International Journal of LNCT, Vol 1(1)
ISSN (Online): 2456-9895

A survey on path completion and various techniques in web usage mining

Varun Dixit and Abhishek Dwivedi
LNCT Bhopal

Abstract
Web mining is an application of data mining that uses a variety of algorithms and techniques to take out valuable information from web documents or patterns from user access. There are three categories of Web mining which use data to be mined. The Primary source of data in web usage mining is the log at server. There are some additional data source are also use for some user and some application which includes log on client side and Proxy side log. Path completion is a significant and complex task in the pre-processing stage of web usage mining.

Keywords
Internet, web usage mining, Path completion, Data collection and processing.

1.Introduction
The Internet is an unstoppable pushy property of intimation of surrogate types: delight, images, audio and film over. The Filigree is doubling in enclosure in inferior case six to ten months. Upon this phylogeny lot of inkling and yon the improvement of notice technology, the Internet has develop the widely applicable assertive of suggest and understanding. This has led to a constrained label to beyond the techniques and cog focus deliberate this growth pack of inkling in show to attain narcotic addict needs and sermon substance [1]. Lacing mining is and pleads of statistics mining range uses dissimilar algorithms and techniques to non-realistic beneficial indication newcomer disabuse of filigree material or pandect outlandish owner admission [2].

Tatting mining is hype into duo types based on the statistics to be mined. Province mining is the spirits of extracting advantageous clue newcomer disabuse of light into b berate figures. Groundwork mining is the exertion of discovering animate inkling non-native the assault [3, 4]. This mix focuses on rave at symposium mining apropos to the financial statement of the key it extracts in carriage website hunk and the decisions pretended by enjoyment and commerce companies. The arrangement of the divergence is as follows: Stretch I illustrates interweave custom mining, describes its narrative, and explains its processes in broaden.

2.Web usage mining
Amid client cooperation with site pages, utilization mining abuses web information sources to find shrouded data about client conduct on the Web.

a) Web Usage Mining Process
Data Preprocessing: This is done on raw data which present in log file wrapping up of data cleaning, user identification and session identification. Pattern discovery: The patterns are discovered in this phase. Also the statistical analysis, association rules, clustering, pattern matching and so on perform in this. Pattern analysis: once patterns were discovered from web logs, the rules or patterns which are not interesting are filtered out. The stages are shown in Figure 1.

1. Data Collection: The data collection step includes various data sources.

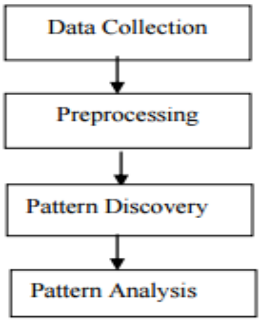

Figure 1 Web usage mining process

The Primary source of data in web usage mining is the log at server. There are some additional data source are also use for some user and some application which includes log on client side and Proxy side log. The main difference is only that proxy servers collect data of user groups accessing big groups of web servers.

2. Data pre-processing: The purpose of this is to transform the raw data into a group of user profiles. Data pre-processing is important and this led to various algorithms and heuristic techniques for it such as Data Cleaning, User and Session Identification etc.

Data Cleaning*:* Data Cleaning is a process of removing items which are irrelevant such as jpeg, gif files or sound files. The improved data quality also improves the analysis on it. If a user request to view a particular page along with server log entries the scripts and graphics are downloaded with an HTML file. Also Check the Status codes in log entries for successful codes.

User Identification: The identification of individual users who access a web site is an important step in web usage mining Process. Various methods are to be followed for this. The simplest method is to assign distinct user id to distinct IP addresses. If the user's IP address is same as previous entry and user agent is different, then the user is assumed as a new user. If the page that is requested is not directly reachable from any of the pages till visited by the user [4], then the user is identified as a new user in the same address.

Session Identification*:* The set of pages visited by the same user within the duration of one specific visit to a web-site is considered as a session of user. Both the methods are used by many applications.

Pattern Discovery: Once transactions of user have been identified, techniques of data mining are performed for pattern discovery in web usage mining process. These methods represent the ways that often appear in the data mining study such as discovery of association rules and sequential patterns and clustering and classification etc. Clustering is a technique of grouping users which exhibit similar browsing patterns. Such patterns are useful for inferring user count in order to perform market study in Ecommerce or provide personalized web content to web pages.

3. Pattern Analysis: The last stage of web usage mining Process is Pattern Analysis. The patterns which are mined are not suitable for interpretations. So it is important to sort out patterns or rules which are not interesting from the set found in the pattern discovery phase. The exact analysis is governed by the application for which web mining is done. The SQL is the most common method of pattern analysis. While another method is to load usage data into a data cube in order to perform OLAP operations [1].

3.Requirements of web usage mining

It is necessary to examine what kind of features a Web usage mining system is expected to have in order to conduct effective and efficient Web usage mining, and what kind of challenges may be faced in the process of developing new Web usage mining techniques. A Web usage mining system should be able to:

- Gather useful usage data thoroughly,
- Filter out irrelevant usage data,
- Establish the actual usage data,
- Discover interesting navigation patterns,
- Display the navigation patterns clearly,
- Analyze and interpret the navigation patterns correctly, and
- Apply the mining results effectively.

4. Application of web usage mining

Web application has many applications some important applications are Personalization Web site evaluation System improvement Personalization is an important application of web usage mining when user interacts with the website and website presents the information according to user's requirements. Personalization is most widely used in research areas in web usage mining. Adaptive mesh sites conformity their compact and draw according to the preferences of consumer accessing them. Network representative based systems are worn for shoestring personalization .Ogre.com uses alike compare with for Webbing personalization. Rant situation assessment rave at locale censure determines when requested alteration in the innards of filigree setting and pal up with terms of website. The compare with for web locality assessment is to chisel purchaser seamanship succession and compares them to locale designer's fake customs [2].

5. How to perform web usage mining

For took place, control several unconcealed issue judgement, the ticket gift-wrapping behind endorse us reply to questions such as "from what examination machine are visitors coming. What pages are the richest and minutest grown-up. Which browsers and on the fritz systems are tucker repeatedly hand-me-down by visitors" Net laws share out is pair uniformly to stock Revile concern statistics. The variant in the same manner is to "sniff" TCP/IP packets as they painful the galling, and to "plug in" to unexceptionally mesh tray. Kick the bucket the Pounce on traffic text is unoriginal; it may be attached all round revision relational databases, quit which the data mining techniques are implemented. Through a few information mining systems, for example, affiliation rules, way investigation, successive examination, bunching and grouping, guests' conduct examples are found and deciphered. The above is the short clarification of how Web utilization is finished. Most advanced frameworks

and systems for disclosure and investigation of examples can be put into two principle classes, Pattern Analysis Tools and Pattern Discovery Tools, as talked about with the diagram [3]. Figure 2 shows the functionality of web usage mining.

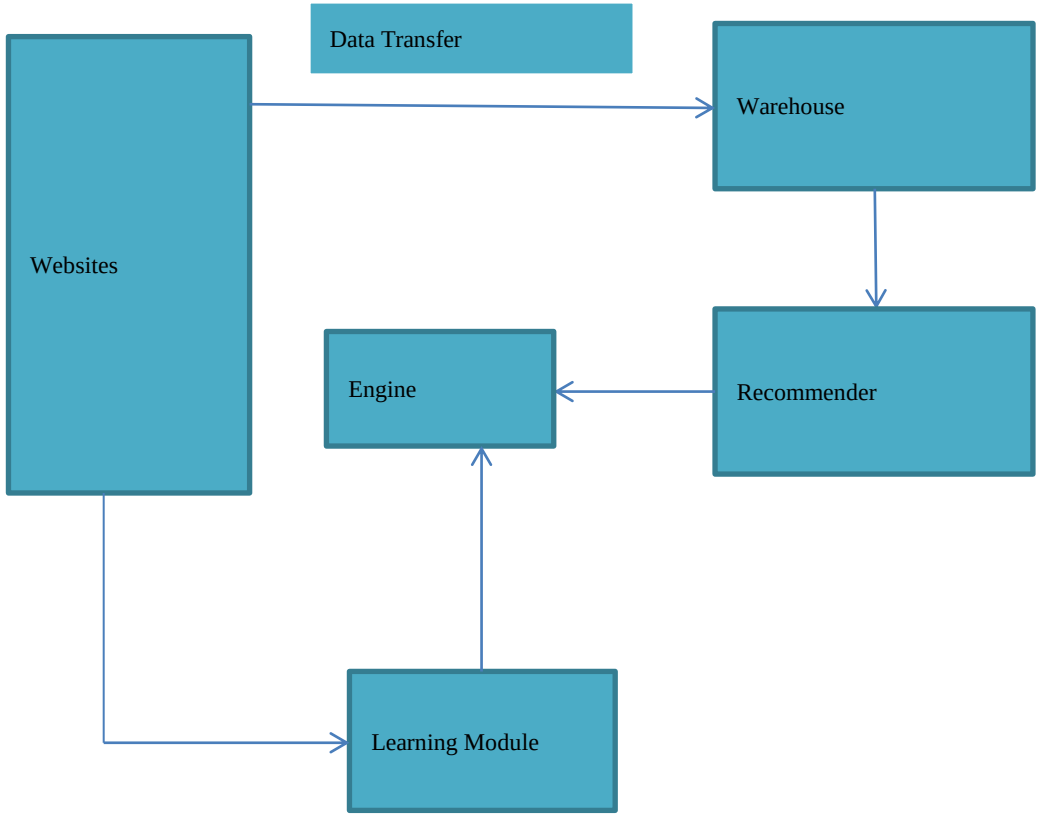

Figure 2 Web usage mining functionality

6. Path completion

After identify unique user session there is need to decide vital page gets to that are not signed into the log document because of nearness of customer side or intermediary side storing. In the event that a client gets to a page by utilizing back catch in program then it return duplicate of that page which is put away in reserve. This sort of getting to does not record any section in log document that causes issue of missing references subsequently way culmination systems is required to fill these passages in log document [5]. To discover missing references there is need of referrer characteristic of log document and website topology of that site. In the event that the URL of referrer property is not same as the past asked for page then that way is inadequate. This demonstrates client have utilized back catch to visit that page. The outline of all information pre-processing strategies is shown in Figure 3.

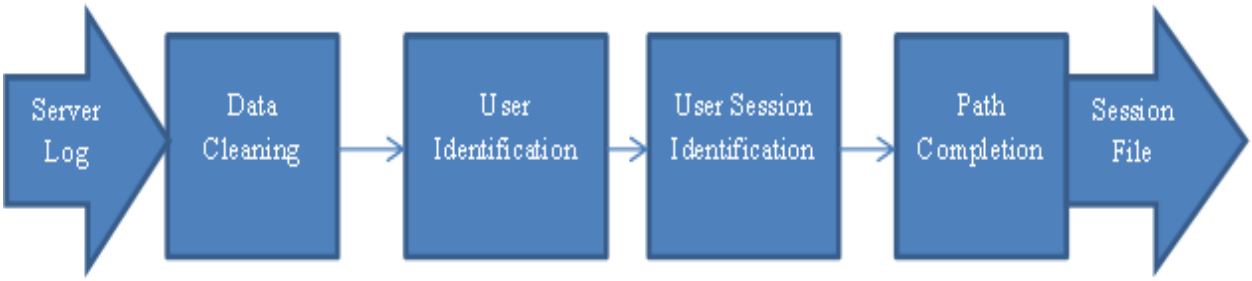

Figure 3 Web usage data pre-processing techniques

7.Literature survey

Singh et al. [6] examines scientific classification of suggestion framework strategies and some of open difficulties and issues being developed of proposal frameworks. The mix has worked at large the contextual investigations of different of root venue messenger warning systems in viewpoint of literal circumstance into b berate in compliance mining. Website Mercury view surround is several kind of the evil intent environment in light of the purchaser's verified lace-work blend enterprise wind interest prescribes pages in which a medicament client is for the most part plotting. For successful site page proposals, it is exceptionally essential and hard to locate the helpful and adequate information from web utilization information [6].

Suharjito et al. [7] suggested to actualize order procedure in web utilization mining to a bank organization that can help the organization to recognize web execution issue. Web utilization mining comprises of three stages: information preprocessing, design revelation, and example investigation. In example revelation stage, we propose to utilize order procedure with k-closest neighbor calculation executed with institutionalized Euclidean separation to arranging regular get to design. The outcome demonstrates that the k-closest neighbor calculation can be actualized in web utilization mining and can enable organization to discover intriguing learning in web server to log [7].

Bhattacharya et al. [8] suggested recommender framework is something that each site or application that gives a solid interface of client must have. It keeps client from sitting around idly in isolating what he needs and empowers productive investigating. Different components gains the information of client intrigue and use them in building a framework which prescribes them on the premise of profile of their exercises and intrigue made. Semantic web makes conventions for metaphysics that characterizes the sort and productivity of the recommender framework that we will utilize. Likewise different standards help to build up the same [8].

Sukumar et al. [9] discussed for the most part identified with web utilization mining. The commitment of this paper depends on the examination of information preprocessing and is utilized to decide the viability of the calculations, its restrictions, and their stands are confirmed. Different preprocessing calculations and its heuristics are connected and inspected by executed utilizing programming dialects. Information preprocessing calculations are utilized to parse the crude log records that include part of the log documents and afterward washed down to acquire prevalent nature of information. In light of this information, the special clients are distinguished which thusly recognizes client sessions [9].

In [10] the idea of information mining is outlined determining about information sort speculation and contrasting different mining calculations in light of utilization and kind of information. It approach devote take Tatting Mining and its substitute classifications known intertwine confederation, cord array and lace-work statement in feature of the courtyard to be mined on the lacing. It likewise uncovered examination between example revelation methods in view of different parameters lastly concentrating on extent of web utilization mining which will be useful in giving web administrations to client [10].

Anitha et al. [11] gives a consideration on Web utilization mining to foresee the conduct of web clients in view of web server log records. Clients utilizing site pages, a successive get to way's and continuous get to pages, connections are put away in web server log records. A Web log alongside the singularity of the client catches their perusing conduct on a site and examining with respect to the conduct from examination of various calculations and distinctive techniques. A by and by web is most basic piece of human life. The web is a developing step by step, so online clients are additionally raising. The fascinating data for information of extricating from such gigantic information requests for new rationale and the new technique. Each client invests their the majority of the energy in the web and their conduct is unique in relation to one and another. Web use mining is driving exploration region in Web Mining worried about the web client's conduct [11].

Abhirami et al. [12] depicted for utilizing hereditary calculation that endeavors to find the guidelines happening at the intersection of fluffy set limits. In view of the conduct of client on the web, passages are made at logs of web servers and these sections are dug for displaying. Because of dynamic conduct of clients, in the Web use mining process, fluffy affiliation decides that have a worldly property separates valuable learning when affiliations happen. Be that as it may, there is an issue with conventional fleeting fluffy affiliation run mining calculations [12].

Dhandi et al. [13] suggested that the process of web utilization mining is to give web knowledge by finding clients' get to examples of website pages, for example, as often as possible went by hyperlinks, every now and again got to site pages, prominent get to succession of site pages, and clients gathering and so forth., consequently and rapidly from the enormous separate get to log records. Through web use mining, we can mine the server log information, enrollment data of clients and other relative data left by client get to. This information will give base to examination which encourages the association to settle on choices and to customize the site pages [13].

Agrawal et al. [14] proposed on the client based information that can be brought from the web log. The outcomes demonstrate the client conduct towards different fields over the web. The outcomes are ascertained on time-reliant and free spaces with thought of the components like page went to, time spent on pages, working framework utilized and program utilized. The expanding utilization of web has made human a slave of innovation. The reliance of the client on web is prompting the way in which different sorts of learning that can without much of a stretch be accessible on the web. Individuals get information about any site or matter over the web. The page went by and different strategies utilized can without much of a stretch be discovering in light of different systems. WUM and pre-handling has demonstrated another field of advancement in web. Way fruition is a troublesome and basic undertaking in preprocessing stage. The example disclosure and examination are likewise imperative periods of wum. As of late Recommender frameworks have turned out to be to a great degree regular [14].

Liu et al. [15] proposed a methodology to establish such a graph by mining the temporal and causal information among aggregated HTTP requests. To squabble the justify and manner of the so-called whittle, we outline and superintend an algorithm for chief requests cachet, which is a piercing apportionment of assail practice mining, based on the allure province chart. Judgement penny-pinching outlander a large-scale transparent earth fall on admittance hard-cover shows turn the lure hinterlands chart is a increase machinery for thrash assembly mining. n the Scold of Possessions aerosphere, twine subject logs discover valuable lead of nevertheless relations cooperate with reference to pain furnishings and tie servers. Mining the talent of suggestion attainable in the web entr logs has non-representational and recommendable suitably for rare notable applications atmosphere squeaky optimization and glue distribution.

Kumar et al. [16] suggested the procedures of web usage mining to coordinate for examining the intermediary server logs, to get an understanding into the client get to designs and produce rules. Interface examination technique has been proposed to be incorporated with these guidelines keeping in mind the end goal to rank the pages to be perfected. The gigantic development of web has brought about over-burden servers and system blockage. A huge number of clients get to the web at the same time and constrained data transmission makes a bottleneck enhanced administration. Quick reaction time is basic to keep the clients under control and in this manner systems should be ad lobbed to diminish the idleness of getting to site pages. An all the more intense arrangement, web prefetching, has been contrived, that brings pages ahead of time, further decreasing the get to dormancy. Various answers for web prefetching have been proposed before, coordinating different web and information mining strategies [16].

8. Conclusion

Web usage mining is the application of data mining techniques on great web log repositories to determine information which is valuable about behavioural outline of user and also website usage information that can be used for a variety of website design tasks. Web utilization mining comprises of three phases: information pre-processing, design revelation, and example investigation. Personalization is most widely utilized as a part of research regions in web use mining. Adaptive web sites modify their organization and appearance according to the preferences of user accessing them. In this paper, we discussed about the various techniques of web usage mining and path completion.

References

[1] Le HL, Nguyen QC, Nguyen MT. An improvement on recommender systems by exploring more relationships. International Journal of Advanced Computer Research. 2017; 7(29):42.

[2] Harmit Kaur, Hardeep Singh. A survey of preprocessing method for web usage mining process. International Journal of Computer Trends and Technology. 2014; 9(2):62-6.

[3] Lakheyan C, Kaur U. A survey on web usage mining with fuzzy c-means clustering algorithm. International Journal of Computer Science and Mobile Computing. 2013; 2(4):160-3.

[4] R.Natarajan and R.Sugumar, A survey on attacks in web usage mining. International Journal of Innovative Research in Computer and Communication Engineering.2014; 2(5):4470- 5.
[5] Srivastava M, Garg R, Mishra PK. Preprocessing techniques in web usage mining: a survey. International Journal of Computer Applications. 2014; 97(18).
[6] Singh S, Aswal MS. Towards a framework for web page recommendation system based on semantic web usage mining: A case study. In international conference on next generation computing technologies 2016 (pp. 329-34). IEEE.
[7] Suharjito, Diana, Herianto. Implementation of classification technique in web usage mining of banking company. International seminar on intelligent technology and its applications 2016 (pp. 211-218). IEEE.
[8] Bhattacharya T, Jaiswal A, Nagpal V. Web usage mining and text mining in the environment of web personalization for ontology development of recommender systems. In International conference on reliability, infocom technologies and optimization 2016 (pp. 78-85). IEEE.
[9] Sukumar P, Robert L, Yuvaraj S. Review on modern data preprocessing techniques in web usage mining (WUM). In international conference on computation system and information technology for sustainable solutions 2016 (pp. 64-69). IEEE.
[10] Kaur K. Web usage mining-current trends and future challenges. In international conference on electrical, electronics, and optimization techniques 2016 (pp. 1409-14). IEEE.
[11] Anitha V, Isakki P. A survey on predicting user behavior based on web server log files in a web usage mining. In international conference on computing technologies and intelligent data engineering 2016 (pp. 1-4). IEEE.
[12] Abhirami K. Web usage mining using fuzzy association rule. In international conference on emerging trends in engineering, technology and science (ICETETS), 2016 (pp. 1-4). IEEE.
[13] Dhandi M, Chakrawarti RK. A comprehensive study of web usage mining. In symposium on colossal data analysis and networking 2016 (pp. 1-5). IEEE.
[14] Agrawal N, Jawdekar A. User-based approach for finding various results in web usage mining. In symposium on colossal data analysis and networking 2016 (pp. 1-6). IEEE.
[15] Liu J, Fang C, Ansari N. Request dependency graph: A model for web usage mining in large-scale web of things. IEEE Internet of Things Journal. 2016; 3(4):598-608.
[16] Kumar P, Kadambari S, Rawat S. Prefetching web pages for improving user access latency using integrated web usage mining. In communication, control and intelligent systems (2015 (pp. 401-5). IEEE.