ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 11 , November 2015

A Process of Web Usage Mining and Its Tools

G.D.Praveenkumar,R.Gayathri

Asst.Professor,Department of computer science, Bharathiar university arts and science college, Gudalur.
Asst.Professor,Department of CA&IT, Kaamadhenu arts and science college,sathy.

ABSTRACT: Now a days most of people are searching data in internet, to search a user needed data from different search engines ,that search data forming a queries format then only big data retrieve from the data base. The big data is the term used to describe the exponential growth and availability of data both structure and unstructured data. This paper deals with log file , overview of web mining process, various tools for web usage mining, filtering process, merits and demerits of web usage mining.

KEYWORDS: Data mining, web mining, web usage mining tools.

I.INTRODUCTION

A. DATA MINING

The data mining is one of the fastest growing fields in computer industry. Data mining can involve the uses of different kinds of software packages such as analysis tool .data mining can automate the process of finding relationship and pattern in raw data and reset either utilized in an automated decision support system or accessed by a human analysis[6]. Data mining can be used in science and business area to analyze a large amount of data to discover trends which they could not otherwise found.

B.WEB LOG FILES

Web log file are files that contain information about website visitors activity,log files are created by web server automatically, each time a visitors request any file from the site information on his request is appended to a current log file. Most of the log file have text format and each log entry is saved as on a line of text[10].

B.1) Content of a web log files

- Username
- Visiting path
- Path traversed
- Time stamp
- Page last visited
- Success rate
- User agent
- URL
- Request type

B.2) Data source for web usage mining

Data source include server log, client log, proxy server log
1) Server log:When a person request a particular page in web ,when ever an entry is logged into a special file is called server log
2) Client log:It is also known as web server log. Web log can also be gathered from user machine by interacting java applet to the web sites, writing java script or even modified browser.
3) Proxy server:It is a server which act as an inter less on between the user request to the other web server.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 2, Issue 11 , November 2015

Web log information can be integrated with web content and web structure mining , when user access a particular page entry is entered in web log server. The interaction details use with web site are recorded automatically in web server as the form of web logs. Web logs is in form of line of text in web server, proxy server and browser.

B.3) Types of web server logs:

Web server logs are plain text files and are independent from the server .there are four types of server logs
1) Access log file:Data of all incoming request and information about client of server. Access log records all requests that are processed by the server.
2) Agent log: Information about user browser versions
3)Error log:List of internal error ,when ever an error is occurred ,the page is being requested by client to web server the entry is begin requested by client to web server the entry is made to error log
4) Referrer log: This file provides information about link and redirect visitors to site

B.4) Weblog file format

Web log file is a simple plain text file which record information about each users, display of log file data in three different format [7]

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

1) Application server log: It contains information of user activities like IP address, request , source.
2) Application level logs: It maintain information of user at application level like number of hits on web page old reference and new reference

C. FILTERING METHODS

Pattern analysis mechanism used for some filtering approaches

- Rule based filtering : provide content to user based on predefined rules
- Collaborative filtering :give recommendation to user based on ratings of similar user
- Content based filtering: track which page user visits and recommended other pages with similar element.
- Hybrid method: usually a combination of content based collaborative
- Site filter: this technique is used in web miner system. The filter uses the site topology to filter out rules and patterns that are not interesting. Any where that conforms direct hyper link between pages is filtered out.

II.WEB MINING

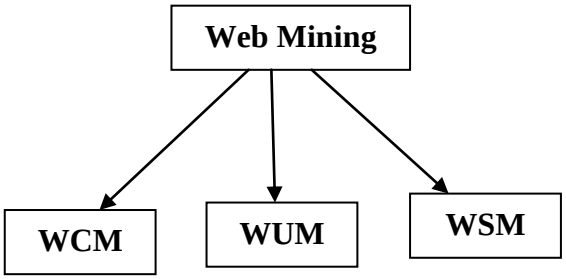
Web page created with xml and html. The World Wide Web has become a very popular medium publishing. The web is rich information on gathering and making a sense of this data is difficult because publication on the web is unorganized. Using a web browser or search engine user needed data to find specific information on the web. The user needed data is convert to query format in web browser or search engine or retrieve a big data from data base these big data may be structure or semi structure[11] .

Search engine have become the most helpful tool for web obtaining useful information from the internet. The search result returned by even the most popular search engine not satisfactory .The data on the web is dynamic .the implicit and explicit structure of data is updated frequently. The data on the web page is no rigid and noun form data structures or no schema as are not followed. This shows that the data on the web is unstructured.

- Resource finding: retrieving the document related to our search.
- Information selection: selecting an appropriate document from the displayed document
- Generalization : discovering patterns from individual web sites
- Analysis: checking the validity of the mined patterns.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 2, Issue 11 , November 2015

Web mining can be divided into three types, they are web content mining(WCM),web structure mining(WSM), web usage mining(WSM).



III.WEB MINING TYPES

A. WEB CONTENT MINING

In this web content mining process attempts to discover all the links of the hyper links in a document, so generate the structural report on a web page. There are two group of web content mining strategies, first strategy is to directly mine the content of document and the second one are those that improve on the content search of other tool like search engine the web content mining is mostly depend on text mining.

B.WEB STRUCTURE MINING [11]

Web structure mining consists of web pages and hyperlinks connecting related page. It analyses the structure of each page contained the website .it represent the structure of web pages and interest with each other. based on structural information is, further divided into two categorized

- Extraction of patterns from the hyper link on the net
- Mining of the structure of the document.

C.WEB USAGE MINING[12]

The web usage mining has emerged as the essential tool for realizing more personalized user friendly and business optional web services. Based on data logs of user interaction of the web including referring pages and user identification .the web log may be web server logs, proxy server logs, browser log. Data usage in web mining is web server log, site content .data about visitors gathered from external channels future application data. It is useful habit which can assist in organizing a web site so that high quality of service may be provided. A large part of web usage mining is about processing usage stream of data ,after that various data mining algorithm can be applied in three phases.

- Preprocessing
- Pattern discovery
- Pattern analysis

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 2, Issue 11 , November 2015

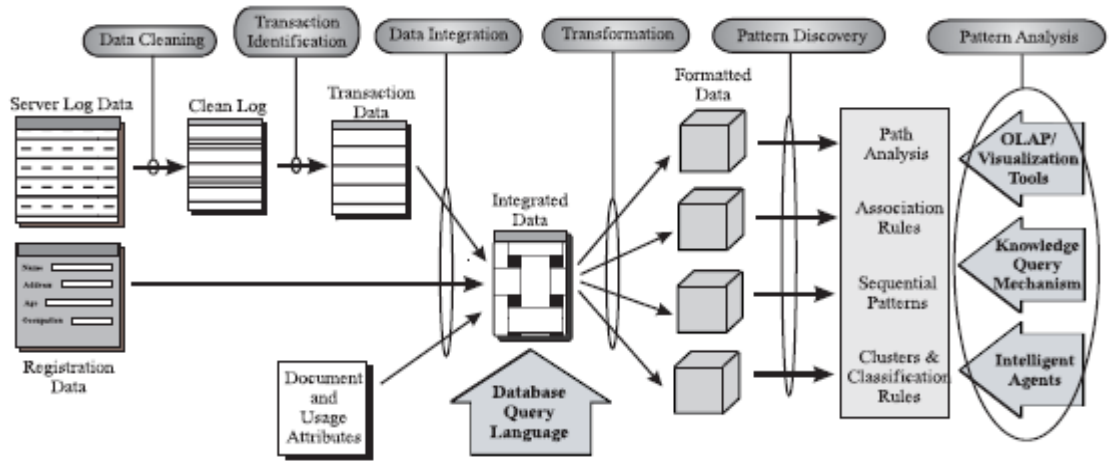


Fig 1. Overview of web usage mining

IV.WEB USAGE MINING PROCESS

A. PREPROCESSING

The preprocessing is an activity to reformatting the web log data before processing. These is also called cleaned data .the unwanted data are removed and minimized log file is obtained. The preprocessing has different sections like

- Data cleaning: Removes outliers and irrelative data
- Session identification: Divide all page accessed by a user into session.
- Data conversion: This is conversion of the log file data into format needed by the mining algorithms.

B.PATTERN DISCOVERY

Data mining techniques are applied to discover the interesting character tics in the hidden pattern. After preprocessing of log file into formatted data the pattern discovery process is undergone. The existing data of the log files may useful patterns are discovered with some files.

Pattern discovery have different sections like

- Path analysis
- Association rules
- Special patterns
- Cluster and classifications

C. PATTERN ANALYSIS

Pattern analysis is the final stage of web usage mining which can validate interested pattern from the output of pattern discovery that can be used to user behavior.

- Knowledge query mechanism: SQL is almost commonly used language for knowledge query mechanism. The language is applied in order to extract the useful patterns from discovered patterns.
- Visualization tool: OLAP provide an integrated frame work for analysis is which allows change in aggregate level.
- Intelligent agent: Various agent are also devised that helps in examining the pattern in web usage mining, these agent perform the work of analysis the discovered patterns.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 2, Issue 11 , November 2015

V. WEB USAGE MINING TOOLS

1) Web miner [2] : A general and flexible frame work for web usage mining , the application of data mining techniques, such as the discovery of association rules and sequential patterns, to extract relationship from data collected in large web data repositories. Restructure a web site ,and analyzing user access patterns to dynamically present information tailored to specific group of users.
2) Web log miner: It coined by zaiane(1998), use data mining and OLAP on treated and transformed web access files. Mining a web server log files.
3) Web mate:The user profile is inferred training examples, proxy agent provides effective browsing and searching help.
4) Web tools:It uses sequential pattern mining which relies on PSP algorithm developed by massegila for usage profiling
5) Web usage miner:It exploits an innovative aggregated storage representation for the information in the web server log. It discovers patterns comprised of not necessarily adjacent events. Mining interesting navigation patterns in form graphs
6) Data miner: It is a tool automating web data extraction and manipulation.
7) Koinotites: A system which uses data mining techniques for the construction of user communities on the web
8) i-miner:To optimize the concurrent architecture of fuzzy clustering algorithm and fuzzy inference system to analyze the trends, pattern discovery and trend analysis from web usage data mining
9) Web quit: web logging visualization system that heils web design team capture usage traces which can be aggregated and visualizes in a zooming interface that shows web pages can be viewed.
10) Speed tracer:reconstructs theuser transversal path for session identification by using referrer page and the url of the requested page as traversal step. Mining web server log files.

VI.APPLICATION OF WEB MINING

- E-commerce
- Information retrieval search on the network management
- Business intelligence
- Modification of web site.
- Marketing and fraud detection.

VII. MERITS AND DEMERITS OF WEB USAGE MINING

A. MERITS:

- Enabled e-commerce to do personalized marketing.[3]
- Ranking is based on combination of several factors
- Ranking is sensitive query[9]
- Query time cost is low
- Redundancy is exploited
- Add all prefixes of class path to the feature pools.

B. DEMERITS:

- Avoid irrelevant link
- Topic drift –document in based set may be too general[4]
- Provide more efficiency
- The classes of hyperlink neighbors and better representation of hyperlink

VII. CONCLUSION

The web usage mining have depend on different search engine, that search engine supported for view the data used user supportable web browser. The user needed information is dynamic, every day new content needed to be added as web page. This method is load to organize and increase the profit. It is now also used by search engine to

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 11 , November 2015

improve search quality and to evaluate search result and many other application implement. This web usage mining strongly shows the user goal identification or data. New algorithm and techniques should be developed scope of repository.

REFERENCES

[1].S.K.Madria,S.S.Bhowmick,W.K.Ng, and E.Lim, Research Issues in Web Data Mining, Data Warehousing and Knowledge Discovery,1990,303-312.

[2]D.shen,J.Sun, Q.Yang and Chen(2006)"building brigges for web query classification,'proc, 29thann int'1 acmsigirconf.Research and development in information retrial

[3]O.nasraoui, M.Soliman, E.Saka, A.Badia and R.Germain , A web usage mining frame work for mining evolving user profile for dynamic web sites, IEEE Transcation on knowledge and data engineering 20(2),2008,202-215.

[4] Chen H. andDumais S. (2000), 'Bringing Order to the Web: Automatically Categorizing Search Results', In Proc. SIGCHI Conference on Human Factors in Computing Systems (SIGCHI '00), Vol. 17, No. 3, pp. 145-152.

[5] Rajini Pamnani, Parmila Chawan : "web usage mining : a resw\earch area in web mining" department of computer technology VJIT university, Mumbai.

[6] Jones R., Rey B., Madani O. and Greiner W. (2006), 'Generating Query Substitutions', In Proc. 15thInt 'l Conference on World Wide Web (WWW '06), Vol. 62, No. 6, pp. 387-396.

[7] Lee U., Liu Z. and Cho J. (2005), 'Automatic Identification of User Goals in Web Search', In Proc. 14thInt 'l Conference on World Wide Web (WWW '05), Vol. 51, No. 3, pp. 391-400.

[8] M.Eirinaki and M.Vazirgiunnis "web mining for web personalization" ACM trans.Internet technology vol3, nov, 2003,pp 1-27.

[9] Wang X. and X Zhai C. (2007), 'Learn from Web Search Logs to Organize Search Results', In Proc. of 30th Annual Int 'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), Vol. 71, No. 5, pp. 87-94.

[10] Zheng Lu, HongyuanZha, Xiaokang Yang, Weiyao Lin and ZhaohuiZheng (2013), 'A New Algorithm for inferring User Search Goals with Feedback Sessions', IEEE Transaction on Knowledge and Data Engineering, Vol. 25, No. 3, pp. 502-522.

[11] S.R.SriAbirami, Dr.C.Nalini and A.P.Ponselvakumar (2015), 'Inferring the User Search Results Using Feedback Sessions and Evaluation Methods', In Proc. National Conference on Computational Intelligence and Fuzzy System (NCCIFS'15), Anna University, pp. 8-15.

[12] S.R.SriAbirami and A.P.Ponselvakumar (2015), 'Restructuring the User Search Results Using Feedback Sessions and Evaluation Methods', In Proc. National Conference on Intelligent Computing (NCIC 2015), Pondicherry Engineering College, pp.1240-1246.

AUTHOR'S BIOGRAPHY

G.D.Praveenkumar is currently working in Assistant Professor in Department of Computer science in Bharathiar University Arts and Science College, Gudalur. He has completed BCA in Kongu Arts and Science College, and MSC (IT) in Kongu Engineering, Erode. He has published 2 paper in National Conference and 1 paper published in international conference and 3 paper published in international journals. His areas of interest are Data Mining, Mobile Computing.