Web Usage Mining: An Implementation View

Sathya Babu Korra, Saroj Kumar Panigrahy, and Sanjay Kumar Jena

Department of Computer Science and Engineering
National Institute of Technology Rourkela, 769 008, Odisha, India
{ksathyababu,panigrahys,skjena}@nitrkl.ac.in

Abstract. This paper describes the implementation of Web usage mining for DSpace server of NIT Rourkela. The DSpace log files have been preprocessed to convert the data stored in them into a structured format. Thereafter, the general procedures for bot-removal and session-identification from a Web log file have been applied with certain modifications pertaining to the DSpace log files. Furthermore, analysis of these log files using a subjective interpretation of recently proposed algorithm EIN-WUM has also been conducted.

Keywords: Data mining, Web data, Web usage mining, DSpace.

1 Introduction

Web usage mining (WUM) is the process of applying data mining techniques to the discovery of usage patterns from Web data, targeted towards various applications [1]. WUM involves mining the usage characteristics of the users of Web applications. This extracted information can then be used in a variety of ways such as— improvement of the application, checking of fraudulent elements etc. The major problem with Web mining in general and WUM in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format and needs a lot of preprocessing and parsing before the actual extraction of the required information. This paper describes about the work in which, a small part of the WUM process has been taken up, that involves preprocessing, user identification, bot-removal and analysis of the log files of DSpace Web server at NIT Rourkela.

2 Data for Web Usage Mining

In Web Mining, data can be collected at the server-side, client-side, proxy servers, or obtained from an organization's database (which contains business data or consolidated Web data for business intelligence [2]). Each type of data collection differs not only in terms of the location of the data source, but also the kinds of data available, the segment of population from which the data was collected, and its method of implementation.

Web Data: The various kinds of data that can be used in Web mining are *Content*— usually consists of multimedia contents such as text, graphics, etc; *Structure*— describes the organization of the contents, i.e., HTML or XML tags, hyperlinks, etc.; *Usage*— the pattern of usage of webpages such as IPs, page references, and the date and time of access; and *User Profile*— demographic information about users of the website such as registration data and customer profile information [1].

Data Sources: The data sources may include Web data repositories— *Web Server Logs*, i.e., a history of page requests [3, 4]; *Proxy Server Logs*, i.e., proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers and serve as a data source for characterizing the browsing behavior of a group of anonymous users, sharing a common proxy server; *Browser Logs*, i.e., client-side data collection can be done by using a remote agent (such as JavaScript or Java applets) or by modifying the source code of an existing browser (such as Mozilla) to enhance its data collection capabilities [1].

Abstract Data: The information obtained by the data sources described above can be used to identify various abstract data— number of hits, number of visitors, visitor referring website, visitor referral website, time and duration, path analysis, browser type, cookies, and platform [5].

Possible Actions: The data collected can be analysed and the following possible actions can be taken— shortening paths of high visit pages, eliminating or combining low visit pages, redesigning pages to help user navigation, and helping in evaluating effectiveness of advertising campaigns [5].

3 Web Usage Mining

There are three main tasks for performing WUM— preprocessing, pattern discovery and pattern analysis [1]. These are briefly explained as follows.

Preprocessing: Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. The different types of preprocessing in WUM are— *usage, content,* and *structure* preprocessing.

Pattern Discovery: WUM can be used to uncover patterns in server logs but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. The various pattern discovery methods are— *Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns,* and *Dependency Modeling.*

Pattern Analysis: The need behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The most

common form of pattern analysis consists of a knowledge query mechanism such as SQL. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

4 Implementation Details

This section describes the various operations that have been done for finding web usage patterns of DSpace server of NIT Rourkela. Different web server log analyzers like Web Expert Lite 6.1 and Analog 6.0 have been used to analyze various sample web server logs obtained. The key information obtained was— total hits, visitor hits, average hits per day, average hits per visitor, failed requests, page views total page views, average page views per day, average page views per visitor, visitors total visitors average visitors per day, total unique IPs, bandwidth, total bandwidth, visitor bandwidth, average bandwidth per day, average bandwidth per hit, average bandwidth per visitor; access data like files, images, referrers, user agents etc.

4.1 Collection of DSpace Log Files

The DSpace server log files were collected and the features found are shown below. The Common log contains the requested resource and a few other pieces of information, but does not contain referral, user agent, or cookie information. The information is contained in a single file. The following example shows these fields populated with values in a common log file record:

host	log	user	date:time GMToffset	request	status	bytes
125.125.125.125	-	-	[10/Oct/1999:21:15:05 +0500]	“GET /index.html HTTP/1.0”	200	1043

4.2 Analysis of Web Server Logs

First part of analysis was preprocessing. Preprocessing segregated all the details provided in the log file into a structured form. JAVA is used for this. Data structures used are linear arrays— ip, time, content, httpmethod, httpstatus, bandwidth, browser etc.

4.3 Key Constraints and Solutions

Not much Variation in IP: As we are considering the DSapce log files, which are specific to NIT Rourkela, it is observed that there is not much variation in IP addresses in the entries recorded in the log file.

Usernames and Aliases not Provided: The second and third entries in the common log format are the usernames and aliases which are mainly recorded in a login based website. These information are not there in the DSpace log files.

Web Crawlers: The various types of crawlers found in the DSpace log files are— MSN bots, Yahoo slurps, Google bots, Baidu spiders etc.

Bot Identification: After much analysis of bot identification and removal, a method has been used specific to DSpace log files to do the same. The pseudocode for the method is as follows:

```
BotId()
{ while(!EOF)
   { readLine();
     Check for keywords (bot,slurp,spider) in browser[] array
     if the array contains keyword
      { botflag=true;   botcounter++; }
     else
        botflag=false;
}  }
```

Identification of User Sessions: User sessions in WUM generally refers to the usage or access of any content of the website from a fixed IP over a fixed period of time. The period of time is subjective to the analyzer. Considering the above requirements, a method specific to DSpace log file has been used to identify user sessions in the log file. The pseudocode for the method is as follows:

```
SessionId()
{ while(!EOF)
  { i=1;
    add first not bot entry to session i;
    for each (next entry)
    { if(entry != bot)
        if(IP == previous IP)
          if(time[this entry] - time[this entry -1] < x)
            add entry to session i;
          else
          { i++; add entry to session i; }
        else
        { i++; add entry to session i; }
} } }
```

4.4 Using EIN-WUM Algorithm

After preprocessing, bot identification and removal, and session identification, the EIN-WUM (Enhanced Immune Network Web Usage Mining) algorithm [6] is used. Our interpretation of the algorithm subject to DSpace website of NIT Rourkela is as follows:

– Limit value of no. of antibodies to 6 (based on the category from DSpace Website).

- We define the category of each entry in the Server Log by assigning it a number (0 through 6). The numbers signify—0 - default value, 1 - content searched by title, 2 - content searched by author, 3 - content searched by date, c - Content searched by author, 5 - content accessed by handle, and 6 - content accessed by bitstream.
- The antibodies are initialized from the first 10 sessions. For each session an entry goes to the corresponding number of antibody as its category is. So each antibody contains only one category of server log entry.
- For each incoming session, compare with each existing antibody. If (similarity of antibody > threshold), replace old session with new session, else if (similarity < threshold) update antibody with most similarity.
- Put a limit on the size of antibody. If (antibody crosses limit), delete old entries.

The various Utilities of the above interpretation are found as:

a. At the end of the program, the ten most interesting antibodies will remain.
b. The contents accessed in the antibodies will be the most frequently accessed contents in the whole website.
c. Based on (b) the following changes can be brought to the concerned site:
  i. improvements on frequently accessed pages.
  ii. deletion or merging of unused pages.
  iii. improvement of content.
  iv. improvement of interaction with referral sites.

4.5 Results

The results obtained from the analysis are given below.
Preprocessed Information from Log Files: The preprocessing program collected the details in the appropriate data structures and also identified whether an entry is a bot entry or a valid user entry (shown below).

```
1    true    203.129.199.129    10/Jan/2010:04:04:26    GET    200
    17013B    0    /dspace/browse-title?top=2080%2F905
2    true    203.129.199.129    10/Jan/2010:04:04:29    GET    200
    14295B    0    /dspace/browse-author?bottom=Misra%2C+M
```

Summary of the Log File and Sessions: The summary of the log file giving overall details, the sessions and the different log file entries that constitute the sessions are shown below.

```
****************Summary****************          session 1   182
number of hits = 14274                            session 1   183
number of visitor hits= 7923                      session 1   191
number of spider hits= 6351                       session 2   193
Number of days= 5                                 session 2   194
Average hits per day = 2854                       session 2   195
Total Bandwidth used = 1494419892 Bytes           session 2   196
Avenrage Bandwidth= 298883978 Bytes               session 2   197
****************************************          session 2   198
```

Usage Patterns: The different frequently accessed contents in the DSpace website is shown below.

```
16  0 1 /dspace/browse-author?starts_with=Das%2C+Atanau
92  0 1 /dspace/browse-author?bottom=Wai%2C+P+K+A
181 0 1 /dspace/browse-author?starts_with=Verghese%2C+L
227 1 1 /dspace/browse-author?top=Joshi%2C+Avinash
364 1 1 /dspace/browse-author?top=Joshi%2C+Avinash
527 5 1 /dspace/browse-author
530 5 1 /dspace/browse-author?starts_with=C
532 5 1 /dspace/browse-author?top=Chatterjee%2C+Saurav
536 5 1 /dspace/browse-author?starts_with=S
569 7 1 /dspace/browse-author?starts_with=S
571 7 1 /dspace/browse-author?top=Chatterjee%2C+Saurav
715 8 1 /dspace/browse-author?top=Bal%2C+S
748 8 1 /dspace/browse-author?starts_with=Das%2C+B+M
831 8 1 /dspace/browse-author?starts_with=Karanam%2C+U+M+R
```

5 Conclusions

The proposed methods were successfully tested on the log files for bot removal and user sessions identification. The results which were obtained after the analysis were satisfactory and contained valuable information about the log files. The methodology and implementation presented in this paper are purely DSpace Website specific. Analysis of above obtained information proved WUM as a powerful technique in Website management and improvement. However, this subjective interpretation of the algorithm EIN-WUM is very ingenious and proposes a lot of scope to be extended on to other problem domains.

References

1. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explor. Newsl. 1(2), 12–23 (2000), http://portal.acm.org/citation.cfm?id=846188
2. Abraham, A.: Business intelligence from web usage mining. Journal of Information & Knowledge Management, iKMS & World Scientific Publishing Co. 2(4), 375–390 (2003), http://www.worldscinet.com/jikm/02/0204/S0219649203000565.html
3. W3C: Logging control in w3c httpd, http://www.w3.org/Daemon/User/Config/Logging.html
4. W3C: Extended log file format. w3c working draft wd-logfile-960323., http://www.w3.org/TR/WDlogfile.html
5. Gupta, G.K.: Introduction to Data Mining with Case Studies. Phi Learning, 1st edn. (2008)
6. Rahmani, A.T., Helmi, B.H.: Ein-WUM: an AIS-based algorithm for web usage mining. In: Ryan, C., Keijzer, M. (eds.) GECCO. pp. 291–292. ACM (2008), http://doi.acm.org/10.1145/1389095.1389144