Efficient Identification of Users and User Sessions from Web Log Repository Using Dimensionality Reduction Techniques and Combined Methodologies

G. Shivaprasad, N.V. Subba Reddy, U. Dinesh Acharya and Prakash K. Aithal

Abstract Web Based Applications are data intensive. In addition to web content and structure, they collect huge amount of data in the form of User interactions with the web, forming Web Log Repository. Application of data mining techniques over the Web Log Repository to extract useful knowledge is referred to as Web Usage Mining. Web Usage Mining consists of three phases—Web Log Preprocessing, Knowledge Discovery and Pattern Analysis. In this paper, an efficient implementation for Web Log Pre-processing using Dimensionality Reduction Techniques and Combined Methodologies is presented.

Keywords Web usage mining · Web log repository · Web log Pre-processing

1 Introduction

Web Usage Mining is the process of extracting useful information from Web Log Repository by the application of Data Mining technique. Extracted patterns represent user browsing behaviors. Accurate analysis of these patterns leads to understanding of users visiting the web site thereby improved user satisfaction. Improved customer satisfaction is the key to success of business. Thus, Web Based Applications can improve their business by the application of Web Usage Mining to Web Log Repository. Targeted Marketing, Location Based Marketing, Web Personalization, Fraud Detection and Improved Web Administration are some of the application areas of Web Usage Mining.

G. Shivaprasad (✉) · N.V. Subba Reddy · U. Dinesh Acharya · P.K. Aithal
Department of CSE, Manipal Institute of Technology, Manipal University, Manipal 576104, India

N.R. Shetty et al. (eds.), *Emerging Research in Computing, Information, Communication and Applications*, DOI 10.1007/978-81-322-2553-9_16

Web Usage mining consists of three main steps: Web Log Preprocessing, Knowledge Discovery and Pattern Analysis. Among these tasks, Web Log Preprocessing is most complex and critical for the efficient extraction of useful patterns. Especially, the Web Log Cleaning is more demanding in order to eliminate noisy and irrelevant data and to make the Log Data suitable for Knowledge Discovery. Also the Web Log is memory intensive and pruning irrelevant data reduces the input load of the Knowledge Discovery phase. This paper presents dimensionality reduction techniques to eliminate the noisy data and combined methodologies to efficiently identify users and user sessions from Web Log Repository.

The paper is organized as follows: Sect. 2 presents a brief literature review, Sect. 3 presents the structure of Web Log, and Sect. 4 presents the Web Log Preprocessor, Sect. 5 presents the Experimental Setup and Result analysis and finally, Sect. 6 presents the conclusion.

2 Literature Review

The method of extracting useful information from server log files and different application areas of Web Usage Mining is presented in [1]. A framework for Web Usage Mining consisting of Preprocessing, Pattern Discovery and Users classification, is proposed in [2]. This framework classifies the users based on country, site entry and access time. Information extraction from user navigation history using Web Usage Mining is explored and discussed in [3, 4]. A detailed survey on data collection and pre-processing stage of web usage mining is discussed in [5]. Several data preparation techniques of access stream to identify the unique sessions and unique users are presented in [6]. Educational data mining techniques to analyze learners' behavior, to help in learning evaluation and to enhance the structure of a given course is implemented in [7]. A new algorithm for preprocessing and clustering of web log is proposed in [8]. A specific methodology to extract useful information from an e-commerce website is proposed in [9]. A critical analysis and comparison of the common web robot detection approaches is presented in [10].

3 The Web Log Repository

Web Log Repository is a pool of user activities on a web site. When activated by the web site administrator, it automatically collects the user navigation activities on the web, the moment he enters the web site till the moment he leaves the web site. In Extended Common Log Format (ECLF), a web log usually contains entries with regard to Host IP Address, User Authentication, Date and Time of visit, HTTP Request, Referrer Field and User Agent Field. Details with reference to each field are given below:

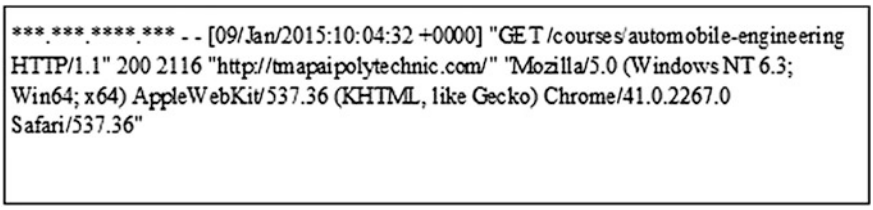

Fig. 1 Extract of experimental data

- *Host IP address*—Used to identify the user visiting the web site.
- *User Authentication*—Contains the Username and password of the user visiting web site, usually empty due freeform of websites.
- *Date and time of the visit*—Tells when the user has visited the web site
- *HTTP Request*—Represents collective information—the Request Method (GET, POST, HEAD, etc.), the Requested Resource (a HTML page, an Image, a CGI program, or a script, etc.) and the Protocol Version (HTTP protocol being used along with version number).
- *Request Status*—Status of the request (200 Series—Successful Transmission, 400 Series- Client Error, etc.).
- *Page size*—Size of the document downloaded in Bytes.
- *Referring Agent (RA)*—Gives the details of the web site from which the user has traversed to the web site. If the user has directly enters this website by typing the web site URL, this field will be "-".
- *User Agent (UA)*—Gives the details with regard to the browser and operating system of the client.

The web access log was collected from the web server of Dr. T.M.A. Pai Polytechnic, Manipal web site [11]. The web site hosts information about courses offered, admission details, facilities available and the placement details, etc. A sample of web log record is given in Fig. 1.

The above log entry indicates that user with IP Address ***.***.***.*** (masked here) requested the link automobile-engineering under courses on 9th Jan 2015 at 10:04:32 AM and he traversed from the link http://tmapaipolytechnic.com. The request was successful and a total of 2116 bytes have been downloaded. Also, it indicates that Mozilla (compatible) 5.0 was the browser and Windows NT 6.3 was the operating system used.

4 Web Log Preprocessor

Web Log Preprocessing plays an important role in Web Usage Mining. The data collected in Web Log Repository is not suitable for Data Mining algorithms. The Log Data needs to be cleansed and converted into structured format before being

processed by Knowledge Discovery Phase. Web Log Preprocessor takes Web Log Repository as input and identifies the Users and User Sessions. We begin Pre-processing phase by Feature extraction and Time Stamp Creation. It is then followed by Data Cleaning employing Dimensionality Reduction Techniques. The original Web Log Repository is blended with relevant and irrelevant information leading to huge log size. Direct processing of this raw data puts unnecessary burden on the Knowledge Discovery Phase. Hence, efficient cleaning of Web Log Repository is necessary to extract useful patterns from Web Log. Once the log is cleansed effectively, Users visiting the web site are identified. Then, User activities in the Web are grouped into meaningful sessions before being processed by Knowledge Discovery Phase. Thus, Web Log Preprocessor mainly contains—Feature Extraction and Time Stamp Computation, Data Cleaning, User Identification, and User Session Identification.

4.1 Feature Extraction and Timestamp Computation

In Feature Extraction step, features are extracted from fields representing collective information so that preprocessing algorithms can be applied. Also, from the date and time entries of web log, time stamp is computed so as to estimate the duration of the user's visit to the Web site and to maintain the sequence of web requests across days. The steps for creation of time stamp are as follows:

(i) Compute the number of days between the web log entry date and a reference date.
(ii) Multiply this number of days by 86,400.
(iii) Find the time in seconds since midnight that is represented by the time in the web log entry.
(iv) Add (ii) and (iii).

4.2 Web Log Cleaning

In Data Cleaning, all irrelevant entries from the log record are eliminated to minimize the burden on the processor. A web log usually contains all the requests to the web server. This includes actual user requests and automated requests. The automated request represents the requests from automated programs like web bots, spiders and crawlers. Similarly, when a user requests a page from the web server, along with the page, any images associated in the requested page is also downloaded and a record for each such image downloads is created in the log. As Web Usage Mining intends to model the user browsing patterns, all such requests need to be eliminated. Similarly, unsuccessful requests are also eliminated. Also, only the request for getting a resource from the web server is retained. Thus, Web Log cleaning consists of the following sub steps:

Fig. 2 IsRobotRequest algorithm

```
Input: Web Log-Record from Web Log Repository
Output: TRUE/FALSE
Algorithm IsRobotRequest
Begin
  For the current Log-Record
      Check whether the Requested Resource Field = 'robot.txt'
      If YES, return TRUE.
      Otherwise, check whether the User Agent Field contains
        words like 'crawler', 'spider', 'bot' or 'robot'.
      If YES, return TRUE. Otherwise, return FALSE.
End
```

Fig. 3 IsImageRequest algorithm

```
Input: Web Log-Record from Web Log Repository
Output: TRUE/FALSE
Algorithm IsImageRequest
Begin
  For the Log-Record
      Check whether the URI field contains file extensions
        like .JPG or .CSS or .png or .js
      If YES, return TRUE.
      Otherwise return FALSE.
End
```

Robot Request Eliminator: Robot requests can be identified using 2 methods:

1. Robot identification based on Requested Page Field
2. Robot identification based on User Agent field.

For efficient identification of web robots, combined methods were employed. The algorithm IsRobotRequest takes as input each log record and returns TRUE/FALSE is given in Fig. 2.

Image Request Filter: The files with the extensions like GIF, JPEG, CSS are also downloaded along with requested page. They are not actually the user interested web page; rather it is just the documents embedded in the web page. So, it is not necessary to include in identifying the user interested web pages. So, the cleaning process eliminates these unnecessary entries from web logs by scanning the Uniform Resource Identifier (URI) field of every record. This step drastically reduces the size of web log. The algorithm for filtering out the Image Requests is given in Fig. 3. The algorithm checks each web log record and returns TRUE/FALSE.

Unsuccessful Request Remover: Successful web requests represent the user actual request to the web server using which the user profile can be modeled. Hence, log records with status codes other than 200 (successful request) are removed. This cleaning process will further reduce the evaluation time for

Fig. 4 IsSuccessfulRequest algorithm

```
Input: Web Log-Record
Output: TRUE/FALSE
Algorithm IsSuccessfulRequest
Begin
    For the Log-Record
            Check whether the Status Code = 200.
            If YES, return TRUE.
            Otherwise, return FALSE.
 End
```

Fig. 5 IsGetMethod algorithm

```
Input: Web Log-Record
Output: TRUE/FALSE
Algorithm IsGetMethod
Begin
    For the Log-Record
        Check whether the HTTP Request Method = GET.
        If YES, return TRUE.
        Otherwise, return FALSE.
 End
```

determining the user interested patterns. The algorithm for Removal of Unsuccessful HTTP requests is given in Fig. 4. The algorithm checks each record of web log and returns TRUE/FALSE.

nonGET Request Remover: A GET method in the HTTP Request Field indicates that the user has requested a resource from the web server. Hence, log records having the value of GET in the Method field of HTTP Request are retained, while all other records are eliminated. This step again reduces the volume of the data to be processed further. The algorithm for elimination of non GET methods from web log is given in Fig. 5. The algorithm takes each record and returns TRUE/FALSE.

The modified Data Cleaning algorithm using above algorithms is given below. The algorithm scans each log record and either retains or discards the record by calling the above algorithms.

```
Input: Web Log Repository obtained after Time Stamp Computation Method
Output: Cleaned Log File
WebLogCleaner Algorithm
Begin
While not eof (WebLogRepository) Do
    Read the next record of WebLogRepository into Log-Record
        If (NOT( IsRobot(Log-Record)) AND NOT(IsImageRequest(Log-Record)) AND IsSuccRequest(Log-Record)
        AND IsGetMethod(Log-Record))
                Write Log-Record to New-LogFile.
        End If
End While
End
```

4.3 Modified User Identification

Identification of each distinct user visiting the website is important and complex task in Web Usage Mining. Apart from the user-id field, the IP address, UA and RA fields can be employed for user identification. In this paper, User Identification based on combined methods using all the three fields has been implemented to uniquely identify the users. In UA field, both the browser and operating system are considered for distinguishing between two users. The Modified User Identification algorithm is given below:

Input: N records of cleaned web log file (New-LogFile)
Output: User set U
ModifiedUserIdentification Algorithm
Begin
User-Count = 0
While not eof (New-LogFile) Do
Read the current record and next record of New-LogFile into Cur-Record and Next-Record.
Let IP_{cur}, IP_{next}, UA_{cur}, UA_{next}, RU_{cur} and RU_{next} be the IP Address, UA and Referrer URL in Cur-Record and Next-Record.
If IP_{cur} <> IP_{next} Identify both entries as belonging to different user, Increment the User-Count by 1.
Else If UA_{cur} <> UA_{next} [Both browser and OS are unique]
Identify both entries as belonging to different user; Increment the User-Count by 1.
Else If RU_{cur}<>"-"and RU_{next} = "-"
Identify both entries as belonging to different user; Increment the User-Count by 1.
Else Assume as same user.
End If
End If
End If
End While
End

4.4 Modified User Session Identification

User Session identification is the process of segmenting the access log of each user into individual access sessions. For Session Identification heuristics based on Time and Navigation are employed. Time based methods are not reliable because users may involve in some other activities after opening the web page. Hence in this paper, a combined technique based on both the heuristics is employed for Session Identification. This method uses web topology and page stay time. The Session Identification algorithm is given below.

```
Input: User sets with N records
Output: Constructed Sessions SessionSet
ModifiedSessionIdentification Algorithm
Begin
    Let PageStayTime ←10 minutes
    SessionSet = { }
    K←J←0
    Let Lj, URIj, tj , RUj and Uj denote log entry, URI, time stamp, Referrer URL and user respectively.
    For each unique user Uj do
        For each Lj do
            If RUj = "-" and (tj- tj-1)>PageStayTime
                K←K+1; Sk ← URIj; SessionSet = SessionSet U Sk
            Else If RUj is Present in any Si of Uj, where i=1,2,3,…sessions of User Uj,  Si ←URIj
                Else  K←K+1;  Sk ← URIj; SessionSet = SessionSet U Sk
                End If
            End If
        End For
    End For
End
```

5 Experimental Setup and Results

The web access log was collected from the web server of Dr. T.M.A. Pai Polytechnic web site from 31st Dec 2014 12:09:56 through 11:18:07 15th Jan 2015, a total of 15 days. A total of 5817 requests were recorded during this period. The algorithms were implemented in MATLAB.

5.1 Web Log Cleaning

The Web Log Cleaning Algorithm was applied to the log data after feature extraction and time stamp computation. The algorithm eliminated a total of 4648 records containing multimedia objects, robot requests and failed requests with a total of 1169 clean log records ready for further processing. This means that the size of the log file was reduced to 20 % of the original log size. The Tables 1 and 2 shows the statistics about individual request category and aggregated results of Data Cleaning Step. Figure 6. depicts the distribution of irrelevant Data in Web Log. It is observed that a major portion of Web Log usually consists of irrelevant and redundant data which has to be eliminated to speed up the upcoming mining process.

Table 1 Statistics of individual request category

Request Category	Number of records	Percentage
PNG	408	7
JPEG	2501	43
CSS	291	5
JS	582	10
NOT GET	58	<1
NOT 200	1164	20
Robot	931	16

Table 2 Aggregate results of data cleaning

Statistics	Number of records
Original size	5817
Failed requests (other than 200)	1164
Multimedia objects	3782
Robots	931
Cleaned log size	1169
Percentage in reduction	80

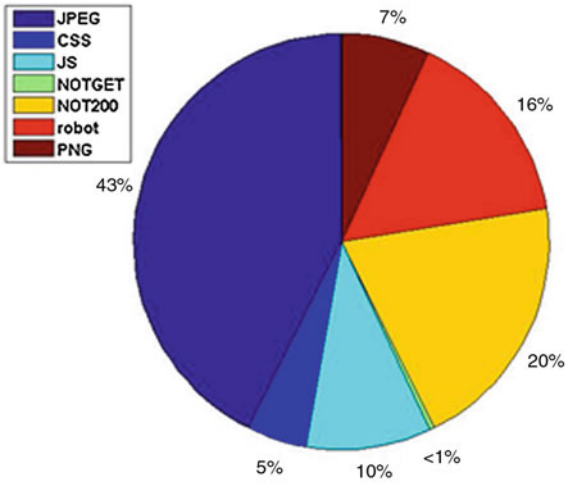

Fig. 6 Distribution of irrelevant data in web log

5.2 User Identification and User Session Identification

The User Identification Algorithm uniquely identifies the users of Web Site. A total of 235 users were identified in the given log. The session identification splits all the pages accessed by each user into individual access sessions using the combined technique based on time oriented and navigation oriented heuristics. It was

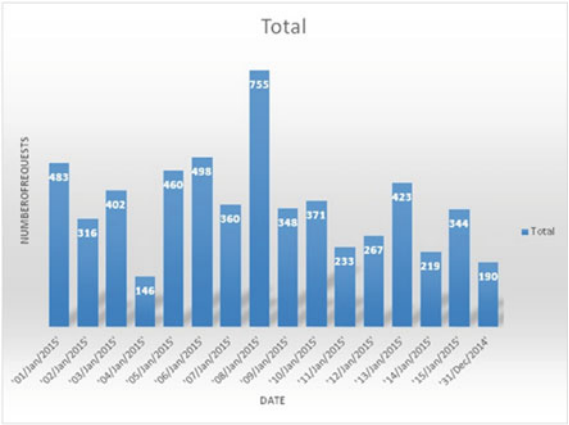

Fig. 7 Days versus no. of requests

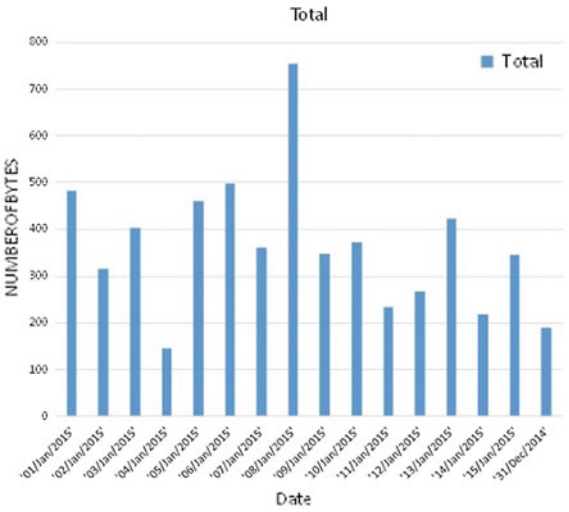

Fig. 8 Days versus no. of byes downloaded

observed that each user having one session with maximum number pages in a session = 26.

5.3 *Analysis of Web Log Pre-processing Results*

The simple analysis of Web Log Repository, after Web Log Preprocessing, could be useful for the web site administrator. The chart of Requests across days and chart of downloads across days is given in Figs. 7 and 8.

6 Conclusion

Web Log Preprocessing is one of the complex tasks of Web Usage Mining. Modified Web Log Preprocessing eliminates the noisy data and drastically reduce the input log thereby lessen the burden on the further tasks. In this paper, the Web Log Preprocessing algorithms based on Dimensionality Reduction Techniques and Combined Methodologies on Web Log Repository from a real time web server is been implemented. The Web Log Preprocessing Algorithm has identified around 16 % robot requests in the log. Results of preprocessing have shown that the input web log size is reduced by 80 %. The results show that the Web Log Preprocessing techniques based on various dimensionality reduction techniques and combination of methods improve the performance of Web Log Preprocessor.

References

1. Neelima, G., et al.: 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, vol. 2, Advances in Intelligent Systems and Computing, vol. 338, pp. 647–655. Springer International Publishing (2015)
2. Bhargav, A., et al.: Pattern discovery and users classification through web usage mining. In: Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on, IEEE, pp. 632–635 (2014)
3. Eltahir, M.A., et al.: Extracting knowledge from web server logs using web usage mining. In: Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on, IEEE, pp. 413–417 (2013)
4. Malik, S.K., et al.: Information extraction using web usage mining, web scrapping and semantic annotation. In: Computational Intelligence and Communication Systems, 2011 International Conference on, IEEE, pp. 465–469 (2011)
5. Varnagar, C.R., et al.: Web usage mining: a review on process, methods and techniques. In: Information Communication and Embedded Systems (ICICES), 2013 International Conference on, IEEE, pp. 40–46 (2013)
6. Sudheer Reddy K., et al.: An effective data pre-processing method for web usage mining. In: Information Communication and Embedded Systems (ICICES), 2013 International Conference on, IEEE, pp. 7–10 (2013)
7. Sael, N., et al.: Web usage mining data pre-processing and multi level analysis on moodle. In: Computer Systems and Applications (AICCSA), 2013 ACS International Conference on, IEEE, pp. 1–7 (2013)
8. Maheswari, B.U., et al.: A new clustering and pre-processing for web log mining. In: Computing and Communication Technologies (WCCCT), 2014 World Congress on, IEEE, pp. 25–29 (2014)
9. Carmona, C.J., et al.: Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Syst. Appl. **39**(12), 11243–11249 (2012). Elsevier
10. Doran, D., et al.: Web robot detection techniques: overview and limitations. In: Data Mining and Knowledge Discovery, pp. 1–28. Springer, US (2010)
11. T.M.A. Pai Polytechnic Web Site: http://www.tmapaipolytechnic.com/