ISSN: 0976-3104
SPECIAL ISSUE: Emerging Technologies in Networking and Security (ETNS)
Krishnagandhi and Dhas.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

WEB LOG MINING - A STUDY

Geetha Krishnagandhi [1]and Suresh Gnana Dhas[2]

[1] Bharathiar University, Coimbatore,Department of BCA, G.T.N. Arts College (SSC), Dindigul, TN, INDIA
[2]Dept.of Computer Science and Engineering, Vivekanandha College of engineering for women (Aut.)Tiruchengode, Namakkal Dt. TN, INDIA

ABSTRACT

The analysis of web log files may give information that are useful for improving the services offered by web portals and information access and retrieval tools, giving information on problems occurred to the users and gain knowledge from the web. A particular useful kind of knowledge, which can be applied to improve the performance of web service. Web log file analysis began with the purpose to offer to web site administrators a way to ensure adequate bandwidth and server capacity to their organization. It is way to evaluate the effectiveness of a Web site and its information access tools is through the mining of web log files.This study reports on initial findings on a specific aspect that is highly relevant to web mining and extracting useful patterns from the web.

Published on: 08th– August-2016

KEY WORDS

Web Mining; Web Log Mining; Web Structure Mining; Web Content Mining; Web Usage Mining

***Corresponding author: Email:** geethachouthri@gmail.com; **Tel.:** +91 8760821422

INTRODUCTION

Data Mining

Data Mining is an important research area as there is a huge amount of data available in most of the applications. To extract useful information and Knowledge from that large amount of data. It is an interdisciplinary research field to database systems, statistics, machine learning, information retrieval etc. Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge presentation are the important data mining processes [1].

The World Wide Web has become one of the most important media to store, share and distribute information. Web Mining is the data mining technique that automatically discovers or extracts the information from web. Mining the web is discovering knowledge from hypertext and WWW. The www is one of the longest rising areas of intelligence accessible internet and the number is still increasing. Every websites attract millions of users and visitors. These visitors behind vast amount of website.

Web log mining is a promising tool to study user behaviours, benefit web-site designers with a better organization and services. Effective web log mining system consists of data processing, sequential pattern mining and visualization [2].

Many existing systems that can be used to analyse the traversal path of web-site visitors, their performance is still far from satisfactory. In often unclear where a specific document is located. And usually a great portion of time is needed to look for and find the appropriate information. [3]

When user accesses websites are recorded in web log file, web server log file is a simple plain text file. Log file contain noisy and ambiguous data which may affect results of mining process [4].

THE IIOAB JOURNAL
ISSN: 0976-3104

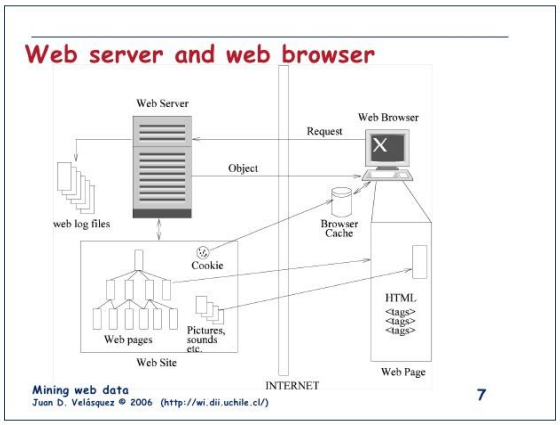

Fig: 1 - Web Server and Web Browser – Mining Web Data

WEB MINING CATEGORIES

Web mining can be broadly classified into three domains

- Web Structure Mining
- Web Content Mining
- Web Usage Mining

Fig: 2. Web Mining Classification

Web Structure Mining

Web Structure Mining is to deal with the structure of the hyperlinks within the web itself. It is the process to discover the model of link structure of the web pages. To catalogue the links, generate information such as the similarity and relations among them by taking the advantage of hyperlink topology. The goal of structure mining is to generate structured summary about the website and web page.
It is the technique to analysis and explain the links between different web pages and web sites. It mainly focuses on developing web crawlers. It works on hyperlinks and mines the topology of their arrangement. Web structure mining can classify the web pages and produce results such as the similarity and relationship between different web sites [5].

Web Content Mining

Web Content Mining is the process of retrieving the information from web into more structured forms and indexing the information to retrieve quickly. It focuses mainly on the structure within a document i.e. inner document level. It is also related with text mining because much of the web contents are text, but is also quite different from these because web data is mainly semi structured in nature and text mining focuses on unstructured text [6].

THE IIOAB JOURNAL
ISSN: 0976-3104

It focuses on extracting knowledge from the contents or their description of the web documents. It involves techniques the summarising, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behaviours.

Web Usage Mining

Web Usage Mining is the process of discovering a meaningful patterns from data generated by client server transactions on one or more web servers [7].

It focuses on digging the usage of web contents from the logs maintained on web servers, cookies logs, applications server logs etc. It works on how and when user moves from one type of content to others. Thus, it can provide association between different contents.

OVERVIEW OF WEB MINING

With the rapid and explosive growth of information available over the internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of web data, web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc., As a result, web users are always drowning in an ocean of information and facing the problem of information overload when interacting with the web.

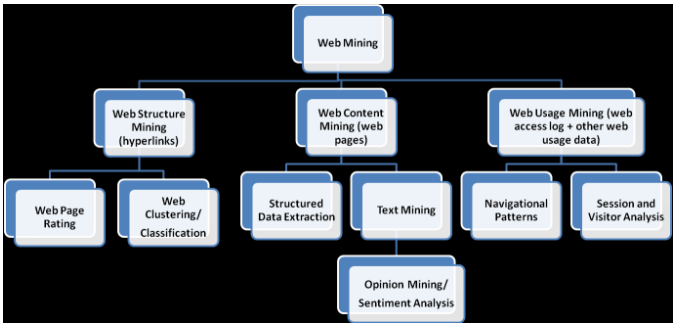

Fig: 3. Overview of Web Mining

……..

Web site is a collection number of web pages grouped under the same domain name. In web mining, data is distributed on various places and web pages may contain data is not only text form it could text, audio, images video and images and navigate between them via hyperlinks. When user accesses website, log files are created. Log file recorded information about the each user. Tremendous uses of web, Web log files are growing at faster rate.

Problems in Web related Search

- Finding relevant information
- Finding needed information
- Learning useful knowledge
- Recommendation of information

Web involves three types of data, the actual data from WWW, the web log data obtained by the users who browsed the web pages and the web structure data.
Web Server maintains the web log file. Log files are located in different locations like web server, web proxy server, and client browser.

THE IIOAB JOURNAL
ISSN: 0976-3104

Web Log File

Web server log file is a simple plain text file which record information about each user. Log file contain information about user name, IP address, date, time, bytes transferred, access request. A time a user requests a resource from the particular site. When user submit request to a web server that activity are recorded in web log file. Log file range 1KB to 100MB [8].

Log file gives significant information to web server.log file information about:

1. Which pages were requested in website?
2. How many bytes sent to user from server?
3. What type of error occurs?

When user submit request to a web server that activity are recorded in web log file. Log file used for debugging purpose. Analysing log file are used to detecting attacks on web.

Fig: 4. Sample Web Log File

Location of Log File

Web log file is located in three different locations.

Web server logs

Web log files provide most accurate and complete usage of data to web server. The log file do not record cached pages visited. Data of log files are sensitive, personal information so web server keeps them closed.

Web proxy server

Web proxy server takes HTTP request from user, gives them to web server, then result passed to web server and return to user. Client send request to web server via proxy server. The two disadvantages are: Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction.The request interception is limited.

Client Browser

Log file can reside in clients' browser window itself. HTTP cookies are pieces of information generated by a web server and stored in user's computer, ready for future access.

Types Of Web Server Logs

Generally four types of server logs.

Access log file

Data of all incoming request and information about client of server. Access log records all requests that are processed by server.

THE IIOAB JOURNAL
ISSN: 0976-3104

Error log file

List of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log [joshila]. Access and error logs are mostly used, but agent and referrer log may or may not enable at server.

Agent log file

Information about user's browser, browser version. Referrer log file: This file provides information about link and redirects visitor to site.

Log File Format

Web log file is a simple plain text file which record information about each user. Display of log files data in three different format

- W3C Extended log file format
- NCSA common log file format
- IIS log file format

NCSA and IIS log file format the data logged for each request is fixed. W3C format allows user to choose properties, user want to log for each request [9].

W3C Extended Log File Format

W3C log format is default log file format on IIS server. Field are separated by space, time is recorded as GMT (Greenwich Mean Time). It can be customized that is administrators can add or remove fields depending on what information want to record. In W3C format of year is YYYY-MM-DD. Omitting unwanted attributes field when log file size is limited [W3C].

NCSA Common Log File Format

National Centre for Supercomputing Application format. NCSA is recorded basic information about user request such as user name and remote host name, date, time, request type, HTTP status code and numbers of bytes send by server. NCSA is fixed format, it cannot customized. It is available for website but not for FTP site. Format of year is DD/MMM/YYYY. Fields are separated by space, time is local time.

IIS Log File Format

IIS format is not customized, it is fixed ASCII format. Fields are separated by comma, easy to read. Time recorded in local time. Its records more information than NCSA format. Fields in IIS are client IP address, user name, date, time, service and instance, server name, server IP address, time taken, client bytes sent, server bytes sent, service status code, windows status code, request type, target of operation, parameters.

PRE-PROCESSING

The aim of data pre-processing is to select essential features, clean data from irrelevant records and finally transform raw data into sessions. In the pre-processing, the data cleaning process includes removal of records, graphics, videos and the formal information. The failed records with the failed HTTP status code and robots cleaning. The Web usage Mining phases are

1) Data Pre-processing.
2) Pattern Discovery.
3) Pattern Analysis Stage

THE IIOAB JOURNAL
ISSN: 0976-3104

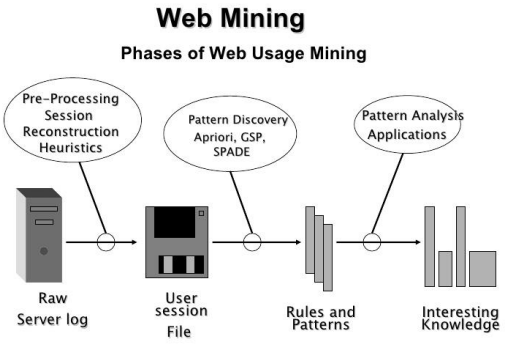

Fig: 5. Phases of Web Usage Mining

……

First, the data pre-processing phase includes the web log data cleaning, user identification, session identification and data transformation [10]. Condensed and transformed in order to retrieve and analyse significant and useful information.

So the main steps of this phase are:
1) Extract the web logs that collect the data in the web server.
2) Clean the web logs and remove the redundant information.
3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in pattern analysis to create summary reports.

Second Pattern Mining can be performed on log records to find association patterns, sequential patterns and trend of web accessing.

The pattern discovery phase involves the discovery of frequent sequences.

The pattern analysis phase involves the analysis of the frequent patterns.

Data Preprocessing

Before using data for pattern discovery it need to be cleaned to get data in specific format. Pre-processing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data pre-processing is to improve data quality and increase mining accuracy. Pre-processing consists of field extraction, data cleansing. This phase is probably the most complex and ungrateful step of the overall process. The main task is to "clean" the raw web data such as web logs. Data pre-processing involves data cleaning, user identification and data transformation [11].

In data cleaning phase, the web log is examined and irrelevant on redundant items such as image, sound, video files and executable gif files and HTML files, removal of HTTP errors recorded created by crawlers, removed from the log.

The user's identification phase involves identification of users from the log data.

1. A new IP identifies a new user.
2. If the same IP is used, but different operating system in terms of type and version is being used, then that is considered as new user.

Pre-processing is necessary, because log file contain noisy & ambiguous data which may affect result of mining process. Some of web log file data are unnecessary for analysis process and could affect detection of web attack.

Data pre-processing is an important steps to filter and organize only appropriate information before applying any web mining algorithm. Pre-processing redice log file size also increase quality of available data. The purpose of data pre-processing is to improve data quality and increase mining accuracy. Pre-processing consists of field extraction, data cleansing, user identification, session identification. In this paper main task is to "clean" the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process. So the main steps of this phase are:

- Extract the web logs that collect the data in the web server.

THE IIOAB JOURNAL
ISSN: 0976-3104

- Clean the web logs and remove the redundant information.

MATERIAL AND METHODS

Web log data pre-processing is a complex process and takes 80% of total mining process. Log data is pre-treated (cleaning) to get reliable data. There are four steps in pre-processing of the log data.

Field Extraction

The log entry contains various fields which need to be separate out for the processing. The process of separating field from the single line of the log file is known as field extraction. The server used different characters which work as separators. The most used separator character is',' or 'space' character.

Data Cleaning

Data Cleaning is the removal of outlies or irrelevant data. It is the process to remove noisy and unnecessary data. Remove log entry nodes contain extension like jpg, gif means remove request such as multimedia files, image, page style file [12].

Data cleaning is usually site specific, and involves extraneous references to embedded objects that may not be important for purpose of analysis, including references to style files, graphics or sound files. Therefore some of entries are useless for analysis process that is cleaned from the log files. By data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data.

Analysing the huge amounts of records in server logs is a cumbersome activity. So initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, are also downloaded which are not useful for further analysis are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be removed from log files [9]. Thus removal process in the experiment includes the records of graphics, videos and the format information: The records have filename extension of GIF, JPEG, CSS and so on, which can be found in the URI field of the every record, can be removed. This extension files are not actually the user interested web page, rather it is just the documents embedded in the web page. So it is not necessary to include in identifying the user interested web pages. This cleaning process helps in discarding unnecessary evaluation and also helps in fast identification of user interested patterns.

The records with the failed HTTP status code: The HTTP status code is then considered in the next process for cleaning. By examining the status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. This cleaning process will further reduce the evaluation time for determining the used interested patterns. Method fields, Robots cleaning are the software tool to extract from the website.

This helps in accurate detection of user interested patterns by providing only the relevant web logs. Only the patterns that are much interested by the user will be resulted in the final phase of identification if this cleaning process is performed before start identifying the user interested patterns.

USER IDENTIFICATION

This step identify individual user by using their IP address. If new IP address, there is a new user. If IP address is same but browser version or operating system is different then it represents different user.
The log file after cleaning is considered as Web Usage Log Set – WULS. The next important and complex step is unique user identification. The complexity is due to the local cache and proxy servers. To overcome this cookies are used. But users may disable cookies. Another Solution is to collect registration data from users. But users neglect to give their information due to privacy concerns. So majority of records does not contain any information in the user-id and authentication fields. The fields which are useful to find unique users and sessions are :

1) A new IP Address indicates a new user.

THE IIOAB JOURNAL
ISSN: 0976-3104

2) User Agent - The same IP but different web browsers or different operating systems in terms of type and version means a new user.
3) Referrer URL – Suppose a the topology of a site is available, if a request for a page originates from the same IP Address as other already visited pages, and no direct hyperlink exists between these pages, it indicates a new user.

Users and sessions are identified by using these fields as follows. If two records has same IP address check for browser information. If user agent value is same for both records then they are identified as from same user.

Session Identification

Each user spends total time in each web page. Session means time duration spent in web pages. A referrer-based method is used for identifying sessions. If IP address, browsers and operating systems are same, the referrer information should be taken. The CS referrer is checked, a new user session is identified if the URL in the refer URL-field is a large interval usually more than 30 minutes between the accessing time of this record.

The goal of session identification is to divide the page accesses of each user into individual sessions. These sessions are used as data vectors in various classification, prediction, clustering into groups and other tasks. If URL, in the referrer URL, field in current record is not accessed previously or if referrer URL field is empty then it is considered as a new user session. Reconsturction of accurate user sessions from server access logs is a challenge task and time oriented heuristics with a time limit of 30 min is followed.

From WULS, the set of user sessions are extracted as referrer based method and time oriented heuristics. Every record in WULS must belong to a session and every record in WULS can being to one user session only. After grouping the records into sessions the path completion step follows.

Path Completion

Path completion step is carried out to identify missing pages due to cache and 'Back' Path set is the incomplete accesses pages in a user session. It is extracted from every user session set.

Path Combination and Completion : Path Set (PS) is access path of every USID identified from USS. It is defined as: PS = {USID,(URI),Date, RLength)…(URI, Date, RLength)}

Where RLength is computed for every record in data cleaning stage. After identify path for each USID path combination is done if two consecutive pages are same. In the user session if any of the URL specified in the Referrer URL is not equal to the URL. In the previous record then that URL in the Referrer URL field of current record is inserted into this session and thus path completion is obtained.

Pattern Discovery

The Pattern Discovery Phase is the key component of the Web mining. Pattern discovery converge the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition, etc. applied to the Web domain and to the available data.

Content Clustering

At the final stage of pre-processing of content data have m web pages that have used to mining content of these pages and integrating them with usage and structure patterns. The symbols m, n, and k to denote the number of documents, the number of terms, and the number of clusters, respectively. The symbol S to denote the set of n documents that we want to cluster, Cj, C2, • • •, Ck to denote each one of the k clusters, and n1 n2 , . . . , nk to denote the sizes of the corresponding clusters. The K-means clustering algorithm is one of the clustering algorithm that is used the vector-space model to represent each document. The best known approach that is based on partitioning is k-means clustering, a simple and efficient algorithm used by statisticians for decades. The idea is to represent the cluster by the centroid of the documents that belong to that cluster (the centroid of cluster C is defined as. The cluster membership is determined by finding the most similar cluster centroid for each document. After clustering done, similar pages are assigned to same cluster that can be used in recommendation process.

THE IIOAB JOURNAL
ISSN: 0976-3104

Page Ranking

Finally, by employing the HITS algorithm on structure data system generate ranked pages. In HITS concept, Kleinberg identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to Kleinberg, “Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs”.[13]

Pattern Anaylsis

Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data, although is in some cases considered to be nearly synonymous with machine learning. Pattern recognition systems are in many cases trained from labeled "training" data (supervised learning), but when no labeled data are available other algorithms can be used to discover previously unknown patterns (unsupervised learning).

The terms pattern recognition, machine learning, data mining and knowledge discovery in databases (KDD) are hard to separate, as they largely overlap in their scope. Machine learning is the common term for supervised learning methods and originates from artificial intelligence, whereas KDD and data mining have a larger focus on unsupervised methods and stronger connection to business use. In pattern recognition, there may be a higher interest to formalize, explain and visualize the pattern; whereas machine learning traditionally focuses on maximizing the recognition rates. Yet, all of these domains have evolved substantially from their roots in artificial intelligence, engineering and statistics; and have become increasingly similar by integrating developments and ideas from each other.

In machine learning, **pattern recognition** is the assignment of a label to a given input value. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of *classes* (for example, determine whether a given email is "spam" or "non-spam"). However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values and parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence.

Pattern recognition algorithms generally aim to provide a reasonable answer for all possible inputs and to perform "most likely" matching of the inputs, taking into account their statistical variation. This is opposed to *pattern matching* algorithms, which look for exact matches in the input with pre-existing patterns.

CONCLUSION

In this paper, I presented a preliminary analysis of the web log mining according to the methodology for gathering information from log files, mining information to extract knowledge from them, analyse the useful patterns and how to presented. The aim of this work was to report on initial findings about the study of web content mining, web structure mining and web usage mining and how to extract patterns from the data from web log files and discover the knowledge

FUTURE WORK

A new web page recommendation framework is to be proposed to make efficient Web Log Mining. First, users' navigational patterns are extracted from web usage data simultaneously web content data and web structure data is also taken after pre-processing and pattern discovery is performed on a server data’s and based on the pattern discovery the recommendations are generated. Proposed framework is to combine the special features of web mining algorithms and to create a new algorithm to maintain efficient web log mining, so it is advantageous as compare to the previous hybrid recommendation frameworks.

THE IIOAB JOURNAL
ISSN: 0976-3104

CONFLICT OF INTEREST
None

ACKNOWLEDGEMENT
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] J Han, M Kamber.[2000] Data Mining: Concepts and Techniques Morgan Kaufmann.

[2] SVeeramalai,N.Jaisankar and A.Kannan.[August - 2010] Efficient web log mining using enhanced Apriori algorithm with Hash Tree and Fuzzy.

[3] Ramakrishnan Srikant.[2011] Mining web logs to improve website organization. , IBM Almaden Research Center.

[4] B.Madasamy, J.Jebamalar Tamilselvi.General web knowledge mining framework (IJCSE).

[5] Priyanka Patil and Ujwala Patil, Preprocessing of web server log file for web mining. (NCETCT 2012).

[6] Renata Ivancsy, Istvan Vajk.[2006] Frequent Pattern Mining in Web Log Data, Acta Polytechnica Hungarica .3(1)

[7] Qiang Yang,Charles X.Ling and JianFeng Gao. [2013]Mining web logs for actionable knowledge.

[8] Zhengtu Yang,Yitong Wang,Masaru Kitsuregawa.[2011] An effective system for mining web log.

[9] J Vellingiri and Chethur Pandian. A Novel technique for web log mining with better data cleaning and transaction identification [ISSN 1549-3636].

[10] Rajashree Shetlar.[2006] Sequential Pattern Mining from web log data. [ISSN 2250-3676].

[11] Jiawei Han and Micheline Kamber.[2015] Data Mining concepts and technologies - Morgan Kaufmann Publishers, 2nd edition-Delhi.

[12] Margaret H.Danham, S.Sridhar [2007] Data Mining Introductory and Advanced Topics –-Dorling Kindersley India PVT Ltd. -New Delhi.

[13] Bing Liu.[2007] Web Data Mining Exploring Hyperlinks, contents and usage data –Springer –New York.