IJARIIT
INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 3)
Available online at: www.ijariit.com

Predicting user behaviour through the sessions of web mining

Ujjwala Pandurang Patil
ujjwalapatil52@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon Maharashtra

Akanksha Sahebrao Shejawal
akankshaankita6@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon Maharashtra

Namrata Govind Ambekar
ambekarnamrata007@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon Maharashtra

Nilu Dilip Shinde
nilushinde08@gmail.com
Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon Maharashtra

ABSTRACT

Nowadays users are concentrate on the complex task-oriented goal such as e-commerce site, making finances, making educational details. User divides the particular task into the number of small tasks and solves the multiple requests at the same time. The search engine keeps the track of queries depends on searched history. By extracting the user sessions from the log files and also sessions are extracted. Server-side log and client-side log are commonly used on the web. The server-side log is automatically generated according to each user request. Client-side logs will capture accurate, comprehensive usage data for usability analysis. The process includes the method of data cleaning, user identification, threshold selection, user profile generation, session identification. The mining is performed according to the frequency of user visiting each page. According to the generated sessions, the user behavior can be analyzed depending on time spent on each page.

Keywords: *Weblog, User identification, Session identification, Profile generation*

1. INTRODUCTION

Many users performing some complicated tasks on the web. To support the users in their information search engine keeps the track of all the queries of the user. Search engine groups all the queries of the user into groups. Some of the users are searching social sites, some of the users search educational sites and some are searching for sports sites etc. so, according to different search requests, sessions of the user will be generated. According to that, we can easily find out the behavior of the user, which kind of the user he is. A web may allow users to interact and collaborate with each other in a social media. So, World Wide Web becomes more popular and user-friendly for transferring information. Therefore users are more interested in analyzing log files which is more useful for website usage. Data mining is the extraction of knowledge from the huge amount of data sets, to find a relationship and patterns in data that have been not previously been discovered to summarize the data that will be easily understood and useful to the users. Web mining is the data mining technique which is used for extracting the useful information from the data. Web usage mining is one of the applications of data mining technology to extract information from weblog to analyze the user access to websites. Our goal is to generate the sessions according to the user request and generate the user profile and find the behavior of the user.

2.WEB LOG FILES

Web log files are the files which contain the complete information about user browser activities. These log files are automatically created by every user. [1] [6] [3].These log files are automatically created by every user corresponding to the web servers. These log files are in the text format.

2.1 Types of log files:

There are three types of log files:

- Web Server Logs
- Proxy Server Logs
- Browser Logs

1. Web Server Logs: Web servers are costly and the most common data source. They collect large information into their log files. These logs contain the name, IP, date, and time of the request, the request line exactly came from the client, etc. This data is bound together in a single text file and also divided into no of the different log like access log, refer log and error log [10].

2. Proxy Server Logs: Proxy caching is used to decrease the loading time of a web page as well as the reduce network traffic at the server and client side. The actual HTTP request from multiple clients to multiple web servers are tracked by the proxy server [5]. The proxy server log is used as a data source for browsing behavior characterization of a group of unauthorized users that shares a common proxy server.

3. Browser Logs: On the client side, browser history is collected using javascript or java applets. To implement client-side data collection user cooperation is needed [6]. Web server logs are used in the web page recommendation to improve the e-commerce usability.

2.2 Types of log files format:

Web log file is a simple plain text file which records information about each user. Display of log file data is in the three different formats by [6] [9] [10].

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

3. METHODOLOGY

There is a number of methodologies which are used for the implementation like data pre-processing, data cleaning, session identification, user profile generation and classification of generated web log files. In data Pre-processing, it takes web log data as input and process on it for the generation of reliable data. In pre-processing primary features are extracted, and remove unwanted information and transform into the sessions. To achieve its goal Data pre-processing is divided into Data cleaning, user identification and session identification [7] [1].

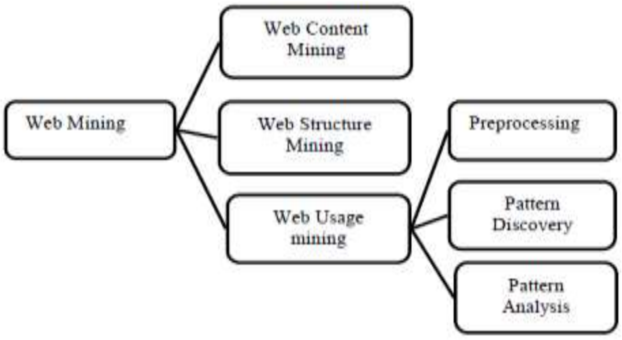

Fig. 1: Web mining classification

- **Data cleaning**: It is the process of removal of unwanted data used in data analysis and mining. For increasing data mining efficiency data cleaning is necessary. The unwanted data include noisy data, graphics records, format efficiency, elimination of HTTP status code records, robot cleaning.

```
Algorithm for data cleaning:
Input: log table
Output: refine log table
Begin
    1. Read records in log table
    2. for each record in log table
    3. Read fields (Status code)
    4. If Status code=200, Then Get all fields.
    5. If suffix UR Link = {*.gif,*.jpg,*.css,*.ico} then,
    6. Remove suffix.UR link
    7. Save fields in the new table.
    End if
    Else
    8. Next record
    End if
    End
```

- **User Identification**: The goal of user identification is to retrieve every user's access characteristics. After this making user clustering and provide the requested service to the users. Each user is identified according to his/her IP addresses.

```
Algorithm for User identification process:
Input: refine log table
Output: identification of the user
 Begin
    1. Read records in the log table
    2. for each record in dataset do
    3. If current IP is not in List of IP then add the current IP in List of IP mark whole record as a new user and assign a user ID
    4. Else assign the old user ID.
     End else
     End if
     End
```

- **Session identification**: Session is a sequence of pages viewed by a user during one visit. Session of each user is found in pre-processing of data and it also defines the number of times the user has accessed a particular web page. Sessions are recorded in log files. In this session identification, it takes all the page reference of a particular user in a log and divides them into a number of user sessions. These generated sessions will be used as input data vector in classification, clustering, prediction etc.

```
Algorithm for session identification:
Input: user identifier table
Output: identified sessions
Begin
    1. Read records in the log table
    2. for each record in dataset do
    3. If the time required > one hour assign new session ID for that log entry
    4. Increment session ID
    5. Else assign the old session ID.
    End else
    End if
    End
```

- **Experimental Evaluation and Analysis:** By using web logs dataset, evaluation of session identification is done. For building User Normal Profile same data set is used. By using ‘μ + σ *α’ and ‘μ - σ *α’ threshold range is generated. For normal Distribution, the value of ‘α’ ranges is between 1 and 3. Detection rate and the false positive rate is evaluated for the different values of ‘α’.

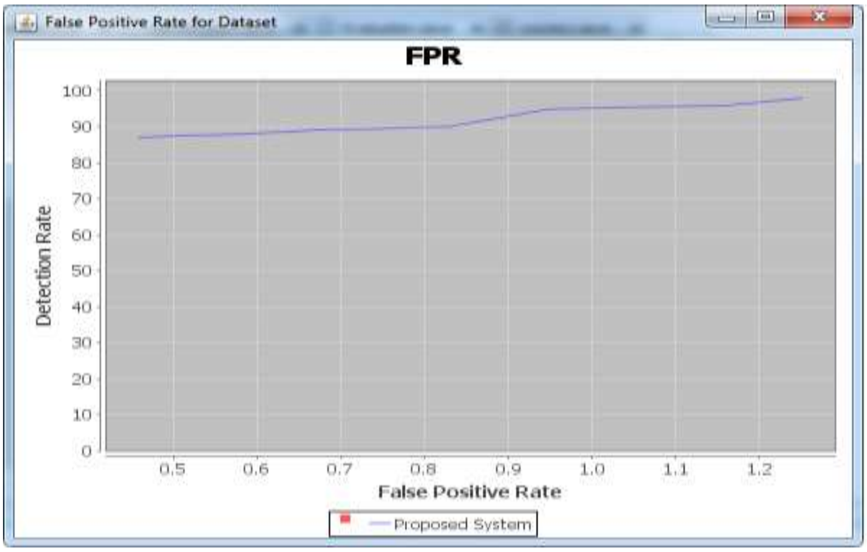

Fig. 2: Graph for detection of false positive rate vs. detection rate

Advantages of proposed system:

1. More accuracy.
2. Less time consuming as compared to the previous system.
3. Classification is done with more accuracy as compared to the previous system.

The disadvantage of proposed system:

Processing speed depends on the machine configuration.

4. FUTURE SCOPE

For checking accuracy it can be implemented with other algorithms also. In order to increase and improve accuracy, it can also be implemented using a hybrid approach. It can be developed using real-world dataset.

5. CONCLUSION

Web mining is the process of extracting, emerging and mining useful information from the web page and discovering and matching the patterns from the WWW. Web mining structure is used to generate the summary of the website and web page. It is necessary to conduct pre-processing step effectively and efficiently. For predicting user behavior, various algorithm and number of important steps are to be carried out such as data cleaning, user identification and profile generation, session identification. After performing all the pre-processing steps of data mining, we can apply data mining techniques like clustering, association, the classification for the applications of web usage mining such as business intelligence, e-commerce, e-learning personalization etc. Whenever a user searches for anything from the web search history is generated. According to that search history, we can find out the behavior of the user. In this paper, we show how such information can be used effectively or the task organizing user searched history into query groups. According to the searched request, different session identification and the user profile will be generated. We also deal with weblog to maintain the history of page requests.

6. REFERENCES

[1] G. Neelima and Sireesha Rodda, “An Overview on Web Usage Mining”, Springer International Publishing Switzerland December 2015.

[2] Gan Teck Wei, Shirly Kho, Wahidah Husain, Zurinahni Zainol “ A Study of Customer Behavior Through Web Mining” Volume 2, Issue 1 available awww.scitecresearch.com/journals/index.php/jisct/index,February, 2015.

[3] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, “Web-Page Recommendation Based on Web Usage and Domain Knowledge”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 10, October 2014.

[4] V. S. Dixit • Shveta Kundra Bhatia,” Refinement and evaluation of web session cluster quality”, Springer transaction Received: 20 February 2014 / Revised: 2 May 2014.

[5] Muhammad Muzammal · Rajeev Raman,” Mining sequential patterns from probabilistic databases”, Received: 11 April 2013 / Revised: 11 May 2014 / Accepted: 3 July 2014 © Springer-Verlag London 2014.

[6] Hongzhou Sha, Tingwen Liu, Peng Qin, Yong Sun, Qingyun Liu,”EPLogCleaner: Improving Data Quality of Enterprise Proxy Log for Efficient Web Usage Mining” Available online at www.sciencedirect.com, Information Technology and Quantitative Management ITQM 2013.

[7] Pani, S.K., Panigrahy, L.: Web Usage Mining: A Survey on Pattern Extraction from Web Logs. International Journal of Instrumentation, Control & Automation (IJICA) 1(1) (2011)

[8] Romero, C., Ventura, S., Zafra, A., de Bra, and P.: Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems (received January 8, 2009) (Received in revised form May 4, 2009) (Accepted May 4, 2009)

[9] Siau, K.: Health Care Informatics. IEEE Transactions on Information Technology in Biomedicine 7(1) (March 2003).

[10] R.Shanthi, Dr.S.P.Rajagopalan, “An Efficient Web Mining Algorithm to Mine Web Log Information”, IJIRCCE Vol. 1, Issue 7, September 2013.