ISSN:0975-9646

Web Log Analyzer for Semantic Web Mining

Fakhrun Jamal , Mamta Bansal

Shobhit University Meerut India

***Abstract*: Web mining is the application of data mining techniques to extract the knowledge from web data, i.e. web content, web structure and web usage data. Website personalization is the process of customizing the content and structure of a website for specifically needs.**

Steps of personalization as

a) **The collection of web data**
b) **Modeling and categorization of these data.**
c) **Analysis the collected data**
d) **Determination of the actions that should be performed.**

Personalization of visitor's experience makes their time on your site or application more productive and engaging. Personalization is also valuable for any organization (business) such as increasing visitor's response or permoting customer retention. And Log analyzer will perform much better to provide the information to visitors. Web log analyzer will observe particular visitor and provide the accurate result and response to his request. Web Log Analyzer will embed to the website or server and one dummy file of it will store on the computer. Web log analyzer will make easy surfing and searching on internet to the visitor in fewer times. Website will also collect technical information, about size of computer screen and type of browser. This information will help web designers formatting website in a way, also collect the information related to your activity on the web Such as your IP address. This Technology will embed on the websites and also on your computer to collect the information about your activity.

***Keyword:* Web personalization, Cookies, Beacons, Google analytic, Web log analyzer.**

INTRODUCTION:

Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure and web usage data. Web personalization is the process of customizing the content and structure of a website for specifically needs. It involves application of data mining techniques on the contents of WWW but is not limited to it. Web site personalization can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs of each user taking advantage of the user's navigational behavior. Websites collect technical information about your computer, such as the size of your screen or type of the browser you use. This information helps web designer's format websites in a way that is websites also collect information related to your activity on the web, such as your internet protocol (IP) address. The time you clicked on a link, how much time you spent on a particular web page before moving onto the next one, and the web page you were reading prior to clicking the link. This information, when aggregated with similar information about other users, is extremely valuable to advertisers.

LITERATURE REVIEW:

A study, work was done on big data that starts from data mining to web mining to big data mining. The researcher reviewed the journey on how on big data was evolved. The size and variety of data pushed us to think ahead and develop new and faster methods of data mining which used the parallel computing capability of processors. She concluded this term as big data. She had also provided with the applications of different methods of mining. (Richa Gupta, April 2014).

Study deals with E-Commerce website to provide security for e-commerce websites. Researchers concluded web mining algorithms page rank and trust rank to develop web mining framework in e-commerce websites. They used four phases for web mining framework. Web structure mining used page rank and trusted rank algorithm. Web content mining used hierarchical clustering and k-means clustering algorithm. Decision analysis used trust calculation of web site application of suitable statistical technology. Finally security module provided security to web site. (M.Karthik, S.Swathi, 5 May2013)

The study focused on the efficient application of the web mining algorithm. Researchers proposed web page collection of web mining algorithm as the best performer to manage time and space complexity. They concluded the strategy for automatic web log information mining by Web Page Collection algorithm that was proved to be more effective. With the mined results, the web application was developed and provides adaptive user interface. Further the other type of explore in web applications will focus in the future work. It also included the information integration of content knowledge and knowledge extraction from the various web sites. (R.Shanthi, Dr.S.P.Rajagopalan, Sept 2013).

In the study researchers worked about the security when mining the web for information. They proposed Web mining combines two hot fields data mining and www together. They also proposed a closer view on the Web mining security, Website security against attacks specially (Denial of service) DoS attack. Attacker launched their attacks conveniently and attracting more users who were most susceptible to computer threats. They also proposed human element was responsible to maintain confidentiality, integrity and availability of organization's information resources. They concluded the ignorance of employees towards organizations security requirements lead to more security break-ins like (Denial of service) DoS attacks which affected the network bandwidth or connectivity. Thus they concentrated on the security against such an attack and methods to overcome it. The initial framework was described which was evaluated by security expert to

check its feasibility. (Ajay Jangra, BhumicaVerma, Feb 2012).

A study was carried out agent based frame work for semantic web contents employing clustering techniques. Researcher proposed agent based solution for mining semantic web content with the aim to provide context based knowledge oriented results to the user. She concluded web mining techniques with agent technology to improve performance, reduced network traffic, and better results. However, implementation of this work was still under progress and is left as future work. (Aarthi Singh, April 2012).

The study was carried out in which comparative statements of various page ranking algorithms with link editing, General Utility Mining and Topological frequency Utility Mining. Researchers also provided Model by taking constraints such as Web Mining activity, topology, Process, Weighting factor, Time complexity, and Limitations etc. This also helped in comparing WPs-Tree and WPs-I tree structures. They concluded the page ranking algorithms play a major role in making the user search Navigation easier in the results of a search engine, which helps in best utilization web resources by providing required information to the Navigator. They also concluded WPs-Tree and WPs-I tree provide better storage representations. The association between web pages could be found easily in an efficient way. This survey could be helpful for understanding various page ranking algorithms along with different storage representation to correlate web pages. As a future direction, the new metric could be developed which may be still better than this, so that users could have quick response, resources on the network could be used efficiently thus promoting green computing. (Prasad Reddy, Shashikumar G.Totad, Geeta R. Bharamagoudar, Sept – Oct 2012).

In this research study was carried out about the log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that could be used in the log files which in turn give way to an effective mining. Researcher proposed a detailed description of how the file is being processed in the case of web usage mining process. They conclude various mechanisms that perform each step in web mining the log file is being discussed along with their disadvantages. (L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai, January 2011).

This study carried out the Security and defense networks, proprietary research, intellectual property and data based market mechanism that depend on unimpeded and undistorted access call all be severely compromised by malicious intrusions. Researchers proposed data mining for counter terrorism and cyber security application. They concluded Web server logs are mostly captured the behavior of machine, not the behavior of end user, off line analysis of intersection of log files had allowed us to identify some host IP addresses that most probably belongs to intruders. They also concluded that the firewall was set in such a manner that those IP will be banned from accessing our network. Intersection of firewall log files coming from different machines could be a source for IP address that belongs to intruders. (Mahesh Malviya, Abhinav Jain, Nitesh Gupta, May 2011).

In this study researchers proposed the security and effectively manage the data from unauthorized access the data. They concluded a document might be represented in XML or RDF. We needed to apply privacy-preserving data mining for XML and RDF data. They also concluded Policy consistency could be determined using data mining techniques. Much work was necessary to prepare the way for the integration of the semantic web, data mining and security. Privacy- preserving information integration were fruitful areas of research in security informatics. (BhawaniThuraisingham, Latifur Khan, 2010 IEEE).

This study was concerned with the in-depth analysis of Web Log Data of NASA website to find information about a web site, top errors, potential visitors of the site etc. Which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web usage mining? Researcher concluded the user access log files of NASA Web server were analyzed to help system administrator and Web designer to arrange their system by determining occurred systems errors, corrupted and broken links. Similar studies could be done for any others web sites to increase their performances. Web usage patterns and data mining could be the basis for a great deal of future research. (K. R. Suneetha, Dr. R. Krishnamoorthi, April 2009).

In this study researchers proposed the powerful algorithm that mined the web log efficiently. They also described proposed algorithm firstly converted the web access data available in a special doubly linked tree. Each access was called event. Our algorithm was efficient than GSP (general sequence pattern). Researchers concluded for low support threshold and for large data base doubly linked mining tree performance was much better than GSP. While for higher support threshold and small size of data set science only few events quality the criteria of frequent event there was no significant difference in both the algorithms. (Ratnesh Kumar Jain, Dr. R.S. kasana, Dr. H.S. Gour, 2009).

This study carried out in which various data mining techniques that were successfully applied for cyber security. They proposed various data mining techniques including link analysis and association rule mining was explored to detect abnormal patterns. They concluded developing techniques that could dynamically adapt to new detection strategies and continue to monitor the adversary. They had developed to reduce false positive and false negatives. Furthermore, they were exploring the applicability of our techniques to distributed pervasive environment. (BhawaniThuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W.Hamlen, 2008 IEEE).

A study was carried out in which Apriori algorithm was used to mine frequent patterns. Researchers proposed a new pattern mining algorithm. And they concluded the following facts after implementing Apriori algorithm and new proposed algorithm. When the size of data increased the developed model performance poor result and some were provide more accurate result. In frequent set mining, our modify Apriori algorithms for search patterns and

building data models. Memory was used directly proportional to the size of data in both kinds of algorithms used in web mining. Accuracy was not much depends on the size of data it was most of the time depends upon type of data. (Birla, Sachin Patel, 2014).

In this study researchers presented a survey of the use of Web mining for Web personalization. More specifically, they introduce the modules that comprise a Web personalization system, emphasizing on the Web usage mining module. A review of the most common methods that are used as well as technical issues that occur is given, along with a brief overview of the most popular tools and applications available from software vendors. Moreover, the most important research initiatives in the Web usage mining and personalization area are presented. The researchers proposed that Web personalization is the process of customizing the content and the structure of a Web site to the specific and individual needs of each user, without requiring from them to ask for it explicitly. This can be achieved by taking advantage of the user's navigational behavior, as it can be revealed through the processing of the Web usage logs, as well as the user's characteristics and interests. They also include the overall process of Web personalization consists of five modules, namely: user profiling, log analysis and Web usage mining, information acquisition, content management and Web site publishing. The main component of a Web personalization system is the usage miner. Log analysis and Web usage mining is the procedure where the information stored in the Web server logs is processed by applying statistical and data mining techniques, such as clustering, association rules discovery, classification and sequential pattern discovery, in order to reveal useful patterns that can be further analyzed. Such patterns differ according to the method and the input data used, and can be user and page clusters, usage patterns and correlations between user groups and Web pages. Those patterns can then be stored in a database or a data cube and query mechanisms or OLAP operations can be performed in combination with visualization techniques. The most important phase of Web usage mining is data filtering and pre-processing. In that phase, Web log data should be cleaned or enhanced, and user, session and page view identification should be performed. Web personalization is a domain that has been recently gaining great momentum not only in the research area, where many research teams have addressed this problem from different perspectives, but also in the industrial area, where there exists a variety of tools and applications addressing one or more modules of the personalization process. Enterprises expect that by exploiting the information hidden in their Web server logs they could discover the interactions between their Web site visitors and the products offered through their Web site. Using such information, they can optimize their site in order to increase sales and ensure customer retention. Apart from Web usage mining, user profiling techniques are also employed in order to form a complete customer profile. Lately, there is an effort to incorporate Web content in the recommendation process, in order to enhance the effectiveness of personalization. (MagdaliniEirinaki, MichalisVazirgiannis, 2003).

PROPOSED WORK:

log analyzer is stored on the computer and also embedded in to website. Log analyzer provides information and counts unique and repeat visitors. Tracks how each page of their website is used and tracks how user entered the site. This helps to understand the behavior of their customers as they visit the site. It is used for website optimization. Log Analyzer gives the optimum data to a visitor according his request. Log Analyzer also is used for downloadable applications. This is typically used to monitor and record the activity of a site.

- Log analyzer allows a website to recognize the user when he/she makes multiple visits on a website.
- Log analyzer is an important authentication function may be used by website operator.
- Log analyzer also enables to personalize the website to suit their individual needs.
- Log analyzer is used by those sending emails to identify which recipients open a message, whether they act on the message and what action they took in response. Log analyzer tells sender how many times a message is forwarded.
- Many email programs allow you to manage how you receive pictures from the internet, you can disallow requests for pictures or displaying pictures form internet or be prompted to allow or disallow each request.
- Website owners use visitor's information to improve their websites and to tell their potential advertisers, how many visitors they get, where those visitors are located and which other sites they visited previously.

Servers store following information for every request.

- IP address.
- Date/time stamp.
- Status of request.
- Referring URL.
- Status of request.
- Type of user agent used software manufacturer and version no.
- Type of operation system.
- Network location and IP address: can include country, city or any other geographic data as well as the host name.
- Time of visit.
- Page visited.
- Time spent on each page of the website.
- Referring site statistics: can include the website you can through to reach this website and search engine query that brought you there.

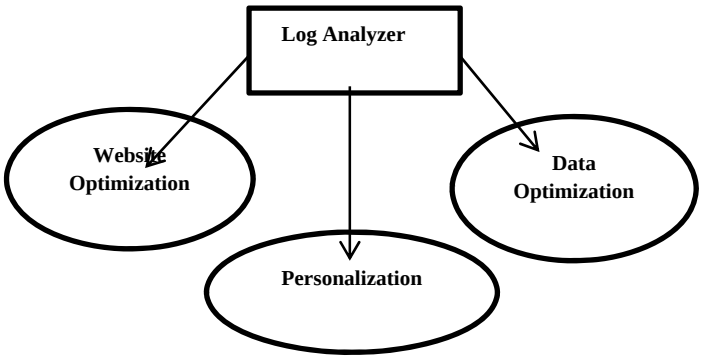

Fig.3.4 Architecture of Log Analyzer

Website Optimization:

- This can be used to collect the information across different domains.
- Typical web statistics are only sored locally in the server where the domain is hosted.
- This also reduces the size of log file.

Personalization:

- This allows a website to recognize a user.
- This is an authentication function.
- This also enables to personalize the website.

Data Optimization:

- It is widely used by websites, which send their raw data to Google and receive an analysis in return.
- Reduce the time spent on each page of the website.
- It has the capability to perform more insightful, detailed reporting on the effectiveness of common online marketing activities such as search engine, pay-per-click advertising, and banner advertising.

ALGORITHM:

Step 1: As visitor connects to the internet, web analyzer counts the visitors.
Step 2: Web analyzer does the task of authentication function with the help of previously recorded information.
Step 3: Web analyzer identifies and records the behavior of the customer if visitor visits the site first time and if the customer is an old, it will find the previous records of the customer and will present the site in accordance with the likening of the customer.
Step 4: Web analyzer stores time stamp of the customer visit.
Step 5: Web analyzer observes the total online expenditure divided by the number of clicking.
Step 6: Web analyzer keeps record of the request status.
Step 7: The same analyzer analyzes marketing investment.
Step 8: Web analyzer reduces the size of log files
Step 9: Updates the visitor record.
Step 10: End

Advantages of Log Analyzer:

- Automatic login: web analyzer remembers your password and username.
- Optimized, personalized information: based on your previous activity on the site, the website can show you information, you may find interesting.
- Location memory: web analyzer records where you stopped accessing a site. This makes it possible for you to re-enter a site and picks up exactly where left off. Example if they were in the middle of making a purchase, you would not have to reselect your purchases or re-enter certain information.
- Site understanding: web analyzer records how do you interact with the website, which helps the owner to develop the sites that better utilize your time.
- Market understanding: web analyzer helps websites learn which offering interest the majority of visitors.
- Customers understanding: web analyzer helps websites infer what interest you are in particular. This allows them to represent you with the information, product or services that might interest you.
- Metrics: web analyzer collects the data such as the number of unique visitors, which have viewed a particular ad or visited a particular web page.
- Log analyzer can optimize your logs just to track and extract what you need to know about visitors without complex filtering. This also reduces the size of log files.
- The same analyzer can be used to collect information across different domains and websites. Since most processing is on the server side, you can just collect all the information on the server.
- Total online expenditure divided by the number of click through the site.
- The percentage of the total number of visitors who make a purchase, sign up for a service or complete another specific action.
- Returns on marketing investment.
- The number of user that visits only a single page divided by the total number of visits.
- Log analyzer can make their shopping and site navigation easier and more enjoyable.
- Log analyzer enables them to better understand and respond to the interests of visitors to their website.
- This can eliminate the need for users to enter their username and password every time they access the site.
- Log analyzer allows e-commerce sites to recognize visitors' generated form online and email advertising campaign.

CONCLUSION:

In this paper a technique to observe the visitors and to collect the information of visitors on the website and then provide the semantic data to the request of visitors is proposed that is web log analyzer. It will work better than the cookies and beacons etc. There is no need to enter password and username every time. Web log analyzer will remember password and username and also personalization

information, Location memory and site understanding. Web log analyzer will optimize your logs just to track what you need to know about visitors without complex filtering. It will also reduce the size of log file. Web log analyzer will be used to collect the information across different domains and websites.

The percentage of the total number of visitors who make a purchase on the site Log analyzer will enable them to better understand and respond to the interests of visitors to their sites. Log analyzer will allow e-commerce sites to recognize visitor's generated form online and email advertising campaign.

REFERENCES

1. Aarti Singh {Agent Based Framework for Semantic content mining} *International Journal Advancement in Technology*, Vol.3 No.2 April 2012, PP 108-113.
2. Ajay Jangra, BhumicaVerma {Web Mining Security, A Survey} *International Journal of Advances of in Information and Technology*, published on February 2012, PP 46-52.
3. Bhavani Thuraisingham, Latifur Khan, and Murat Kantarcioglu, {Semantic Web, Data Mining, and Security} 2010 IEEE.
4. Bhaiyalal Birla, Sachin Patel, {An Implementation on Web Log Mining} *International Journal of Advanced Research in Computer Science and software Engineering* Volume.4 Issues.2 February 2014, PP 68-73.
5. BhavaniThuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen {Data Mining for Security Application} 2008 IEEE/IFIP *International Conference on Embedded and Ubiquitous Computing*, PP 585-589.
6. Geeta R. Bharamagoudar, ShashikumarG.Totad, Prasad Reddy PVGD {Literature Survey on Web Mining} *IOSR Journal of Computer Engineering (IOSRJCE)*, Volume 5, Issue 4 Sep-Oct. 2012, PP 31-36.
7. D.K.Dharmarajan-Scholar, Dr. M.A. Dorairangaswami, {Current Literature Review- Web Mining} September 2014, Volume-1, Special Issue-1, PP 38-42.
8. K. R. Suneetha, Dr. R. Krishnamoorthi, {Identifying User Behavior by Analyzing Web Server Access Log File} *IJCSNS International Journal of Computer Science and Network Security* VOL.9 No.4, April 2009, PP 327-332.
9. L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai, {Analysis of Web Logs and Web User in Web Mining} *International Journal of Network Security & Its Applications (IJNSA)* Vol.3, No.1, January 2011,PP 99-110.
10. M.Karthik, S.Swathi {Secure web mining framework for e-commerce Websites} *International Journal of Computer Trends and Technology (IJCTT)* - volume4 Issue5–May 2013, PP 1042-1046.
11. Mahesh Malviya, Abhinav Jain, Nitesh Gupta, {Improving Security by Predicting Anomaly User Through Web Mining: A Review} *International Journal of Advances in Engineering & Technology*, May 2011, PP 28-32.
12. Magdalini Eirinaki, Michalis Vazirgiannis, {Web mining for web personalization} *ACM*, February 2003.
13. Richa Gupta, {Journey from Data Mining to Web Mining to Big Data} *International Journal of Computer Trends and Technology (IJCTT)* – volume 10 number 1 – Apr 2014, PP 18-20.
14. Ratnesh Kumar Jain , Dr. R. S. Kasana, Dr. Suresh Jain, {Efficient Web Log Mining using Doubly Linked Tree, (IJCSIS)} *International Journal of Computer Science and Information Security* Vol. 3, No. 1, 2009.
15. R.Shanthi, Dr.S.P.Rajagopalan {An Efficient Web Mining Algorithm To Mine Web Log Information} *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 1, Issue 7, September 2013.