Contemporary Engineering Sciences, Vol. 8, 2015, no. 13, 573 - 580
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2015.55154

Sequential Web Page Access Analysis Using Web Log

S. Prince Mary

Department of Computer Science and Engineering
Sathyabama University, Chennai, India

E. Baburaj

Department of Computer Science and Engineering
Narayanaguru College of Engineering, Kuzhithurai, TN, India

Abstract

The world wide web contains an increasing amount of websites which in turn contain an increasing number of web pages. When a client views a new website, it has to go through a huge number of web pages to confront their requirements. Web usage scooping is the process of clipping useful information from server logs. Hence, this research reviews the process of discovering sequential patterns of web files by applying genetic method. This method can be applied to analyse the recent visitor's trend and lead to the creation of repeated and most visited pages. The purpose of using a genetic algorithm is to make evolution approach for knowledge extraction and it is also simpler to implement. It can be implemented in several real time applications.

Keywords: web server, log repositories, user identification, session identification, sequential access, genetic algorithm

1. Introduction

The growth of web sites over the World Wide Web has opened a window of opportunity for organizations to analyse the lifetime value of their customers, and also improve their cross marketing strategies. The new strategies involve analysing a large volume of data. The web data are collected in the form of server

logs generated by the interaction of clients with the web site and stored in the form of transaction logs. Some of the most used algorithms in web usage mining process include association rule formation, sequential pattern formation and clustering.

In data mining, genetic algorithm can be used to either optimize parameters for other kinds of data mining algorithms or discover knowledge by itself. The advantage of genetic algorithm becomes more obvious when the search space of a task is large. In our work, this concept is used to find the best sequential pages visited by the user on the site.

Initially, web log files are collected and analysed and then redundant data deleted. Then web log data is stored into the database. The second step is pre-processing of data which includes user identification and session identification. Third step is to combine both user and session identification tables of data to get a sequence of pages visited. Final step is to apply genetic algorithms on these pages to get better visited sequence of pages.

The process of Ordering the web pages is known as sequential web pages. Mining sequential web pages are an important task in the field of mining web logs. Many traditional methods have used to find a sequential web page.

2. Related Work

Data pre-processing phase presents a lot of challenges that has been discussed in various research papers. Web mining is the application of data mining techniques to large web data repositories [1]. The research carried out by [2] and [3] have discussed the difficulty in the identification of users and sessions from web server logs. In certain cases users are identified only by their IP address [4].Distinct user identification techniques detailed by [5] have been experimented on web log of a library of RK university with DUI algorithm.

Most of the websites are using the rating scheme to rate their pages. Most of the companies have used the soft computing techniques for finding the sequential visits. The demerit of the existing system is that for a large number of data those techniques are not producing a better sequential web page.

3. Proposed System

This paper proposes a web mining techniques to find the best web pages which are visited sequentially on the website via Genetic algorithm. This technique will be more efficient for large number of data.

Steps in proposed system:

1. Data Preprocessing
2. Mining Sequential Patterns

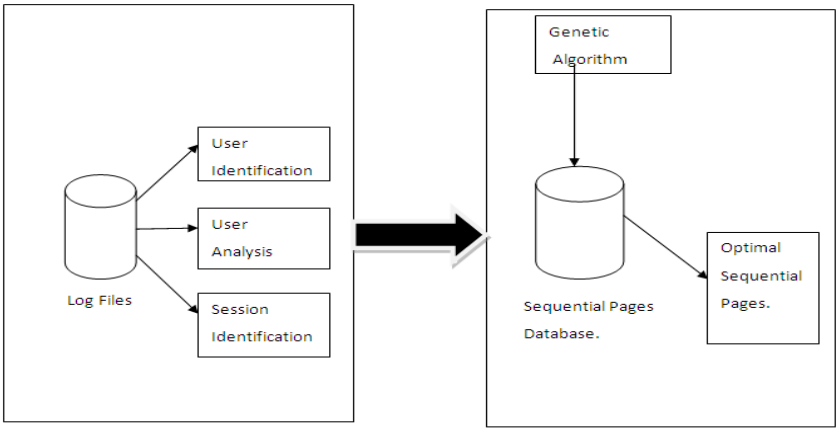


Fig.1. Proposed System Architecture

3.1. Data Preprocessing

This step results in the creation of a suitable user identification and session identification for input into specific data mining operation. Following are the steps to do data preprocessing of log files:

3.1.1. Collection of log files

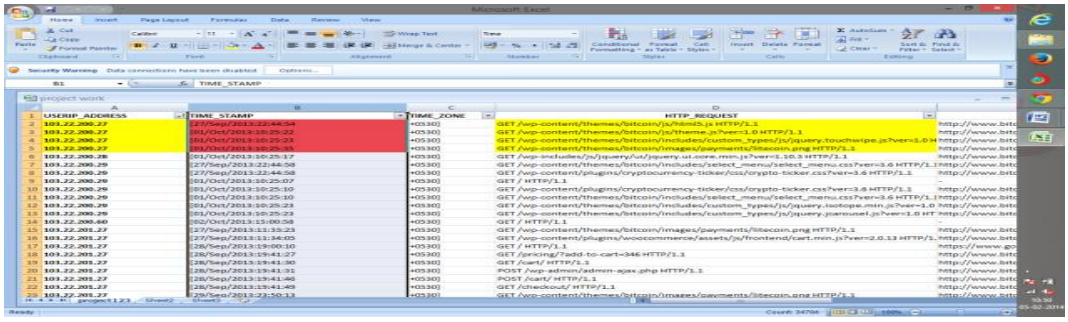
Web log mining got three data sources like web server logs, proxy server logs and browser logs. Here the web server logs are used. The web page requested by the client is stored in a web server as log files. For knowledge retrieval log files are unstructured in database technique. Because of the rapid increase in clients to use the web, data in log files are very important to identify user interest.

```
141.101.99.111 - - [26/Sep/2013:17:44:54 +0530] "GET /pricing/
HTTP/1.1" 200 11118 "-" "Mozilla/5.0 (iPhone; CPU iPhone OS 7_0
like Mac OS X) AppleWebKit/537.51.1 (KHTML, like Gecko)
Version/7.0 Mobile/11A465 Safari/9537.53"
141.101.99.97 - - [26/Sep/2013:17:44:56 +0530] "GET /wp-
content/themes/bitcoin/images/search-2@2x.png HTTP/1.1" 200 3402
"http://www.bitcoinfrenzy.com/pricing/" "Mozilla/5.0 (iPhone; CPU
iPhone OS 7_0 like Mac OS X) AppleWebKit/537.51.1 (KHTML, like
Gecko) Version/7.0 Mobile/11A465 Safari/9537.53"
141.101.99.102 - - [26/Sep/2013:17:45:11 +0530] "GET /wp-
content/uploads/2013/07/favicon.png HTTP/1.1" 200 1739 "-"
"MobileSafari/9537.53 CFNetwork/672.0.2 Darwin/14.0.0"
173.245.62.232 - - [26/Sep/2013:17:46:35 +0530] "GET /robots.txt
HTTP/1.1" 200 88 "-" "Mozilla/5.0 (Windows NT 5.1; rv:6.0.2)
Gecko/20100101 Firefox/6.0.2"
```

Fig.2. Sample Log file

The log file of bitcoinfrenzy website. It has 64000 IP address records. A small part is shown in the fig.2.

Fig.3. Cleaning log files



Log file arrangement in fig.3 gives different IP addresses, time stamp, url request, volume and user agent.

3.1.2. User Identification

The task of this process is to retrieve different user's from the web logs.

User Identification Rules:

- Assorted IP addresses refer to different users.
- Same IP address with different browsers or different operating system considered as different users.

3.1.4. **Session Identification**

In many research Time-out oriented technique was widely used to identify sessions. Based on the predefined time- out threshold interval a session is identified. A new session is identified with time out threshold value of 30 mins.

In the proposed system time out threshold used is 30 mins as well as the system is for online navigational patterns mining, hence a returning visitor request is very important. If the same user visit different pages during their visit, will not be useful for navigational pattern mining. Hence a visitor view some of all pages of their previous visits. To accept a session during 30 mins threshold, there is a need to cross check whether any patterns are shared between sessions. Pages which are common between sessions are known are shared patterns.

Table 1. Shard Pattern

Visitor ID	Session id	Session
V1	S1	p1,p2,p5,p3
V1	S2	p2,p3, p4,p1
V1	S2	p1, p3

In Table 1, for the same visitor there are three sessions given, since the sessions belong to the same visitor, there exists a shared pattern p1, p3. If there is a shared patterns exits between sessions of the same user, then the sessions are accepted. If there is no shared pattern exists, then the corresponding sessions of same visitor will be rejected. Hence the session identification is done by revising the primitive time-out session identification heuristic is to verify that there must be a shared pattern between the identified session from the same sequence. By means of this Quality of sessions is improved.

Algorithm involved in session identification as follows [9]:

```
begin
        sort page sequence according to visitor id and
          timestamp
        for every page sequence do
                divide the current page sequence by threshold
                if the page sequence is single
                        add current page sequence to session list
                else
                        if there exits a shared pattern between the subsequence
                                add subsequence to the session list
                        else
                                skip the sequence
        end for
end
```

To find shared patterns between subsequences efficiently a generalized suffix tree used. Suffix trees are efficient data structure and produce a linear time solution for string problems. This approach in identifying sessions works well with access logs as well as an extended log with referrer formats.

3.2. Mining Sequential Patterns

The combination of the session and user identification will give the sequence of pages categorized by user and session identification rules. In the combination table each IP address has their session time along with their respective reference pages. But for the huge amount of data it will be very difficult to filter the optimal sequence of pages and it is complex as well as it will create confusion too.

3.2.1. GENETIC ALGORITHM

So, for genetic algorithm 4 processes will be there:

1. Selection
2. Crossover
3. Mutation
4. Fitness function

3.2.1.1 Selection

For selection, the combined table of user identification and session identification is taken as input.

3.2.1.2. Crossover

The search of the solution space is done by creating new chromosomes. So, the most important search is crossover. Crossover site is randomly selected and the information is swapped, thus creating new results better than the old ones.

3.2.1.3 Mutation

Each page in each selected crossover is changed with generating a random number between first page number and last page number.

3.2.1.4 Fitness Function

Three steps are there in the fitness function:

- Session

Session time is taken at 15 minutes.

- Support count

Support=∑ {records with session subset} /∑ {session's subset}

- Similarity rate

$$\text{Similarity rate} = \sum_{S=1}^{\text{Sum of sessions}}$$

Fitness=support × similarity rate

4. The Experimental Results

To validate the effectiveness and efficiency of algorithm above, the experiment with web server log of bitcoinfrenzy site [10] has been done. The data source of experiment is from September, 2013 to October, 2013.Experiment was performed on SQL server 2008 and visual studio.

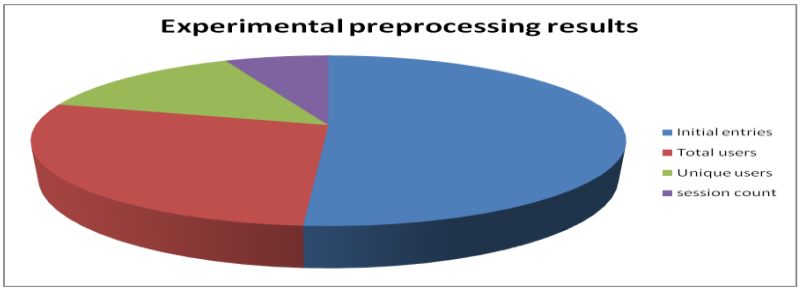


Fig.4. Preprocessing Result

Fig.4.shows the preprocessing result. The 1st section is a graph of initial data which is taken from the website. 2nd section is the total number of users after doing user identifying. 3rd section is the record of unique users after user identification. The final section is the record of total sessions after session identification.

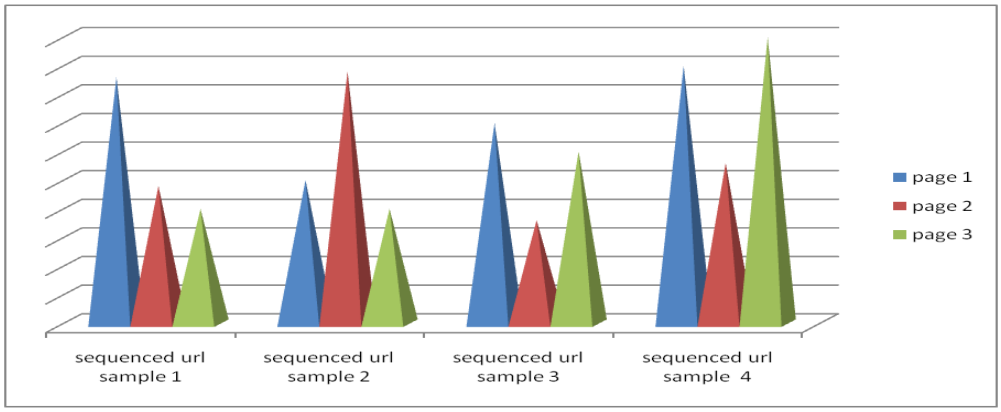


Fig.5. Sequential pages

Fig.5 describes the sample of url in 4 sections. Among them sample 4 is the best sequence of pages that is most visited by users.

5. Conclusion

This paper presents the sequential pages of visited user, which enhances the pre-processing steps of web log usage data in data mining. Firstly, user identification and session identification are applied to the log files. This preprocessing step filters the number of users and the number of distinct users. The sequential access table shows the combination of both user and session identification. All the sequential pages visited by the users can be seen in this table. For getting the best sequential pages, genetic algorithm is used. And lastly best sequential pages are obtained.
In research related to sequential web pages future work can also be done to improve the accuracy of the result.

References

[1] C. C. Aggarwal, J. L. Wolf, and P. S. Yu. ”A New method for similarity indexing of market basket data.” in Proceedings of the 1999 ACM SIGMOD Conference, Philadelphia, PA, June 1999.
http://dx.doi.org/10.1145/304182.304218

[2] P. Pirolli, J. Pitkowand R. Rao, ”Silk from a sow’s ear: Extracting usable structure from the web.” in Proceedings of 1996 conference on human factors in computing systems. Vancouver, British Columbia, Canada, 1996, 118-125. http://dx.doi.org/10.1145/238386.238450

[3] J. Pitkow, ”In search of reliable usage data on the www.” In the sixth international World Wide Web conference. Santa Clara, CA, 1997, 1343-1355. http://dx.doi.org/10.1016/s0169-7552(97)00021-4

[4] Renata Ivancsy, and Sandor Juasz,”Analysis of web user identification ethods.”World Academy of science,Engineering and Technology 2007.

[5] Sheetal A.Raiyani and Shailendra Jain “Enhance preprocessing technique Distinct user identification using web log usage data”.vol 2(4), 526-530.

[6] C. P. Sumathi, R. Padmaja Valli, T. Santhanam, ”An overview of preprocessing of log files for web usage mining.”, vol 34 no.1, Theoretical and Applied Information Technology, 15th December 2011.

[7] Priyanka Patil, Ujwala Patil, “Preprocessing of web server log file for web mining”, National Conference on Emerging Trends in Computer Technology (NCETCT-2012)“ Held at R.C.Patel Institute of Technology, Shirpur, Dist. Dhule, Maharashtra,India. April 21, 2012.

[8] Bhaiyalal Birla, Sachin Patel, Hemlata Sunhare, “Comprehensive Framework for Pattern Analysis Through web logs using web mining.” vol. 2, issue. 4, Monthly Journal of Computer Science and Information Technology ISSN 2320–088X IJCSMC, April 2013, pg. 32 – 37.

[9] S. Prince Mary, Dr. Baburaj. E, "Performance Enhancement in Session Identification", IEEE Conference, ICCICCT-2014. Page(s):837 - 840. http://dx.doi.org/10.1109/iccicct.2014.6993074

[10] For log files- http://www.bitcoinfrenzy.com

[11] Vo B., Hong, T.P., Le, B.: A lattice-based approach for mining most generalization association rules. Knowl. Based Syst. 45, 20– 30 (2013). http://dx.doi.org/10.1016/j.knosys.2013.02.003

Received: May 21, 2015; Published: June 11, 2015