ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 3, April 2016

A Survey on Web Usage Mining– Tools

1. Dr.K. Prabha, Assistant Professor
2. T. Suganya, Ph.D Research scholar, Department of Computer Science
PeriyarUniversity PG Ext.Centre,Dharmapuri, TamilNadu, India.

Abstract- ***Web usage mining is a type of web mining which exploits data mining techniques to discover valuable information from navigations of web uses, application of data mining techniques to web worthwhile data in order to extract usage patterns. The performance of web information retrievals and web based data warehousing are boosted with the extraction of information from the web using web mining tools. And web usage mining is one of the fastest developing areas of web mining, thus this activity that involves automatic discovery of user access patterns from one or more web servers.This paper provides anexploration about web usage mining analysisof concepts, ranking, tools, and applications.***

Keywords:

Data mining,web server, web mining.

I INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the Web. According to analyzetargets, web mining can be divided into three different types, which are Web content mining, Web structure mining andWeb usage mining. The extraction of hidden predictive information from large databasesisa powerful newtechnology with great potential to help companies focus on the most important information in their data warehouses.Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-drivendecisions [12].

There are three general classes of information that can be discovered by web mining:

- **Class A**: Web activity, from server logs and Web browser activity tracking.
- **ClassB**: Web graph, from links between pages, people and other data.
- **Class C**: Web content, for the data found on Web pages and inside of documents.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. The three categories of web mining for the following:

1.1 WEB CONTENT MINING

Web content mining is an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine readable semantic, some approaches have suggested restructuring the document content in is presentation that could be exploited by machines. The usual approach to exploit known structure in documents is to use wrappers to map documents to some data model. Techniques using lexicons for content interpretation are yet to come. There are two groups of web content mining strategies: Those that first group is directly mine the content of documents and the second group improves on the content search of other tools like search engines.

1.2 WEB STRUCTURE MINING

World Wide Web can reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. The Page Rank and CLEVER methods take advantage of this information conveyed by the links to find pertinent web pages. By means of counters, higher levels cumulate the number of artifacts subsumed by the concepts they

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 3, April 2016

hold. Counters of hyperlinks, in and out documents, retrace the structure of the web artifacts summarized.

1.3 WEB USAGE MINING

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools existed but they are limited and usually unsatisfactory. We have designed a web log data mining tool, Weblog Miner, and proposed techniques for using data mining and On-line Analytical Processing (OLAP) on treated and transformed web access files.

Applying datamining techniques on access logs unveils interesting access patterns that can be used to restructure sites in a more efficient grouping, pinpoint effective advertising locations, and target specific users for specific selling ads. Customized usage tracking analyzes individual trends [11, 2]. Its purpose is to customize web sites to users. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. While it is encouraging and exciting to see the various potential applications of web log file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge one can discover from the large raw log data. Current web servers store limited information about the accesses. However, for an effective web usage mining, an important cleaning and data transformation step before analysis may be needed.

II RANKING IN WEB USAGE MINING (WUM)

The web usage mining generally includes the following several steps: data collection, data pretreatment[6] and knowledge discovery and pattern analysis.

2.1 DATA COLLECTION

The first step of web usage mining is the data collection. The works smoothly carrying on and the final recommendation of characteristics services quality is directly affect by the data authenticity and integrity. So the data collection is use the scientific, reasonable and advanced technology to gather various data. Right now, the web usage mining technology has the main three data origin kinds: server data, client data, and middle data(agent server data and package detecting).

2.2 DATA PREPROCESSING

Some databases are insufficient, inconsistent and including noise. The oneness transformation is carry on database by the data pre-treatment. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

2.2.1 DATA CLEANING

To eliminate the irrelevant items is purpose of data cleaning, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web accesslog will be eliminated during data cleaning. Since the target of Web Usage Mining is to get theuser's travel patterns, following **two kinds ofrecords are unnecessary and should be removed**

- **The records of graphics, videos and the formatinformation**. The records have filename suffixes ofGIF, JPEG, CSS, and so on, which can found in theURI field of the every record.
- **The records with the failed HTTP status code**. By examining the Status field of every record in theweb access log, the records with status codes over299 or under 200 are removed. It should be pointed out that different from mostother researches, records having value of POST orHEAD in the Method field are reserved in presentstudy

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 3, April 2016

for acquiring more accurate referrerinformation.

2.2.2 USER AND SESSION IDENTIFICATION

The task of user and session identification is to find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The session identification goal is to divide the page accesses of each user data time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows:

a) The different IP addresses distinguish different users;

b) If the IP addresses are same, the different browsers and operation systems indicate different users;

c) If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty;

d) The session identified by rule 3 may containsmore than one visit by the same user at different time, the time-oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

2.2.3 PATH COMPLETION

In data preprocessing theanothercritical step is path completion. The result in path's incompletion are of some reasons, for instance, local cache, agent cache, "post" technique and browser's back button can result in some important accesses are not recorded in access log file, and the number of URL recorded in log may be less than the original or real one. Christo Ananth et al. [11] discussed about a method, Sensor network consists of low cost battery powered nodes which is limited in power. Hence power efficient methods are needed for data gathering and aggregation in order to achieve prolonged network life. However, there are several energy efficient routing protocols in the literature; quiet of them are centralized approaches, that is low energy conservation. This paper presents a new energy efficient routing scheme for data gathering that combine the property of minimum spanning tree and shortest path tree-based on routing schemes. The efficient routing approach used here is Localized Power-Efficient Data Aggregation Protocols (L-PEDAPs) which is robust and localized. This is based on powerful localized structure, local minimum spanning tree (LMST). The actual routing tree is constructed over this topology. There is also a solution involved for route maintenance procedures that will be executed when a sensor node fails or a new node is added to the network.

Through data pre-processing, web log can be transformed into another data structure, which is easy to be mined.

2.3 PATTERN DISCOVERY

To carry the analysis, use statistical method and mine the pre-treated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

2.4 PATTERN ANALYSIS

A challenge of Pattern Analysis is to filter uninteresting information and to visualize and take to mean the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse. Next use technology of OLAP and so on to carry on the comprehensive mining and analysis. Finally, the pattern analysis, provide the characteristic service to the electronic commerce website.

III WEB USAGE MINING TOOLS

In the past, several WUM projects have beenproposed to for the users' preference and theirnavigation behavior, as well as many recent resultsimproved separately the quality [7] of the personalizationor the user profiling phase.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 3, April 2016

3.1 SPEED TRACER

In Speed Tracer, a usage mining and analysis tool, isdescribed. Its goal is to understand the surfingbehavior of users. The analysis isdone by exploring the server log entries. The speed tracer has a main characteristics is does not require cookies or user registration for session identification. It uses five kind ofinformation: IP, Timestamp, URL of the requestedpage, Referral, and Agent to identify user sessions.Advanced mining algorithms, uncover user's movement through a Web site. The final result is acollection of valuable browsing patterns which helpwebmasters better understand user's behavior.Speed Tracer generates three types of statistics:

- **User-basedstatistics** pinpoint reference counts and durations ofaccesses.
- **Path-based statistics** identify frequenttraversal paths in Web presentations.
- **Group-basedstatistics** provide information on groups of Web sitepages most frequently visited.

3.2 SUGGEST 3.0

SUGGEST 3.0 is a recommender system aimed toprovide the users with useful information about pagesthey may find of their interest. The personalization isachieved by means of a set of dynamically generated page links. Differently from the recommender systems proposed so far, SUGGEST 3.0 does not make use of any off-line component, and is able to manage Web sites made up of pages dynamically generated. To this purpose SUGGEST 3.0 incrementally builds and maintains historical information by means of an incremental graph partitioning algorithm, requiring no off-line component.

3.3 COLLABORATIVE FILTERING

Collaborative filtering is the most successful technology for building personalized recommendation system and is extensively used in many fields. Most web personalization recommendation system adopts two types of techniques: a content-based approach and a collaborative filtering (CF) approach. The content based approach, recommends web objects that are similar to what the user has been interested in the past. The collaborative filtering approach, finds other users that have shown similar tendency to the given users and recommends what they have liked. The collaborative filtering recommendation acts according to other users' viewpoint to produce recommendation tabulates to the goal user. Its basic thought is based on a supposition: If user grade to some product quite similarly, then they grade to other product also similar.

3.4. WEB PERSONALIZER

A list of hyperlinks to users in system is provided in dynamic recommendation. Analysis is based on the usage data combined with structured data formed by hyperlinks of site. Aggregated usage profile isobtained by applying data mining technology inpreprocessing phase. Theactive user session in order to find match among useractivities and discover usage profile in the online phase. Matchingentries are then used to compute a set ofrecommendations which will be inserted into lastrequested page as a list of hypertext links. Thissystem helps in personalization.

3.5 CLASSIFYING USER NAVIGATION PATTERNS USING LONGEST COMMON SUBSEQUENCE ALGORITHM

Prediction of user future movements and intentionsbased on the users' click stream data is a mainchallenging problem in Web based recommendationsystems. Web usage mining based on the users'click stream data has become the subject ofexhaustive research, as its potential for web basedpersonalized services, predicting user near futureintentions, adaptive Web sites and customer profiling is recognized. This model provides online prediction through web usage mining system and proposes a novel approach for classifying user navigation patterns to predict users' future intentions. The approach is based on using longest common subsequence algorithm to classify current user activities to predict user next Movement.

IV APPLICATIONS OF WEB USAGE MINING

4. 1 E-COMMERCE

E-Commerce means two trading parties based on Internet according to certain rules or

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 3, April 2016

standard developing the whole traditional business activity in digital network mode. Services through internetor the buying and selling of products,

E-Commerce generates a huge volume of interactions. This tremendous growth in the E-commerce enterprise, twisted to product surplus.It also supports e-commerce sites to retain the most profitablecustomers [4], improve the functionality of web based applications, provides more custom-made content to visitors. In addition, with the use of Web usage mining techniques e-commerce companies can improve products quality or sales by anticipating problems before they occur. They also provide companies with previously unknownbuying patterns and behavior of their online customers. More importantly, the fast feedback the companies obtainby using Web usage mining is very helpful in increasing the company's benefit [5].

4.2 E-LEARNING

E-learning is a form of electronically supported learning which allows the people to learn any subject at anytime andanywhere. The simplicity in using the tools to browse the resources on the web, its easiness in deploying and maintaining resources made the web as an excellent tool for delivering courses. Web is one and only major choice to manage and maintain learning resources and has become one of the leading choice of modern advanced distanceeducation system. As education becomes more technologically advanced, the complexity of available learning resources also increased accordingly. It is difficult to evaluate the structure of the course content and its effectiveness on the learning process.The pattern analysis capability of web usage mining has an important role in web-based learning system. They can analyze the students and instructors behavior [10] and improve the educational experience. Tracking the activities happening in the course website and mine patterns is also beneficial to improve or adapt the course contents. This allows instructors to appraise the access behavior, assess the learning activities and compare learners. The arrangement of the course contents can be enhanced by analyzing the traversal paths of the course content web pages is another advantage of Web usage mining [1].

4.3 E-GOVERNANCE

E-governance provides a single web portal that integrates all services that includes government, nonprofit and private-sector entities [12]. In such a type of service system which provides ready access to information, the user interface quality is an important factor. This is one of the challenging user-centric parameter since this has to provide information to extensive and various users [3].The patterns of the online behavior of the users can be discovered by using Web usage mining techniques. These patterns reveal the user interests and that can be utilized to fine tune user interfaces and suggest the most appropriate browsing paths. User requirements also are exhibiting in their navigation behavior. Analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalized system. This technology also uses the experience of users of past sessions to provide recommendation to users of current session [10].

V CONCLUSION

Web usage mining is a kind of mining to server logs. Web usage mining plays an important role in realizing, enhancing the usability of the website design, the improvement of customer's relations and improving the requirement of system performance and so on. Web usage mining provides the support for the website design, providing personalization server and other business making decision, etc. Web usage mining can model user behavior and therefore to forecast their future whereabouts. The main aim of web usage mining is to get useful to users for easy access information in logs to make sites perfect with effectual.

REFERENCES

[1] Bart C Palmer; Web Usage Mining: Application to an onlineeducational digital library service; Digital Commons@USU; 2012

[2]GovindMurariUpadhyay, KanikaDhingra, "Web Content Mining: Its Techniques and Uses", IJARCSSE, November, 2013.

[3]G.Rani; S .Chakraverty, ―Boosting Interactivity of EGovernance‖, International Conference on communication Languages and SignalProcessing- with Preference to 4 G Technologies‖,ICCLSP4G,January2012.

[4]http://revistaie.ase.ro/content/51/104%20-20 SIVARAMAKRISHNAN,%20 BALAKRISHNAN.pdf

[5] J. g. Liu, h. h. Huang. Web Ming for Electronic Business Application, Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies, Chengdu, China, 2003:872~876.

[6]Margaret H. Dunham "Data Mining Introductory and Advanced Topics ISBN 978-81-7758-785-2 pp 205-218

[7]Pranit Bari, P.M.Chawan "Web Usage Mining", Journal of Engineering, Computers & Applied Sciences (JEC&AS) Volume-2 no.6, June 2013, pp 37-38. ISSN NO:2319-5606.

[8]Romero C. Ventura S, Pechenizky M , Baker R. S ; Handbook of educational data mining; 2010; CRC Press.

[9] S. Chakraverty 1 , B. G. Rani, C. B. Singla and D. Anand; Experience based recommendations system for e-governance; 2012.

[10]XiaoqingZheng,YilingGu,YinshengLi,"DataExtraction from Web Pages Based on Structural Semantic Entropy",International World Wide Web conference committee (IW3C2),April 2012.

[11]Christo Ananth, S.Mathu Muhila, N.Priyadharshini, G.Sudha, P.Venkateswari, H.Vishali, "A New Energy Efficient Routing Scheme for Data Gathering ",International Journal Of Advanced Research Trends In Engineering And Technology (IJARTET), Vol. 2, Issue 10, October 2015), pp: 1-4

[12]ZakareyaEbrahim and ZahirIrani; —E-government adoption: architecture and barriers‖ Emerald Business Process Management journal, vol.II, No.5 2005, pp589-611, 2005