Web Usage Mining for Discovery and Evaluation of Online Navigation Pattern Prediction

Pradnya Mehta
Student
MMCOE, Pune

Shailaja B.Jadhav
Assistant Professor
MMCOE, Pune

R.B.Joshi
Assistant Professor
MMCOE, Pune

ABSTRACT

Web mining is combination of two activated research areas Data Mining and World Wide Web. Web mining is used for mining the interested knowledge from World Wide Web. Web usage mining is used to discover the user access patterns from web server log files. The first step of web usage mining called as data pre-processing used for gaining an accurate web log mining results and good quality input data. The session identification is accomplished by time oriented heuristics. The focus is on referrer log which will which contain information about referrer page of the current page. An efficient approach for discovery of navigation pattern can be done by density based clustering algorithm. An online navigation pattern prediction is proposed by use of K nearest neighbor algorithm along with inverted index concept. The prediction accuracy of patterns can be increased by modifying TF-IDF values to include time spent on page.

Keywords

Web usage mining, User session analysis, Log File Analysis, Indexing, and cluster analysis

1. INTRODUCTION

Web mining is used to salvage, mine and analyze information for knowledge discovery from web documents and services. Web mining is categorized into three parts. First, web content mining is used to mine structured and semi-structured data. Second, web structure mining focuses on the interior document structure which means discovering the link structure of the hyperlinks at the inter-document level and the last, Web usage mining is used to discover the user access patterns from web server log files. A technique to assess the effectiveness of a Web site and its information access tools is through the mining of web log files [11]. Web usage mining has many applications such as personalization of web data, caching, recommendation systems and ecommerce. One of the application is web personalization, consists of following processes: (a) the collection of Web data, (b) the modeling and categorization of these data (preprocessing phase), (c) the analysis of the collected data, and (d) the determination of the actions that should be performed [14].

There are certain problems faced by web usage mining to discover patterns they are [1]:

1. Backward and forward buttons entries are not logged on to the log file.

2. During a specific session, time spent on last page by visitor is difficult to estimate.

3. To make a differentiation between the visitors is one of the major tasks that have to be handled by web usage mining.

All of the above tasks are focused in this paper. The referrer field will give reference pages of last page visited by proposed algorithm. The web usage mining is enclosed with three steps: pre-processing, pattern discovery and result analysis. Data pre processing is categorized as data cleaning, user identification and sessionization In Data cleaning process, the useless requests are removed from log files. For instance all log entries with filename suffixes such as gif, jpeg, JPG must be removed [5]. The robots request is removed from log file. Because of data cleaning process not only good quality input data can be achieved but also log file size can be minimized so less storage space is required. Log file stores IP address along with agents. This information helps in user identification. User identification is achieved by **referrer** oriented heuristic method. Referrer is added to know the referring of current page. The usage data to be dealt with is the access log along with the referrer log of Apache Tomcat web server.

In this paper, the main concern is on web usage mining for discovery and evaluation of online navigation pattern prediction. There are number of techniques that can be used to mine hidden information and the discovery of pattern is done by density based clustering approach. Finally, pattern analysis prediction is done by use of K nearest neighbor (KNN) and inverted index technique. Section 2 describes literature survey; section 3 describes system architecture & proposed algorithm, section 4 deals with results and section 5 states the conclusion.

2. LITERATURE SURVEY

In Web Usage Mining (WUM), it is essential to make a distinction between diverse visitors rather than user's identification. Session is the series of web pages visited by user while browsing a site. If the visitor is signed with website then it is very easy to determine the user, but if visitor is browsing namelessly then it is very tedious task to distinguish between different visitors. The different ways to make differentiation between the visitors are as follows:

Cookies are used to recognize users. Visitor browse the particular website, server will provide the requested resource along with the cookie. But there are certain disadvantages in cookies they are:

1. The privacy concern will not be fulfilled by cookies.[12]

2. A cookie is associated with the user's browser: If the user decides to operate on multiple browsers, they are alleged as multiple persons by the web server.

3. It does not recognize the boundaries of session's i.e. the end of first session and beginning of next session is not determined by cookies. 4. The

server will not be able to discriminate between the re-visitors if users delete cookies stored on their computer [1].

Another method used to make a differentiation between users is eradication of the entire requests coming from same IPS and proxy servers. The weaknesses of this approach are as follows:

1. Legitimate analysis of dataset turn out to be too undersized.
2. Proxies and cache navigational patterns will not be revealed [1].
3. If the request is coming from same IP address repeatedly then the information is not modified in the proxy. It means we are getting old data.

Session id defined as each time visitor visits to website.

It is necessary to know about visitor what information they are browsing each time when they visit same web site number of times. So it becomes necessity to divide several requests of visitor into number of sessions. Exploit of time gap between the sessions is one of the approaches used to identify sessions. Session identification can be made by use of referrer. Referrer will give the reference page information. Time-oriented heuristics considers margins on the time depleted on a page or in the entire site during a single visit. [10]

Knowledge discovery techniques are absolutely designed for the analysis of web usage data. The hidden useful information can be represented in form of rules, graphs; patterns etc.The different paradigms are association rules, sequential patterns, classification, statistical analysis and clustering [4].

1) Sequential Patterns: are used to discover frequent sub sequences among large amount of sequential data. The problem of this approach is if huge amount of data is present then it is difficult to find interesting pattern from it.

2) Association Rules: Association rules are used to find relations or links between frequent item sets that appear together in user's sessions. The drawback of this approach is that the minimum support and confidence is required. Some useful frequent item sets cannot be used to mine the data because if minimum support is kept too low then the data size becomes too large to discover patterns, so accuracy will not be maintained and lots of useless patterns will be generated.

3) Classification: This approach is supervised learning approach. In this, server logs are classified. Its major shortcoming lies in the need for training data and the need to delineate unimodal spectral classes beforehand.

All the above pitfalls are eliminated by making the use of clustering technique. Clustering is a technique used to group together set of items having similar characteristics. The clustering of user will cluster bunch of visitors who refer similar pages. Clustering of pages based on cliques is also possible.

Clustering based on pages will be helpful in internet search engine. The focus is on density based clustering approach because they detect outliers [7]. Another benefit of using density based clustering is number of clusters are not necessary to determine in advance. Density based clusters can handle various types of shapes and sizes. They are very sensitive to input parameter. Input parameter is given by ordering point to identify the clustering structure algorithm.

Online navigation pattern prediction can be done by two ways. One is to make use of clustering and another is classification approach. With K nearest neighbor (KNN) based approach good quality patterns are achieved but scalability is not supported and the vice versa scenario is possible by the clustering approach. Inverted index is used to speed up the search process.

3. THE PROPOSED SYSTEM

3.1 Problem Statement

For making better websites and consequently helping in visitor's retention an accurate web log mining results and efficient online navigational pattern prediction are important. To achieve this goal the following activities will be carried out.

1) Session identification is done by advanced referrer based heuristic; the usage data we have selected to work on is the data logged by a web server in a file called access log file.
2) We are suggesting the usage of OPTICS [22] algorithm with DBSCAN for navigational pattern discovery.
3) Finally, a new approach for efficient online prediction is suggested by using inverted index with K Nearest Neighbor.

3.2 Description

This section describes, how to accomplish superior quality sessions, through which methodology stated in literature survey patterns can be discovered and in what manner the analysis of pattern can be predicted.

The first step is data cleaning and pre processing.

The data cleaning will be done to remove unuseful data. Because of this, the entire http requests which are having error status such as 400 (page not found) are removed. If visitor is browsing website and want to download information at that time gif, jpg, jpeg images are also downloaded and these entries are stored in web server log file. Removal of robots request is also the part of data cleaning. The main purpose of data cleaning is to reduce the size of input data so that good quality can be sustained.

The crawler will automatic crawl information to avoid this robot.txt file will be removed [17]. In this paper, one lexical pattern is defined in which the robot.txt file will never be crawled.

Session identification is done in pre-processing.

Session identification is done by heuristic methods such as time oriented heuristic and referrer based heuristic. Referrer based heuristic uses the time delay of 10 seconds [10].

The session calculation is done by following code

```
Begin calculateSessions()
  1. For each record r in Dataset
  2. For each record r1 in dataset
       if(r.url==r1.referrer OR r.timestamp-r1.timestamp<10)
  3.       add to Session (r1);
       else
```

```
4.      create New Session(r1);
     end for
     end for
```

The time delay delta is used whenever the referrer is undefined. As there are number of requests coming from client, a need arises to divide the sequence of requested pages into separate sessions. Along with this the main focus is on repeated visits of visitor for particular website so pattern creation is important.

The web server log file is input to the system, this log file is cleaned and preprocessed, and sessions are identified. After this process the focus is on clustering. OPTICS (Ordering Points to Identify Clustering Structure) algorithm will collect the identified sessions and will arrange in form of matrix and these are ordered according to the closeness for output. OPTICS will be represented in 2 ways, binary matrix and non binary matrix. In binary matrix number of rows indicating as sessions and number of websites are represented as columns. It will indicate 1 if particular page is not referred within specific session, otherwise for successful operation it will return 0.Dense regions within dataset will be perfectly identified by OPTICS.

The DBSCAN (density based spatial clustering along with noise) algorithm is used for pattern discovery [7]. As it is sensitive to input parameter, OPTICS algorithm will be used for giving input. It detects outliers and does not require prior knowledge about number of clusters required in advance.

Visitor will have hinted what will be the next interesting pattern.

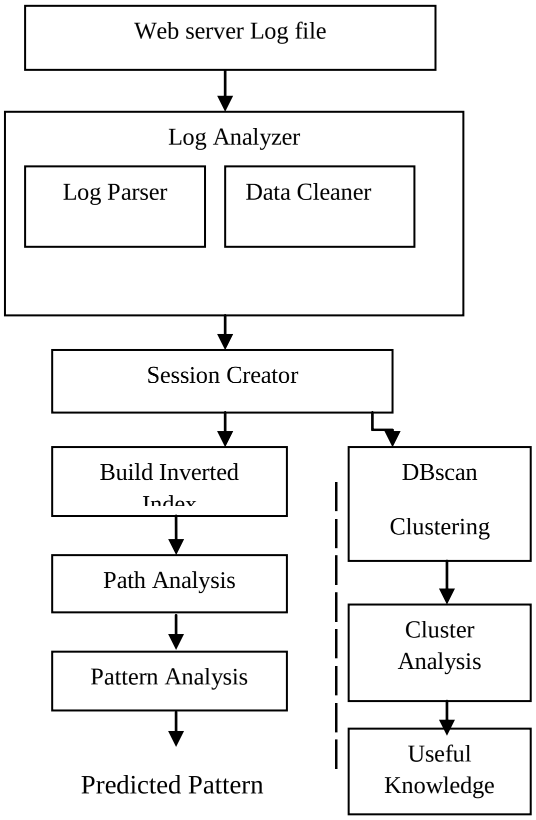


Figure 1: Discovery and evaluation of online pattern prediction

The pattern calculation is done by following method:

```
Begin:
calculateAllPatterns ()
for each session s in SessionList
for each record r in s
  for each record r1 in s
    if(r.url==r1.referrer)
       addtoPatternList (r.url-r1.referrrer)
  end for
end for
End
```

The validity of patterns will be determined by following code

```
Begin
calculateValidPatterns ()
for each Pattern p in PatternList
tfIdf=calculatetf-Idf(p,PatternList);
    if (tfIdf>threshold){
            addtoRecognisedPattern (p);
End
```

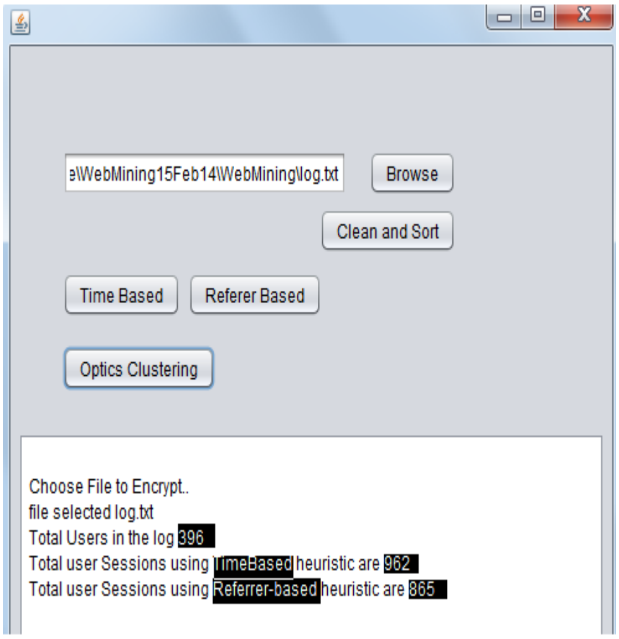
The Tf-Idf value will be calculated as follows:

```
Begin
calculatetfidf (Pattern,PatternList){
            tf=patternCount/PatternList.Size();
            idf=log(PatternList.Size()/patternCount);
            tfIDF=tf*idf;
            return tfIDF;
End
```

Prediction of online navigation pattern will be done by KNN approach but it has the drawback that it will be working for small data size. If large dataset is available the KNN is used along with inverted index concept to increase the prediction accuracy of system.TF-IDF is used to find the closest sessions related to current online sessions. The TF-IDF value will be modified to calculate the time spent on page by visitor. The use of inverted index will speed up the searching process.

4. RESULTS

Mainframe:



Choose File to Encrypt..
file selected log.txt
Total Users in the log 396
Total user Sessions using TimeBased heuristic are 962
Total user Sessions using Referrer-based heuristic are 865

Figure 4.1 Mainframe System

The above figure shows the result of log.txt file (Input of system) which will contain total 396 log entries. We are using referrer based heuristic approach in which total user sessions are 865 and if we are using time based heuristic for creating or identifying sessions then 962 sessions are created. So from this we can say referrer based heuristic is the appropriate methodology to create sessions accurately with a minimum time period. Overall comparative analysis will only be possible after completion of work.

5. CONCLUSION AND FUTURE SCOPE

In this paper, an efficient approach of Web Data Mining for Discovery and Evaluation of Online Navigation Pattern Prediction is defined. In proposed system, referrer log is used for storing the references of current pages. Referrers are determined by referrer based heuristic methodology and time oriented methodology. The clustering of these sessions is done by DBSCAN algorithm which will detect outliers. Clustering technique is highly accurate and efficient. In this paper, the focus is on time spent on the page by visitor. The KNN based approach along with inverted index is suitable for online navigation pattern prediction. Through all these concepts the good quality patterns can be created. Accuracy will be achieved with clustering and speed up process for searching will be achieved through concept of indexing. Though there are lots of advantages in the proposed system still there is a scope to combine DBSCAN with OPTICS to make a single process for efficient pattern discovery.

6. REFERENCES

[1] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley,Reda Alhajj"Effective web log mining and online navigational pattern prediction"(2013)

[2] Juan D. Velásquez "Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments"(2013)

[3] Paweł Weichbroth Mieczysław Owoc Michał Pleszkun "Web User Navigation Patterns Discovery from WWW Server Log Files " (2012).

[4] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood "Web Usage Mining: A Survey on Preprocessing of Web Log File"(2012)

[5] Theint Theint Aye "Web Log Cleaning for Mining of Web Usage Patterns" (2011)

[6] Quanshu Zhou1,"Performance Analysis of Web Applications Based on User Navigation "

[7] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin Ester, Hans-Peter Kriegel, J&g Sander, Xiaowei Xu

[8] Renáta Iváncsy " Different Aspects of Web Log Mining "(2010)

[9] Lara D. Catledge1 "Characterizing Browsing Strategies in the World-Wide (2010)

[10] Bettina Berendt Measuring the accuracy of sessionizer for web usage analysis

[11] M Agosti " Web Log Mining: A Study of User Sessions "(2007)

[12] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou, The impact of site structure and user environment on session reconstruction in web usage analysis, Webkdd 2002 – Mining Web Data For Discovering Usage patterns and Profiles (2003) 159–179.

[13] Prof. Bhupendra Verma1 "Single Level Algorithm: An Improved Approach for Extracting User Navigational Patterns To Improve Website Effectiveness" (2010)

[14] Web Mining for Web Personalization MAGDALINI EIRINAKI and MICHALIS VAZIRGIANNIS(2003)

[15] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Discovery and evaluation of aggregate usage profiles for web personalization, Data Mining and Knowledge Discovery (2002) 61–82.

[16] H. Zhang, A. Ghorbani, The reconstruction of user sessions from a server log using improved time-oriented heuristics, in: Second Annual Conference on Communication Networks and Services Research, Proceedings, 2004, pp. 315– 322

[17] Analia Lourenco,Catching web crawlers in ac

[18] Khalid Hammouda, Prof. Fakhreddine Karray,‖A Comparative Study of Data Clustering Techniques‖ University of Waterloo, Ontario, Canada N2L 3G1

[19] Xin Wang and Howard J. Hamilton " A Comparative Study of Two Density-Based Spatial Clustering Algorithms for Very Large Datasets"

[20] Robert walker Cooley "Web usage mining: Discovery and application of interesting patterns from web data "(2000)

[21] Qingtian Han1" Study on Web Mining Algorithm Based on Usage Mining"

[22] M. Ankerst, M. Breunig, H. Kriegel, J. Sander, OPTICS: ordering points to identify the clustering structure, Sigmod Record 28 (2) (1999) 49–60.