ISSN:2321-1156
International Journal of Innovative Research in Technology & Science(IJIRTS)

A SURVEY ON WEB LOG PRE-PROCESSING AND EVIDENCE PRESERVATION FOR WEB MINING

Richa Chourasia, M.Tech Scholar, Infinity Management & Engineering Institute, Sagar, M.P.;
Prof. Preeti Choudhary, Asst. Professor, Infinity Management & Engineering Institute, Sagar, M.P.;

Abstract

It is known to all that maximum of the time required to dig out any real world data is generally spent on data preprocessing. The data preprocessing phase lays the groundwork for data mining with which, the user extract and identify relevant information from the World Wide Web. In this paper, we discusses data preprocessing methods and various steps involved in getting the required content effectively. An effective web log preprocessing methodology is being proposed for web log preprocessing to extract the user patterns. The data cleaning technique removes the irrelevant entries from web log and filtering algorithm discards the uninterested attributes from log file.

Introduction

With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce in real world business, the world wide web has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses. Most of the organizations emphasizing on studying visitor's activities through web analysis, and identify the patterns in the visitor's behavior. The results obtained from the web analysis, when amalgamate with organizations data warehouses offer great opportunities for the near future. The web usage mining process involves the discovery of patterns from one or more Web servers. It also help organizations to predict the value of any specific customer, cross marketing strategies for various products and the effectiveness of promotional campaigns etc.

During the past few years the World Wide Web has become the biggest and most popular way of communication and information proliferation and promulgation. It provides a platform for exchanging various information. The volume of information available on the internet is increasing rapidly with the explosive growth of the World Wide Web and the advent of E-Commerce. While users are provided with more service options and information, it has become more difficult for them to find the relevant information of their interest, the problem commonly known as information overload.

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Hence, Based on the different criterion and different ways to obtain information, the process of web mining can be divided into two major parts: Web Usage Mining and Web Contents Mining. Web Contents Mining can be described as the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines / web spiders. Web Usage Mining can be defined as the process of identifying and analysis of user access patterns obtained from mining of log files and associated data from a particular Web site

Log Files

Log files are considerable sources for determining the health status of a system and used to capture the events happened within a computer system and networks. Logs are collection of log entries and each entry contains information related to a specific event that has taken place within a system or network. Many logs within an association contain records associated with computer security which are generated by many sources, including operating systems on servers, workstation computers, networking equipments and other security software's, such as antivirus, firewalls, intrusion detection and prevention systems and many other applications. Routine log analysis is beneficial for identifying security incidents, fraudulent activity, policy violations and other operational problems. Logs are also useful for performing auditing and forensic analysis, it also supports internal investigations, identifying operational trends and long-term problems [1].

Initially, logs were used for troubleshooting problems, but now days they are used for many functions within most organizations and associations, such as optimizing system and network performance, recording users actions, and providing data useful for investigating malicious activity. Logs have evolved to contain information related to many different types of events occurring within the networks and systems. In an organization, many logs contain records related to computer security; common examples of these computer security logs are audit logs that track user authentication attempts and security device logs that record possible attacks.

Log file study a users query behavior while user navigates a search site. Understanding the users navigational preferences helps to improve query behavior. In fact, the knowledge of the most likely user access patterns allows service provides to customize and adapt their sites interface for individual users as well as to improve the sites static structure within the wider hypertext system.

The web log files also helps cyber forensic in seizing and probing computer, obtaining electronic evidence for cyber crime investigations and maintaining computer records for the federal rules of evidence.

LITERATURE SURVEY

In the research work shown in [16], Web Personalization is defined, which h is the process of customizing the content and structure of a web site to the specific and individual needs of each user taking advantage of the user's navigational behavior. Web pages belonging to a particular category have some similarity in their structure. This general structure of web pages can be deduced from the placement of links, text and images (including images and graphs). This information can be easily extracted from a HTML document. [17] The main data source in the web usage mining and personalization process is the information residing on the web sites logs. Web logs record every visit to a page of the web server hosting it. The entries of a web log file consists of several fields which represent the date and the time of the request, the IP number of the visitor's computer(client), the URI request , the HTTP status code returned to the client, and so on. The log data collected at Web access or application servers reflects navigational behaviour knowledge of users in terms of access patterns.

Physically, a page is a collection of Web items, generated statically or dynamically, contributing to the display of the results in response to a user action. A page set is a collection of whole pages within a site. User session is a sequence of Web pages clicked by a single user during a specific period. A user session is usually dominated by one specific navigational task, which is exhibited through a set of visited relevant pages that contribute greatly to the task conceptually. The navigational interest/preference on one particular page is represented by its significant weight value, which is dependent on user visiting duration or click number. The user sessions (or called usage data), which are mainly collected in the server logs, can be transformed into a processed data format for the purpose of analysis via a data preparing and cleaning process. In one word, usage data is a collection of user sessions, which is in the form of weight distribution over the page space.

An implementation of data preprocessing system for web usage mining and the details of algorithm for path completion are presented in Yan Li's paper [2]. After user session identification, the missing pages in user access paths are appended by using the referrer-based method which is an effective solution to the problems introduced by using proxy servers and local caching. The reference length of pages in complete path is modified by considering the average reference length of auxiliary pages which is estimated in advance through the maximal forward references and the reference length algorithms. As verified by practical web access log, the algorithm path completion, proposed by Yan LI, efficiently appends the lost information and improves the reliability of access data for further web usage mining calculations.

In Web Usage Mining (WUM), web session clustering plays a key role to classify web visitors on the basis of user click history and similarity measure. Swarm based web session clustering helps in many ways to manage the web resources effectively such as web personalization, schema modification, website modification and web server performance. Tasawar Hussain, Dr. Sohail Asghar[3] proposed a framework for web session clustering at preprocessing level of web usage mining. The framework covers the data preprocessing steps to prepare the web log data and converts the categorical web log data into numerical data. A session vector was obtained, so that appropriate similarity and swarm optimization could be applied to cluster the web log data. Author says that the hierarchical cluster based approach enhances the existing web session techniques for more structured information about the user sessions.

Doru Tanasa[4], in his research brought two significant contributions for a WUM process. They proposed a complete methodology for preprocessing the Web logs and a divisive general methodology with three approaches (as well as associated concrete methods) for the discovery of sequential patterns with a low support.

Huiping Peng[5] used FP-growth algorithm for processing the web log records and obtained a set of frequent access patterns. Then using the combination of browse interestingness and site topology interestingness of association rules for web mining they discovered a new pattern to provide valuable data for the site construction.

In order to solve some existing problems in traditional data preprocessing technology for web log mining, an improved data preprocessing technology is used by the author ling Zheng[6].

The identification strategy based on the referred web page is adopted at the stage of user identification, which is more

effective than the traditional one based on web site topology. At stage of Session Identification, the strategy based on fixed priori threshold combined with session reconstruction is introduced. First, the initial session set is developed by the method of fixed priori threshold, and then the initial session set is optimized by using session reconstruction. Experiments have proved that advanced data preprocessing technology can enhance the quality of data 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 is a matured area, and a number of generic and special purpose compression algorithm and utilities are available resulting good compression rations and timings. General purpose compression utilities like bzip, bzip2[8], gzip [9] use generalized compression algorithms like Burrows-Wheeler Transform [10], Lmlpel Ziv algorithm [11] to name a few. These utilities can provide good compression schemes for large scale cluster event logs. However, the performance of log compression can be further improved, by leveraging specific attributes commonly observed within these large scale cluster logs. 7zip [12] compression utility, available on windows and UNIX platforms, implements many compression algorithms including one PPM(Prediction by Partial matching)[13] which is one of the best performing algorithm on English text, and LZMA which is generally gives good compression ratios than bzip2.

Apart from these generic compression utilities, Bal´azs, Andr´as[14] discusses the log compression for web servers. Sahoo et al [15] discusses the filtering of failure logs of large scale clusters tested on Blue Gene/L data which can be used as a lossy compression technique. Pzip compression proposes a better compression schema for tabular data with fixed length records and fixed column widths. To the best of our knowledge, no work is done specifically to manage large amount of event logs in a lossless manner for large scale clusters while improving the compression ratio and timings.

PROPOSED APPROACH

In this paper, we will emphasize on web usage mining and the reasons are very simple: With the popularity of E-commerce, the way organizations are doing businesses has been changed. The term e-commerce, mainly characterized by performing electronic transactions through Internet, has provided a effective and cost-efficient way of doing business transactions.

Web usage mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data. The web server log files were used initially by the system administrators and webmasters for the purposes of analyzing the traffic. Besides server logs are also used to record and trace the visitors' on-line behaviors.

The proposed system for log preprocessing provides Portions of Web usage data exist in sources as diverse as Web logs, referral logs, registration files and also index server logs. Such information needs to be integrated to form a complete data set for data mining. yet, before the integration of the data, log files need to be filtered/cleaned, using techniques like filtering the raw data to eliminate outliers and/or irrelevant items, grouping individual page accesses into semantic units.

Filtering the raw data to eliminate irrelevant items is important for the analysis of web traffic. Elimination of irrelevant enteries can be accomplished by checking the suffix of the URL name, which informs the system; what format these kind of files are. taking an example, the embedded graphics can be filtered out from the Web log file, whose suffix is usually the form of "gif", "jpeg", "jpg", "GIF", "JPG", "JPEG", can be removed.

The proposed approach after performing all preprocessing steps ensures the integrity, authenticity, admissibility and forensically sound evidence and hence can be used as the evidence to trace out the criminal that commits the cyber crime. The proposed mechanism also provides the security to the log files and makes the log repository to the digital forensics. Web Pre-processing is the process of customizing the content and structure of a web site to the specific and individual needs of each user taking advantage of the user's navigational behaviour.

The steps of the web personalization process include:

- The collection of web data.
- The modeling and categorization of these data (preprocessing phase)
- The analysis of the collected data.
- The determination of the actions that should be performed.

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.