security and log file preservation and management.