International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 15994-15998

Review on Techniques and Applications Involved in Web Usage Mining

B Bhavani[1], Dr. V. Sucharita[2] & Dr. K.V.V. Satyanarana[3]

[1]Research scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Guntur, India.

[1]Orcid: 0000-0002-3219-153X

[2]Professor, Department of Computer Science and Engineering, Narayana Engineering College, Gudur, India.

[3]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Guntur, India.

Abstract

In the era of internet, the number of web users increased enormously along with countless web applications with which large volumes of information is being stored, retrieved and distributed in the web. This huge amount of web data is to be processed to achieve customer needs and satisfaction, because the web data is unstructured or semi-structured, one cannot apply data mining techniques directly on such data, this is where Web Mining has been evolved. Web mining is an interesting discipline in the domain of data mining where information mining strategies are utilized for extracting data from the web servers. Web mining can be classified into three expansive zones of mining. Web Content Mining, Web Structure Mining and Web Usage Mining. Web use mining includes examining use design. Distinguishing the use examples of clients is exceptionally vital as there is immense data accessible in the internet. While the client connects with the web, web usage mining uses the methods which can predict the user behaviour. This paper explains the process in web usage mining and distinctive applications and devices utilized as a part of web use mining.

Keywords: web mining; web usage mining; pre-processing; pattern discovery; pattern analysis; applications.

INTRODUCTION

The web is multiplying in estimate each six to ten months and there are different kinds of vast resources of information available in the World Wide Web. With this growing amount of information web has turned into a critical resource for information and knowledge. To effectuate the user needs and satisfaction, there is an immediate requirement to upgrade the techniques and tools which map the growing information. Web Mining is the way toward applying information mining strategies on web data to automatically and quickly extract useful information and also to discover interesting patterns. Although web mining has its establishes profoundly in information mining, it isn't same as information mining. The unpredictability of web mining depends on the unstructured nature of web data. Web mining entail the analysis of a particular web sites server logs basically called web server logs, which contain the complete interaction list of a particular user when accessing the web site. The knowledgeable information from such analysis from server logs is very helpful in almost all the web applications.

Contingent on the information to be mined, web mining is separated into three categories, web content mining: is the procedure of extricating data or knowledge from the content of web documents such as text and images, web structure mining: is the procedure of extricating data or knowledge from the structure of web pages and impact of this structure on traversal through these web pages, web usage mining: is the procedure of extricating data or knowledge i.e.; valuable use designs from web information, to understand the user's needs and implement them in the web applications. The vital uses of web use mining are personalization, web site design, ecommerce, web advertising or web marketing, fraud detection, transaction analysis etc. Summary of Web Mining and its types are presented in the table 1.

WEB USAGE MINING

Web utilization mining is centred around learn about web clients and their cooperations with sites. The objective of web utilization mining is to find rapidly and consequently, client's entrance designs from web log information. Through web utilization mining, server log, enrollment data, and other relative data left by client access can be mined. Steps involved in internet usage mining process:

1) Web data collection

2) Pre-processing

3) Pattern discovery and

4) Pattern analysis

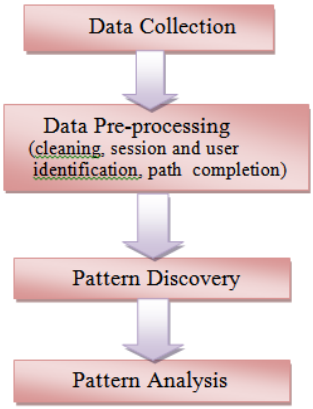

Figure 1: Basic web usage mining phases.

1) Data Collection

The extraction of log information can be done by getting to web log documents where web log information is put away and there are diverse sorts of web log information document stores are available they are.

A) Web server logs(server side log data):

Web server logs are stored on server side. Web server logs contain data like IP address, asked for URL, time stamp, number of bytes, convention utilized and so on. This data is typically displayed in a standard configuration. A standout amongst the most prominent log record arrange is the Common Log Format (CLF), A typical log design document is made by the web server to monitor the solicitations that happen on a site.

Figure 2: Common Log Format

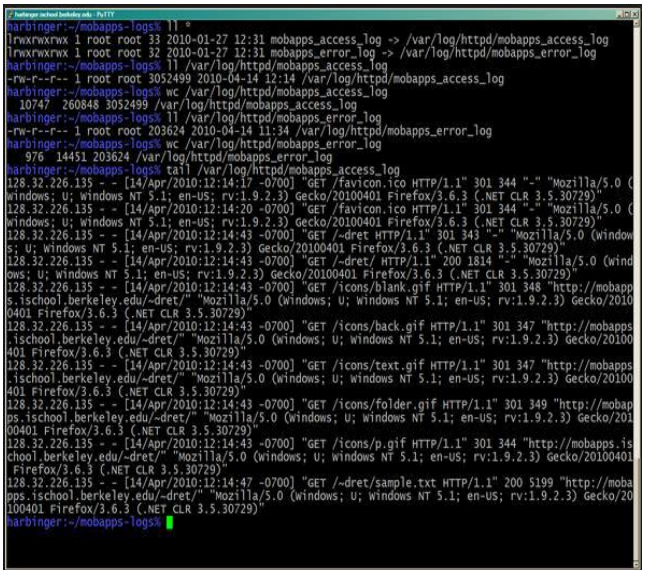

Figure 3: Example of server log.

B) Proxy server logs(proxy side log data):

Proxy server logs are put away on the intermediary server. At whatever point the real server can't react to the client asks for, the solicitations are taken care of by the intermediary servers for the benefit of the principal servers, as of now, intermediary server logs are created. These intermediary server logs contain some extra data identified with intermediary server notwithstanding that of web server log records.

C) Browser logs(client side log data):

Browser logs are collected from the clients machine where clients accesses the websites. Client data is found through sending remote agents enforced in Java or JavaScript and are then appended in web pages; they are used to chunk information from the client such as user navigation history. Client data is more reliable than server data as they avoid the problems like caching and IP misinterpretation. However, this data needs cooperation on the part of users who often restrict the operation of Java and JavaScript programs for security reasons.

2) Data Pre-Processing

Data pre-processing step is very crucial in web usage mining.The immediate phase after having collected huge amounts of various data sources is data preprocessing. Data should be consistent and integrated in order for them to be used in next phase that is pattern discovery. Data preparing involves information cleaning, client distinguishing proof, session recognizable proof and way completion.

A) Data Cleaning:

The objective of information cleaning is to take out unessential things, unimportant records in web get to log will be disposed of amid information cleaning. Since the point of Web Usage Mining is to get the user‟s traversal designs, following two sorts of records are

- pointless and ought to be expelled.
- The documents of designs, recordings

furthermore, the arrangement data, they are found in the fields of URI of each and every file

• The files with the HTTP status code failed. By inspecting the Status field of each record in the web get to log, the files with status codes over 299 or below 200 are removed.

B) User and Session Identification:

Crafted by client and session distinguishing proof recognize the diverse client sessions from the web get to log. Client distinguishing proof is, to recognize who collaborate with the site and which pages are gotten to. The objective of session distinguishing proof is to partition the page gets to of every client at once into singular sessions. A session is a progression

of website pages which the client peruse in a solitary access. The troubles to finish this progression are presented by utilizing intermediary servers, e.g. a server log may contain distinctive clients with same IP address. To unravel such issues a referrer-based strategy is proposed. The principles received to recognize client sessions can be expressed as takes after:

- The various IP addresses distinguish various users;
- On the off chance that the IP addresses are comparative, at that point the diverse programs and working frameworks demonstrate distinctive clients;
- On the off chance that the majority of the IP address, programs and working frameworks are same, the referrer data should be considered;
- The session recognized by run 3 may contain more than one visit by a similar client at various time, the time situated heuristics is then used to isolate the distinctive visits into various client sessions. In the wake of consolidating the records in web sign into client sessions, the way finish calculation ought to be utilized for removing the total client get to way.

C) Path Completion:

Another essential advance in information preprocessing is way finish. There are a few issues that emerge way deficiency, for instance, neighborhood reserve, specialist store, "post" procedure and program's "back" catch can bring about some imperative get to not show up in the entrance log document, and the first number of Uniform Resource Locators(URL) vary from that of recorded in the log. Utilizing neighborhood reserving and intermediary servers likewise build the troubles for way consummation as there is a probability for the clients to get to the pages from a nearby store or from the intermediary servers without leaving any sign in server's entrance log. Subsequently, the client gets to ways are not completely saved on the web get to log. To recognize client's traversal design, the missing pages in the client get way ought to be annexed. The objective of the way consummation is to fulfill this errand. The better results of information pre-preparing will enhance the mined examples' substance and quality likewise spare calculation's running time. The structure of weblog records is not the same as the information in database or information distribution center. They are not organized and finish because of different reasons. So it is particularly important to pre-process web log records in web utilization mining. Through information pre-preparing, the weblog can be changed into another information structure, which is anything but difficult to be mined.

3) Pattern discovery

In pattern discovery phase the data mining techniques such as Association, Clustering, Sequential Analysis, and classification are performed on data for pattern discovery.

A) Association:

This technique relies on generating frequent patterns and rules. After preprocessing the data in web log file presents interesting facts such as number of URL visits by number of users by which one can identify frequently accessed web pages by users which can help to understand user needs. The association rule focuses on discovery of relations between pages visited by users on web site. Association rule can be used to relate the web page that most often used by the single user session. Several algorithms like Apriori, Eclat, Frequent Pattern tree etc. are used to perform association rule mining.

B) Clustering:

Clustering is a technique to combine a group of users or data items (pages) together, which have identical characteristics. It can further help in the advancement and execution of prospective marketing policies. Clustering of users helps to invent the group of users, who have identical navigation pattern. It is very helpful for inferring user statistics to implement market distribution in E-commerce operations or provide personalized Web content to the individual users. The bunching of pages is helpful for Internet web indexes and Web specialist co-ops, since they can be utilized to find the gatherings of pages having related substance

C) Sequential pattern:

In order to perform sequential pattern analysis there are various algorithms such as Apriori, SPADE, GSP, PrefixSpan and Spam. This analysis is used to find that a suspected user visit a particular link X followed by link Y in a time period. By using this analysis we can discover the suspected user psychology which is useful in crime detection.

D) Classification:

This technique map a data element into one of distinct predefined classes, which guide to build in a profile of users relating to a particular class or category. This requires extraction and selection of features that first characterize the properties of a given category or class. directed inductive learning calculations, for example, choice tree classifiers, naïve Bayesian classifiers, k-nearest neighbour classifier, Support Vector Machines etc., can be utilized to perform grouping.

4) Pattern Analysis

A valuable model or standard pattern for specific web usage mining application is discovered in pattern analysis phase. Techniques used for pattern analysis are visualization

technique, OLAP technique, data and knowledge querying and usability analysis.

A) OLAP (Online Analytical Processing Technique):

Deals, showcasing, administration announcing, business process administration, financial reporting etc are some of the applications of OLAP (Online Analytical Processing Technique). It is a powerful model for strategic evaluation of relational database which is very effective for use in trading. Relational journalism, business agility, data mining and many other are associated to OLAP.

B) Data and Knowledge Querying:

Most common method used for analyzing patterns is SQL. This is an important part of internet use mining in which we analyze the several reasons of abnormal behaviors of users. By the use of SQL we find some explicit results from database, for instance suspicious session in database created by the users like failure status code of http protocol in very short span of time.

C**) Usability analysis:**

Is a modeling technique for extricating the behavior of user on the web site.

D) Visualization Technique:

The behavior of web users can be effectively represented by graphical method. Graphical method is representation of visualization technique

WEB USAGE MINING APPLICATIONS

The general objective of Web Usage Mining is to accumulate fascinating data about client's route designs. This data can be utilized later to enhance the site from the clients' perspective. The outcomes created by the mining of weblogs can be utilized for different purposes A) To personalize the delivery of web content;

B) To improve user navigation through prefetching and caching;

C) To improve web design; or in e-commerce sites

D) To improve the customer satisfaction.

Table 1: Summary of web mining

			Main Data	View Of Data	Representation	Method	Application
WEB MINING	**Web Content Mining**	IR View	-Text docs -Hypertext docs	-Structured -Unstructured	-Bag of words, n-gram terms -phrases, ontology or concepts -Relational	-Machine learning -Statistical	-Categorization -Clustering -Finding patterns in text
		DB View	-Hypertext docs	-Semi structured -website as DB	-Edge labelled graph -Relational	-Property algorithms -Association rules	-Finding frequent sub structures -Web site schema discovery
	Web Structure Mining		-Link structure	-Link structure	-Graph	-Property algorithms	-Categorization -Clustering
	Web Usage Mining		-Server logs -Browser logs	-Interactivity	-Relational table -Graph	-Machine learning -Statistical -Association rules	-Site construction -Marketing -User modelling

A) Personalization of Web Content :

Web Usage Mining strategies can be utilized to give customized web client encounter. For example, progressively, it is conceivable to anticipate the client conduct by contrasting the present route design and regular examples which were extricated from past weblog. Here, proposal frameworks are the most widely recognized application; their point is to prescribe fascinating connections of items which could enthusiasm to clients.

B) Prefetching and Caching:

The outcomes delivered by Web Usage Mining can be misused to improve the execution of web servers and online applications. With the utilization of weblogs that store client's entrance history can be utilized to foresee future gets to. Ordinarily, Web Usage Mining can be utilized to create legitimate prefetching and storing methodologies in order to decrease the server reaction time.

C) Support to the Design :

Ease of use is one of the significant issues in the plan and usage of sites. The outcomes delivered by Web Usage Mining procedures can give rules to enhancing the outline of web applications. Versatile Web locales speak to a further advance. For this situation, the substance and the structure of the site can be progressively revamped by the information mined from the clients' conduct.

D) E-commerce :

Mining business insight from web utilization information is significantly essential for online business electronic organizations. Web use mining systems can likewise be valuable in Customer Relationship Management (CRM). The issues particular to business, for example, client fascination, client maintenance, cross deals, and client flight are chiefly in the center.

CONCLUSION

This paper has given an overview of the expediently expanding research territory 'Web Usage Mining'. With the hazardous development of online applications around the world, especially in electronic business, there is a requirement for investigation of web utilization information, for example, get to log document and other client's data to remove learning to better comprehend client's conduct, and afterward apply this extricated information to better serve the necessities of clients. Preprocessing is the extremely urgent stage in web utilization digging for expelling unimportant data from the huge informational index. From this review, it is clear that the proper application of web utilization mining methods and tools largely impact organizations success by analyzing the usage information from their websites which help them to produce productive information appropriate to their business objectives and goals. Also web developers are benefited in developing their websites with enhanced site usability and accessibility. In this research paper we tried to provide a clear understanding of the data preparation and knowledge discovery process.

REFERENCES

[1] Monika Dhandi, Rajesh Kumar Chakrawarti, "A Comprehensive Study of Web Usage Mining" in 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 978-1-5090-0669- 4/16/$31.00 © 2016 IEEE.

[2] Abdullah Gok, Alec Waterworth, Philip Shapira, "Use of web mining in studying innovation" in Scientometrics(2015) 102:653-671 Springer.

[3] N.Pushpalatha, S. Sai Satyanarayana Reddy, "Towards an extensible web usage mining framework for actionable knowledge" in ICICCT , 2017 IEEE.

[4] M. Aldekhail, "Application and Significance of Web Usage Mining in the 21st Century: A Literature Review" in International Journal of Computer Theory and Engineering, Vol. 8, No. 1, February 2016

[5] Neeraj Kandpal, Prof. Ripu Ranjan Sinha, M. S. Shekhawat, "A Study of Processes Involved in Web Usage Mining," in International Journal of Allied Practice, Research and Review, Vol. III, Issue XI, p.n.01-05, Dec, 2016.

[6] V. David Martin, Dr. T. N. Ravi, and "A Literature Survey on Web Content Mining," in International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 4, Issue: 10, October 2016.

[7] Shyam Nandan Kumar, "World Towards Advance Web Mining: A Review" in American Journal of Systems and Software, 2015, Vol. 3, No. 2, 44-6.

[8] Chaitra L Mugali AyeshaAzeema Maniyar Asst. Prof. Padma Dandannavar, "Pre-Processing and Analysis of Web Server Logs" in International Journal of Innovative Research in Advanced Engineering (IJIRAE), ISSN: 2349-2163, Issue 8, Volume 2 (August 2015).

[9] Mehak, Mukesh Kumar, "Web Usage Mining: An Analysis," in Journal of Emerging Technologies in Web Intelligence, Vol. 5, No. 3, August 2013.

[10] Aditi Shriva7stava, Nitin Shukla, "Extracting Knowledge from User Access Logs, " International Journal of Scientific and Research Publications, Volume 2, Issue 4, April 2012.