Analysis of Web Server Log Data by Web Usage Mining

Prof. Amit Narote[1], Sana Afsheen Ansari [2], Sagar Singh Bangari [3], Rakshanda Khan [4], Jay Patel [5]

Abstract — With the creating reputation of the World Wide (Web), tremendous volumes of data are collected normally by Web servers and accumulated in log records. Web Usage Mining (WUM) is one of the data mining techniques to discover the learning or information apparent in the web log record, for instance, customer get outlines from web log data and for examining behavioral cases of customers. Examination of server data can give essential and profitable information. In this paper, we mine client visit designs from at least one Web servers for discovering connections between information and give careful consideration to the intriguing new examples. We change a particularly capable Apriori algorithm for planning and captivating new cases and associated encouragement and assurance to determine the measures of fascinating cases, to this particular setting. Apriori algorithm can be required to convey immense number of candidate sets. To imitate the candidate sets, it requires a couple of compasses over the database.

Index Terms— Data Mining, Frequent Pattern, Web Mining, Apriori Algorithm, Support and Confidence.

IJSER

1 Introduction

eb has transformed into a huge chronicle or limit fragmented and undesirable information as well. Though

W

with regards to recouping, setting, sharing and moreover scatter the data or information. Web utilization mining is an essential progression for understanding client's practices on the web. The era of net, utilization of web programs and the scope of web customers are getting increased day by day. Web has ended up being most outstanding stage for attracting and satisfying the customers. Each and every association of web customer with the web will be recorded or secured in content document which is on a very basic level called as Web Log File. These web log archives will be in ".txt" format. Every single connection of customer with web or server will be recorded as a single record in web archive containing logs. Data mining techniques can be connected on the web log documents to evacuate out pointless information and afterward discovering designs out of pre-handled information for breaking down the information to contemplate web client conduct. These pages can be used to perceive the common lead of the customer and to make assumption about needed pages. Along these lines, personalization for a customer can be refined through web usage mining.

The information which is secured in web log records will comprise of a high measure of data with some sort of

it's a little hard to manage entire information which are in colossal measure of size. Along these lines, undesirable or uninterested information can be evacuated by preparing the information. Data mining strategies can be connected on the web log records to evacuate out pointless information and after that discovering designs out of prehandled information for breaking down the information to think about web client conduct. Mass customization and personalization performed by unique Website by collecting groups of clients with comparative access designs and by including navigational connections. The route toward applying the data mining methods on web information to find the fascinating examples is known as web mining.

2 Web Log Files

Web utilization information is the gathering of information that depicts the use of web assets [2]. The utilization information which is utilized for mining purposes can be gathered at various levels i.e. Server level, Client level or Proxy level [6]. In this investigation we will take the instance of web server.

IJSER

A sample Web Log is given below.

```
103.21.58.28 - [14/Jun/2016:05:34:37 - 0500]
―GET/HTTP/1.0‖ 200 3240
```

Where,

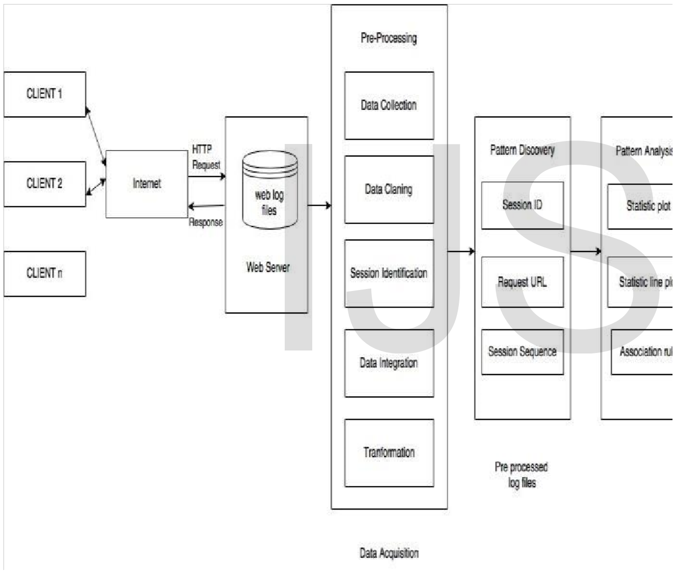
- 103.21.58.28 - IP address
- “- “(hyphen) indicates Anonymous user id
- 14/Jun/2016:05:34:37 - Web page access time
- -0500- The time zone
- GET/HTTP- HTTP request method
- 200- HTTP status code
- 3240- Number of bytes transmitted

The dataset utilized as a part of this work are the web log records. These are the web log documents which are produced as per communication of web client with server or web [2]. Each record line in web log document shows a communication amongst client and server. There are various types of web log documents, which store these sorts of consequently created log information. Data of a typical Web server is shown in figure 1.

```
1. 2016-06-14 05:34:37 W3SVC1378 MDIN-PP-
   WB2 103.21.58.28 GET / - 80 -
   120.63.185.131 HTTP/1.1 Mozilla/5.0+
   (Windows+NT+6.1;+WOW64)+AppleW
   ebKit/537.36+(KHTML,+like+Gecko)
   +Chrome/51.0.
   2704.84+Safari/537.36 - -
   www.iihanaajewellery.com 301 0 0 403 381
   670
2. 2016-06-14 05:34:38 W3SVC1378 MDIN-PP-
   WB2 103.21.58.28 GET /img/panel-logo.png -
   80 -
120.63.185.131 HTTP/1.1
   Mozilla/5.0+(Windows+NT+6.1;
   +WOW64)+AppleWebKit/537.36+(KHTML,
   +like+Ge cko)
   +Chrome/51.0.2704.84+Safari/537.36 -
   http://iihanaajewellery.com/
     iihanaajewellery.com 200 0 0 3903 358
   327
3. 2016-06-14 05:34:40 W3SVC1378 MDIN-PP-
   WB2 103.21.58.28 GET /img/apps/pd-box.gif -
   80 -
120.63.185.131 HTTP/1.1
   Mozilla/5.0+(Windows+NT+6.1; +WOW64)
   +AppleWebKit/537.36+(KHTML,
  +like+Gecko)
  +Chrome/51.0.2704.84+Safari/537.36 -
  http://iihanaajewellery.com/
  iihanaajewellery.com 200 0 0 6175 359
  1544
4. 2016-06-14 05:34:40 W3SVC1378 MDIN-PP-
   WB2 103.21.58.28 GET / - 80 -
   120.63.185.131 HTTP/1.1 Mozilla/5.0+
   (Windows+NT+6.1;+WOW64)+AppleW
   ebKit/537.36+(KHTML,+like+Gecko)
   +Chrome/51.0.
   2704.84+Safari/537.36          -          -
   iihanaajewellery.com 200 0 0 11778 377 2886
5. 2016-06-14      05:34:40      W3SVC1378
      MDIN-PP-WB2
   103.21.58.28  GET  /css/style.css  -  80  -
   120.63.185.131
   HTTP/1.1
   Mozilla/5.0+
   (Windows+NT+6.1;+WOW64)+AppleW
   ebKit/537.36+(KHTML,+like+Gecko)
   +Chrome/51.0. 2704.84+Safari/537.36 -
   http://iihanaajewellery.com/
   iihanaajewellery.com 200 0 0 9396 343
   3182
```

Fig1. Web server log

3 SYSTEM ARCHITECTURE



IJSER

Fig 2. System Architecture

The system architecture for web usage analysis is shown here. There are three main modules in the system namely: ▪ Data Acquisition and Pre-processing ▪ Pattern Discovery

- Pattern Analysis.

A. DATA ACQUISITION AND PRE-PROCESSING

Applications of web usage are mainly based on web data collected from sources like,

- Web servers
- Proxy servers
- Web clients

The steps or methods involved here in web data preprocessing are:

1) Data Cleaning

Data cleaning is the way toward expelling out the irrelevant or pointless information which isn't helpful for future work in mining. Here we are interested on documents or pages or URLs accessed by web client. So, we are simply going to expel out all other web log records which having URLs with file extension other than ".html" like of "gif", "jpeg", "css" etc [4]. After this progression the web log records with file extension ".html" are utilized for additionally process.

2) User Identification

User identification includes procedure of recognizing the web user [1]. User can be related to interesting IP address, Browser utilized likewise alluding working framework utilized by the web clients. Every client is recognized separately.

3) Session Identification

Session identification is the step where in the sessions are identified. Each of the web log record is taken as a single session.

4) Data Transformation

Data transformation is the step in which the information will be changed starting with one frame then onto the next which is a pertinent shape to chip away at [2]. The information can be changed to the least difficult frame which makes simple procedure to manage it. In this stream it's a little difficult to manage long URL frame, each one of the URL is changed as a unique number. So that it will become easy to deal with each URL with that unique number in future process.

B. PATTERN DISCOVERY

Pattern discovery step is performed to discover frequent patterns or knowledge to analyze web user behavior [7]. The discovered patterns or knowledge can be represented in some form like table, graph and charts etc. Variety of techniques used to discover the web patterns;

1) Statistical analysis 2) Association Rules
3) Clustering
4) Classification

❖ Association rules

The age of association rules of given web log information utilizing web mining strategies is done here. In the web usage mining domain, the association rules will allude to the set of web pages which are accessed together and furthermore the set of web page patterns which are accessed as often as possible by the web clients [7]. It additionally briefs to the pages which are referenced together.

An Association rule is an implication, expression of the form X Y, where X and Y are disjoint itemset, i.e., X ∩ Y = Ø. The strength of an association rule can be measured in terms of its support and confidence.

Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain
X.

The formal definitions of these metrics are

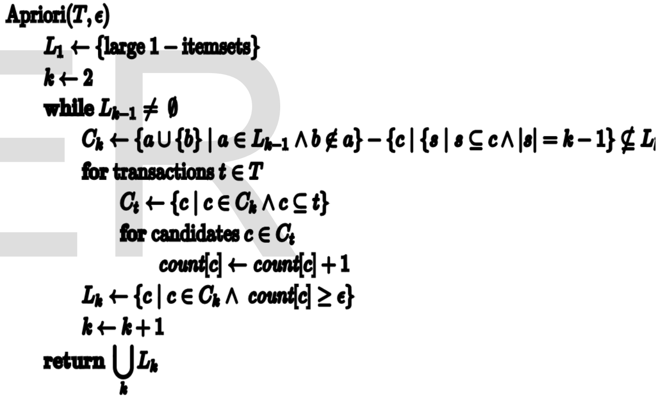
Support , s(X Y) = $\frac{\boldsymbol{Number\ of\ transaction\ in\ which\ X\ appears}}{\boldsymbol{Total\ number\ of\ transactions}}$

Confidence, c (X Y) = $\frac{\boldsymbol{support(X\ U\ Y)}}{\boldsymbol{support(X)}}$

PROPOSED APRIORI ALGORITHM FOR WEB MINING

Apriori is intended to work on databases containing transactions. Apriori utilizes a "bottom up" approach, where visit subsets are broadened one thing at any given moment and gatherings of applicants are tried against the data [5]. Apriori algorithm for mining frequent item sets which are utilized for Boolean association rules. Apriori algorithm is a level-wise, breath-first algorithm which checks transactions. Apriori utilizes an iterative approach known as a level-wise search, in which n-item sets are utilized to explore (n+1)- item sets. To start with, the set of frequent 1-itemsets is found. This set is denoted P1. P1 is utilized to find P2, the frequent 2-itemsets, which is utilized to find P3, until the point when not any more regular n-thing sets can be found. Finding of each P(n) requires one full scan of the database.

Algorithm



```
Apriori($T, \epsilon$)
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L$
        for transactions $t \in T$
            $C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
            for candidates $c \in C_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \mid c \in C_k \wedge count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

The name of the Apriori algorithm depends on the way that the algorithm uses earlier information of frequent itemset property which is that all nonempty subsets of a successive itemset should likewise be visit. The fundamental thought is to find the frequent itemsets [1]. The procedure of the algorithm is as per the following.

Step1. Set the minimum support and confidence according to the user definition.

Step2. Construct the candidate 1-itemsets. And then generate the frequent 1-itemsets by pruning some candidate 1-itemsets if their support values are lower than the minimum support.

Step3. Join the frequent 1-itemsets with each other to construct the candidate 2-itemsets and prune some infrequent item-sets from the candidate 2-itemsets to create the frequent 2-itemsets.

Step4. Repeat the steps likewise step3 until no more candidate item-sets can be created.

C. PATTERN ANALYSIS

The pattern analysis is done to improve the situation by examining the clients to get designs and the way they navigate the sites. This has numerous applications in industry [4]. This pattern analysis will help enhance the business by concentrating on the client exercises and concern. For instance, finding the frequently accessed website which can pre-fetch the arrangement of pages and place them in sequence which will help the client to get to those pages in less time which thus spare the time and fulfil the client necessity.

USABILITY ANALYSIS

Usability Analysis investigation report gives you information, directions and proposals to make your site more usable and easy to use. Next vital thing is the visual viewpoint, which includes the logo and arrangement of components on the home page, and the general look of the site.

IJSER

Usability Analysis is critical for any site, yet in the event that you have an advertising or shopping website page, you have to give careful consideration now. The principal thing your site ought to have is an obviously shown endorsement of trust and believability, on the grounds that if your page isn't reliable, a potential client will squander no time in changing to an old and time-tested site rather than yours. Your items must be indexed productively for introduction, and the shopping basket interface must be strategically located for the client. The enrolment, login, structures and checkout links must be obvious and exact, with the goal that a client has a good experience.

4 IMPLEMENTATION

The log files data consists of Web server of a jewelry store (http://iihanaajewellery.com) from 14 June 2016 to 29 June 2016. The Statistical output of Visitor count and status count is shown in Figure 4 and Figure 5 after the analysis for the Log file data of 12 February 2018.

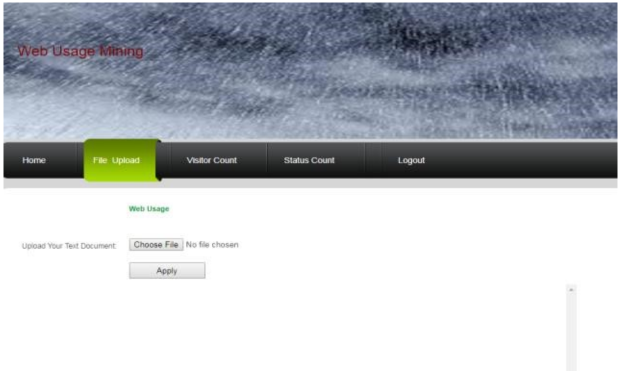


Fig 3. Upload Log File

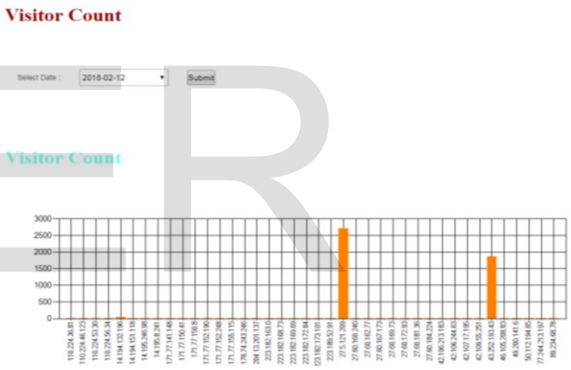


Fig 4. Visitor Count

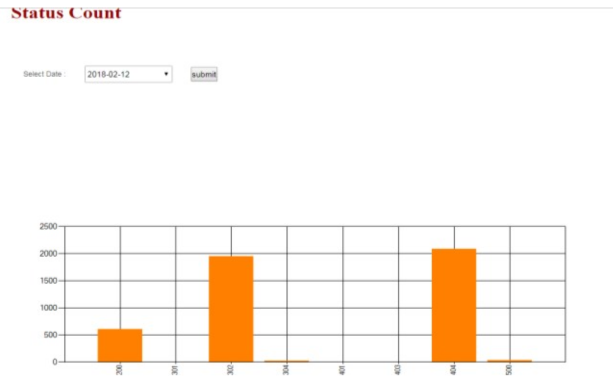


Fig 5. Status Count

5 RESULTS

In this study, we have analyzed the log files of Web server of a jewelry store (http://iihanaajewellery.com). The log files consist the data from 14 June 2016 to 29 June 2016. We have determined different types of errors that occurred in web surfing. Statistics about hits, visitors and bandwidth are shown in table 1. The Website Status statistics is shown in table 2. Different types of errors are shown in Table 3. It is clear from the table that 404 (Table 4) is most frequently occurred error.

Table 1: Summary of statistics

Hits	**Count**
Visitor Hits	134,400
Average Hits per Visitor	9.21
Failed Request	314
Average Page Views per Visitor	1.17
Visitors	
Total Visitors	7,356
Average Visitors per Day	432
Total Unique IPs	4,009
Bandwidth	
Visitor Bandwidth	679.51 MB
Average Bandwidth per Visitor	127.85 KB

IJSER

Table 2: Statistics for Website Status

Sr. No	**Status**	**Hits**
1.	200 OK	28
2.	302 Found	50
3.	304 Not Modified	52
	Total	337

Table 3: Types of errors

Sr. No	**Error**	**Hits**
1.	404 Page Not Found	289
2.	500 Internal Server Error	25
	Total	314

Table 4: 404 Errors (Page Not Found)

No	**Request/Referrer**	**Hits**
1.	/ iihanaajewellery/pad_file.xml No Referrer	45
2.	/img/ssp_screenshot.html No Referrer	65
3.	/ iihanaajewellery /history.html No Referrer	11
4.	/ iihanaajewellery /registration.asp No Referrer	42
5.	/t.php No Referrer	44
6.	/ iihanaajewellery /screenshots.asp No Referrer	16
7.	/)/ No Referrer	29
8.	/ iihanaajewellery /screenshots.php No Referrer	37

6 RELATED WORK

As of late, web usage mining is one of the favored territory of numerous analysts. Web usage mining strategies have been generally used to find intriguing and visit client route designs from web server logs. A novel approach for classifying user navigation patterns and to predict user's future request was introduced in [8]. In another approach, data from a data warehouse and web data can be used to improve marketing activities [9]. A survey about the different categories of web mining e.g. web content mining, web structure mining and web usage mining has done in [10]. A survey on mining interesting knowledge from web logs is presented in [11]. An overview of soft computing techniques (neural network, fuzzy logic, genetic algorithms) used in web usage mining applications is presented in [12, 13].

7 CONCLUSIONS

In order to make a site popular among its clients and counterparts, System executive and website specialist should endeavor to build its adequacy since site pages are a standout amongst the most imperative commercial apparatuses in universal market for business. In this investigation, examination of web server log records of iihanaa jewellery site has done by Datamining. Other sites can be utilized for similar sort of methods to expand their viability. The acquired results can be utilized by system administrator or website specialist and can mastermind their framework by deciding happened framework mistakes, adulterated and broken connections.

8 REFERENCES

Innovative Research in Computer and Communication Engineering, ISSN 2320-9798.

[7] Sandeep Singh Rawat and Lakshmi Rajamani "Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm.", International journal of computer science & information Technology (IJCSIT), August 2010.

[8] Liu, H., and Keselj, V. ," Combined mining of Web server logs and web contents for

Asst. Prof. Padma Dandannavar "Pre-Processing and Analysis of Web Server Logs" , International Journal of Innovative Research in Advanced Engineering (IJIRAE),August 2015, ISSN: 2349-2163.

[2] Navin Kumar Tyagi1, A. K. Solanki2 and Manoj Wadhwa "Analysis of Server Log by Web Usage

Mining for Website Improvement", IJCSI International Journal of Computer Science Iss ues, July 2010.

[3] S. Uma Maheswari1 and S. K. Srivatsa2 "An Application of Preprocessing and Clustering In Web Log Mining", Int. Journal of Philosophies in Computer Science, ISSN 2454-3349, (2015), pp. 21-30.

[4] Tasawar Hussain, Dr. Sohail Asghar, Simon

[1] Chaitra L Mugali, AyeshaAzeema Maniyar, Fong "A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining" unpublished. (Unplublished manuscript).

[5] S.VijayaKumar, A.S.Kumaresan, U.Jayalakshmi "Frequent Pattern Mining in Web Log Data using Apriori Algorithm", International Journal of Emerging Engineering Research and Technology, ISSN 2349-4395 , PP 50-55

[6] Aruna Kumari G K, Sudheer Shetty "Web Usage Mining: Web log Pre-processing and Online Visitor's frequent Pattern Discover", International Journal of

data

classifying user navigation patterns and predicting user's future requests", Data and Knowledge Engineering,2007,Vol 61,Issue 2, pp.304-330.

[9] Arya, S., and Silva, M.," A methodology for web usage mining and its applications to target group identification", Fuzzy sets and systems, 2004, pp.139-152.

[10] R. Kosala, and H. Blockeel," Web mining research: a Survey", SIGKDD Explorations, 2000, 2, pp.1-15.

[11] F.M. Facca, and P.L. Lanzi," Mining interesting knowledge from web logs: a survey", Elsevier Science, Data and Knowledge Engineering, 2005, 53, pp.225-241.

[12] Tug, E., Sakiroglu, and A.M. Arslan, "Automatic

[13] S. Pal, V. Talvar, and P. Mitra," Web mining in soft computing framework: relevance, state of the art and future directions", IEEE Transactions of Neural Networks, 2002, 13 (5), pp.1163-1177.

discovery of the sequential accesses from web log

IJSER