DOI 10.29042/2019-4732-4737

Helix Vol. 9 (1): 4732- 4737

User Behavior Prediction Using Enhanced Pattern Tree Data Structure and Web Usage Mining

[1]G.Neelima, [2]Sireesha Rodda

[1]Vignan Engineering College, Duvvada, Visakhapatanam, A.P, India.

[2]GITAM University, Visakhapatnam, A.P, India

Email: [1]gullipalli.neelima@gmail.com

Received: 08th November 2018, Accepted: 24th November 2018, Published: 28th February 2019

Abstract

As we all know that, in today's era World Wide Web places a measurable role by providing lots of information to the user which will be more useful. Thus, to know the process of discovering and analyzing usage patterns of user, the paper aims to predict the behavior of the user from the web logs. The weblogs is termed as web log mining or web usage mining, every click made by the user will be automatically entered into the weblogs of the corresponding web server. The proposed data structure, pattern tree can be used efficiently to store the usage patterns and their frequency of all the users using the path sharing. The Enhanced Pattern Tree (EPT) maintains the relationship between different patterns and the corresponding users as rules. Search for the specific pattern would yield the corresponding user and vice-versa in minimum no. of searches as rule sharing is used by the pattern tree data structure. This research work aims to develop a framework for analyzing user behaviour through user patterns obtained from the web server logs and supports the use of association rules for representing the relationship between user and patterns.

Keywords

Web Usage Mining, Enhanced Pattern Tree, Association Rule, Loge Files

Introduction

Today's epoch World Wide Web plays a major role by providing a huge amount of information in all aspects so by this the user is very much familiar with the web data and the importance of the data for their daily needs, so this paper aims to predict the behavior of the user by providing the information related to the user based on their daily usage [3]. Web mining is the use of data mining technique to automatically extract the information from the web data which consist of web documents, hyperlinks between documents and usage logs of websites, etc. There are three categories of web mining viz., Web usage mining, Web structure mining and Web content mining. These categories focus on knowledge discovery from the web. Web Usage Mining focuses on extracting useful information and patterns of the user for further use [2]. The web usage mining mainly deals with the web server logs, which consists of information like IP address, client/user id, date, time, method, status code and size of the object. These log files are created and generated by the server and hence also referred to as server logs and are handled by the server administrator [5]. By using these log files this paper aims to predict the user given the pattern and also pattern given the user. By processing these data, either using more complicated data mining techniques, or by using simple statistical methods, we can identify or predict the interesting area, and patterns concerning the activity in the Web site. This process includes 5 stages, namely Data Cleaning, User Identification, Session Identification, Enhanced Pattern Tree Construction [3], and Pattern Recognition [1]. Initially, in data cleaning, erroneous data will be removed from the log file. Then each user is identified according to his/her IP address as specified in the log file which is known as user identification. In session identification, the time spent by each user on a particular website will be identified [4]. The frequency of the users visiting a particular website will be represented in the form of a pattern tree in tree construction step. Finally, the usage pattern for every user is extracted. This process is performed by considering the frequency of users visiting each page [13]. The knowledge thus extracted may be utilized to restructure the website so that efficient utilization of time is done. This work supports the use of association rules [11][14] for representing the relationship between user and patterns.

Problem Statement

This research work aims to develop a framework for analyzing user behaviour through user patterns obtained from the web server logs. This work supports the use of association rules for representing the relationship between user and patterns. The pattern tree maintains the relationship between different patterns and the corresponding users as rules. Search for the specific pattern would yield the corresponding user and vice-versa in minimum no. of searches as rule sharing is used by the pattern tree data structure. The following section deals with the experimental results include 5

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Modules, namely Data Cleaning, User Identification, Session Identification, Pattern tree Construction, and Pattern Recognition.

Types of Log Files:
Display of log file data is performed in three different formats.

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

NCSA Common log file format: Stores basic information about the request received. It is a fixed ASCII text-based format so that no one can customize it[5].
Example
216.67.1.91 - - [01/Jul/2004:12:11:52 +0000] "GET /index.html HTTP/1.1"
Modules to Experimental Work:
In this section, the entire experimental work was shown with outputs. Here there are 5 models where each and every model is link with each other and following figure shows the work flow of the paper.

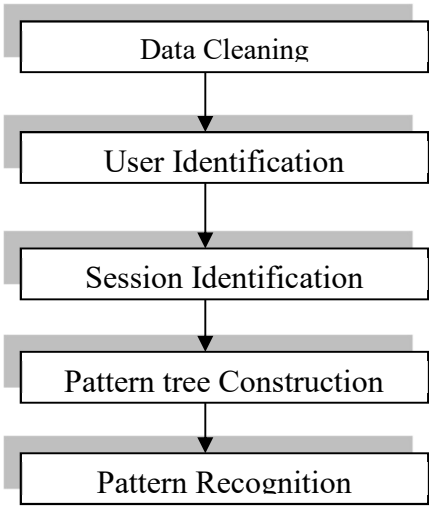

Figure 1: Work Flow

Data Cleaning:
Removes extraneous references to style files, graphics, or sound files that may not be important for the purpose of our analysis. Data cleaning includes [1],

- *Removal of records of graphics, videos and the format information.*
- *Removal of records with the failed HTTP status code.*
- *Removal of records entered during robots navigation.*

Records have a filename extension of GIF, JPEG,CSS, and so on, which can be found in the URL field of the every record, can be removed[2][6]. The records with status codes over 299 or under 200 are removed. Successful transmission (200 series):

- 200: success
- 201: created
- 202: accepted
- 204: no content

The web usage mining mainly deals with the web server logs, which consists of all the information like IP, client, user id, date, time, method, status code and size of the object. As mentioned above the unwanted data is removed from the raw log file dataset and after cleaning the unwanted data the figure1 shows the cleaned data. Following is the statement to get the cleaned data[8].

```
If Status code=200, Then Get all fields.
Ifsuffix.URL_Link={*.gif,*.jpg,*.css,*.ico} then,
Remove suffix.URL_link
```

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

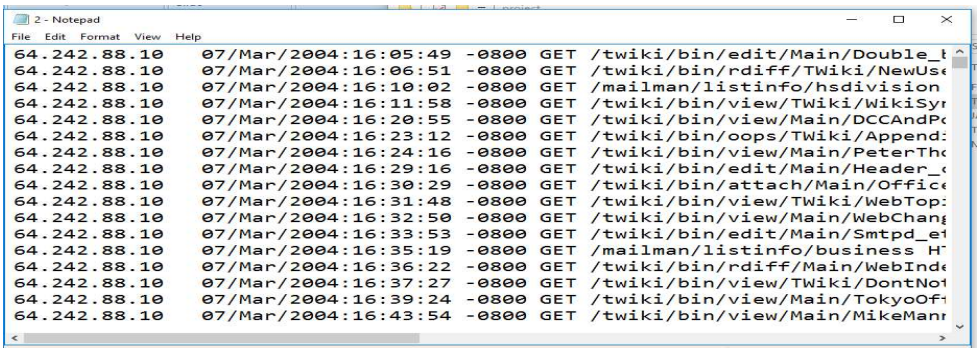

Figure 2: Data Cleaning

User Identification:

In this phase individual users are identified by their unique IP address. The person who has visited the website will have his/her unique IP, that individual IP will be treated as an individual user. Each different user accessing the website is identified. Each individual user is assigned with a unique IP address. Different users are identified by different IP addresses. Following are the statements to find the individual user [5][12].

```
If current IP is not in ListOfIP then add the current IP in
ListOfIP mark whole record as a new user and assign
userID
else assign the old userID.
```

From the figure 2 the user identification was done, here we can see that 14 unique users have identified.

Session Identification:

The person/user who viewed sequence of web pages is treated as a session. The every session is recorded or stored into the log file. Generally, this session identification of each user is one of the very important stages in pre-processing. By finding the session of a user we can define how many number of times the user has accessed a web page. The session takes all the web page reference of a given user in a log and divides them into user session. These sessions will be given or used as the input data in prediction of user behavior [7][6]. Here to find the session the following algorithm statement is used. The session can be known by assigning the time slot.

```
if time_required > one hour assign new sessionID for that log entry
increment sesssionID
else assign the old sessionID.
```

ip	date	zone	method	path	version	response	bytes	username	sessionid
195.11.231.210	12/Mar/2004:03:32:56	0800	GET	/mailman/listinfo/webber	HTTP/1.0	200	6032	10	23
145.253.208.9	12/Mar/2004:04:59:21	0800	GET	/twiki/bin/view/Main/SpamAssassinUsingRazorAndDCC	HTTP/1.1	200	7435	11	24
61.165.64.6	12/Mar/2004:05:25:20	0800	GET	/mailman/listinfo/cncce	HTTP/1.1	200	6208	12	25
145.253.208.9	12/Mar/2004:05:44:35	0800	GET	/twiki/bin/view/Main/SpamAssassinUsingRazorAndDCC	HTTP/1.1	200	7435	11	25
145.253.208.9	12/Mar/2004:05:44:50	0800	GET	/twiki/bin/view/Main/DCC	HTTP/1.1	200	4396	11	25
145.253.208.9	12/Mar/2004:05:51:36	0800	GET	/favicon.ico	HTTP/1.1	200	1078	11	25
67.131.107.5	12/Mar/2004:11:39:25	0800	GET	/twiki/bin/view/Main/WebHome	HTTP/1.1	200	10419	13	25
216.139.185.45	12/Mar/2004:13:04:01	0800	GET	/mailman/listinfo/webber	HTTP/1.1	200	6051	14	25

Figure 3: User Identification & Session Identification

As shown in figure 2,it can be observed that there are 25 sessions done by the different users.

Pattern Tree Construction:

Here the transactions identified from each user session form a collection of paths. Multiple visitors may access the same pages in the same order; we use the tree data structure to merge the paths along common prefixes. Each node corresponds to the occurrence of a specific page in a transaction. It is annotated with the number of users having reached the node across the same trail prefix [4][15].

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Algorithm:
Input: session identified table
Output: frequency representation tree
Begin
1. Read path from table
2. for each path field record in table do
3. copy path in a text file
4. replace / in the text file with space
5. insert the values in the table such that data separated by space in a single column
6. select distinct values from each column and calculate the count
7. display each value along with count starting from first column to last column
End

Figure 4: Enhanced Pattern Tree

From the above tree the patterns are recognized by paths, by considering the log file data set each and every page will be treated as a node , here some of the roots are {twiki,mailman,pipermail,bin.listinfo,..}. Here initial root node will be the root, as the first level and then favicon.ico, pipermail, mailman, twiki as the second level and so on.

Pattern Recognition:

In this phase the patterns in which the users were more interested was identified in this phase by calculating the number of visits for that webpage. Similarly the patterns that are accessed by particular user were identified. This is used to identify the user behavior and the pages that are frequently accessed. From the below figure 4, by giving the user we can get the paths of the user and also the frequency [13][15] , by that we can find the user interested areas and through admin we can provide relevant information to the user.

Algorithm:
Input: refine_logfile
Output: frequency of user visits
Begin
1. enter the path to be recognized
2. calculate the count of path occurred in the table
3. display the path along with ip and frequency of visits
4. enter the ip to be searched
5. display the path accessed by the user with that ip
end

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

username	ip	visits	path
4	10.0.0.153	24	/twiki/bin/view/Main/WebHome
2	128.227.88.79	11	/twiki/bin/view/Main/WebHome
8	142.27.64.35	2	/mailman/listinfo
11	145.253.208.9	4	/twiki/bin/view/Main/SpamAssassinUsingRazorAndDCC
10	195.11.231.210	1	/mailman/listinfo/webber
5	195.246.13.119	9	/twiki/bin/view/Main/WebHome
7	203.147.138.233	1	/favicon.ico
6	207.195.59.160	11	/twiki/bin/view/Main/WebHome
9	208.247.148.12	1	/mailman/listinfo/ppwc
3	212.92.37.62	11	/twiki/bin/view/Main/WebHome
14	216.139.185.45	1	/mailman/listinfo/webber
12	61.165.64.6	1	/mailman/listinfo/cncce
1	64.242.88.10	330	/twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2
13	67.131.107.5	1	/twiki/bin/view/Main/WebHome

Figure 5: Frequency of Paths

Here we are having 14 unique users and the corresponding ip's. By observing the above output screen user 4 visits path /twiki/bin/view/Main/Webhome for 24 times. In the same way we can find the interest of different users by knowing the frequent visits on paths.

```
enter path or -1 to exit
twiki/bin/view
```

username	ip	visits	path
4	10.0.0.153	4	/twiki/bin/view/Main/WebHome
2	128.227.88.79	11	/twiki/bin/view/Main/WebHome
11	145.253.208.9	3	/twiki/bin/view/Main/SpamAssassinUsingRazorAndDCC
5	195.246.13.119	8	/twiki/bin/view/Main/WebHome
6	207.195.59.160	11	/twiki/bin/view/Main/WebHome
3	212.92.37.62	11	/twiki/bin/view/Main/WebHome
1	64.242.88.10	139	/twiki/bin/view/TWiki/WikiSyntax
13	67.131.107.5	1	/twiki/bin/view/Main/WebHome

Figure 6: Path to User Identification

```
enter ip or -1 to exit
195.246.13.119
```

username	ip	path
5	195.246.13.119	/twiki/bin/view/Main/WebHome
5	195.246.13.119	/favicon.ico
5	195.246.13.119	/twiki/bin/view/Main/SpamAssassinAndPostFix
5	195.246.13.119	/twiki/bin/view/Main/KevinWGagel
5	195.246.13.119	/twiki/bin/view/Main/SpamAssassinTaggingOnly
5	195.246.13.119	/twiki/bin/view/Main/SpamAssassinDeleting
5	195.246.13.119	/twiki/bin/view/Main/DCCAndPostFix
5	195.246.13.119	/twiki/bin/view/Main/RelayGateway
5	195.246.13.119	/twiki/bin/view/Main/LinksOfUse

Figure 7: User to Path Identification

By seeing the figure 5 and 6 we can identify the users and frequent patterns. Figure 5 shows that, by entering the path of any transaction we can get different users who are all using that path. Figure 6 shows that, By entering the user i.e., IP address we are getting different paths used by that unique user. So, by this analysis we can identify the user path and can predict the next path immediately by having this pre used web pages.

Conclusion

Web usage mining is indeed one of the emerging areas of research and important sub-domain of data mining and its techniques. Focus on the areas of preprocessing, including data cleaning, session identification, user identification. Used to analyze the user patterns from which information regarding the problems occurred to the users and usage of the website can be obtained within particular intervals of time.

References

[1] Mehra, J., & Thakur, R. S. (2018). An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining. *International Journal of Applied Engineering Research*, *13*(2), 1227-1232.

[2] Neelima, G., & Rodda, S. (2015). An overview on web usage mining. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2* (pp. 647-655). Springer, Cham.

[3] Jadhav, J. N., & Asaithambi, M. (2017). Web Page Recommendation System Using Weighted Sequential Pattern Mining and WLI Fuzzy Clustering.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

[4] Geng, R., & Tian, J. (2015). Improving web navigation usability by comparing actual and anticipated usage. *IEEE transactions on human-machine systems*, *45*(1), 84-94.
[5] Chinnaiyan, R., & Ilango, V. (2015). Analyzing the user behaviours by mining web access log files. *International Journal of Advanced Studies in Computers, Science and Engineering*, *4*(11), 7.
[6] Neelima, G., & Rodda, S. (2016, March). Predicting user behavior through sessions using the web log mining. In *Advances in Human Machine Interaction (HMI), 2016 International Conference on* (pp. 1-5). IEEE.
[7] Like, Z., Zhongbao, K., & Changshui, Z. (2004, October). Session identification based on time interval in web log mining. In *Intelligent Information Processing II: IFIP TC12/WG12. 3 International Conference on Intelligent Information Processing (IIP2004) October 21-23, 2004, Beijing, China* (Vol. 163, p. 389). Springer Science & Business Media.
[8] Mehra, J., & Thakur, R. S. (2018). An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining. *International Journal of Applied Engineering Research*, *13*(2), 1227-1232.
[9] Alphy, M., & Sharma, A. (2018, August). An Improved Hybrid Algorithm for Web Usage Mining. In *International Conference on Wireless Intelligent and Distributed Environment for Communication* (pp. 153-160). Springer, Cham.
[10] Hsu, K. W. (2017). Efficiently and Effectively Mining Time-Constrained Sequential Patterns of Smartphone Application Usage. *Mobile Information Systems*, *2017*.
[11] Aljawarneh, S. A., Vangipuram, R., Puligadda, V. K., & Vinjamuri, J. (2017). G-SPAMINE: An approach to discover temporal association patterns and trends in internet of things. *Future Generation Computer Systems*, *74*, 430-443.Srivastava, M., Garg, R. A. K. H. I., & Mishra, P. K. (2017). A MapReduce-based Parallel Data Cleaning Algorithm in Web Usage Mining. *International Journal of Computer Science and Applications*, *14*(2).
[12] Raphaeli, O., Goldstein, A., & Fink, L. (2017). Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach. *Electronic Commerce Research and Applications*, *26*, 1-12.
[13] Gashaw, Y., & Liu, F. (2017, October). Performance evaluation of frequent pattern mining algorithms using web log data for web usage mining. In *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress on* (pp. 1-5). IEEE.
[14] Shanthi, S. (2017). Survey on Web Usage Mining using Association Rule Mining. *International Journal of Innovative Computer Science & Engineering*, *4*(3), 65-67.
[15] Muruganandam, S., & Srininvasan, N. (2017). Personalised e-learning system using learner profile ontology and sequential pattern mining-based recommendation. *International Journal of Business Intelligence and Data Mining*, *12*(1), 78-93.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495