Chapter 27
Data Preprocessing in Web Usage Mining

Xiang-ying Li

Abstract At present, the study on Web Usage Mining mainly focuses on pattern discovery (including Association Rules, sequence pattern, etc) and pattern analysis. However, the study on the main data sources, that is to say, the study on web-log pre-process is relatively rare. Given that high-quality data helps a lot in improving Pattern mining precision, this paper studies from this aspects, and proposes the high-effective data preprocessing method.

Keywords Client identification · Date cleaning · Path completion · Web usage mining

27.1 Introduction

The main data resource of Web Usage Mining is web log, from which we can know the browsing behaviors of clients. Based on the browsing behaviors of clients (Shao 2009), we can (1) modify the corresponding web link; (2) get to know the interested points of clients, and provide personalized pages for them; (3) subdivide the clients, carry out different promotion strategies for different customers aiming to improve (ROI) return on investment; (4) find out clients' clicks on ads, based on which modify the ads setup (Cooley 1997a).

Data preprocessing is the first part in Web Usage Mining. Whether the data preprocessing is good or bad will directly influence the effect of the following links (Zhao 2003; Wu 2002), such as Association Rules mining, Sequence Pattern discovery and the categorical and clustering of clients and so on.

X. Li (✉)
College of Information Engineering, Shandong Youth University of Political Science, Jinan, China
e-mail: victory_lxy@163.com

E. Qi et al. (eds.), *The 19th International Conference on Industrial Engineering and Engineering Management*, DOI: 10.1007/978-3-642-38391-5_27,
© Springer-Verlag Berlin Heidelberg 2013

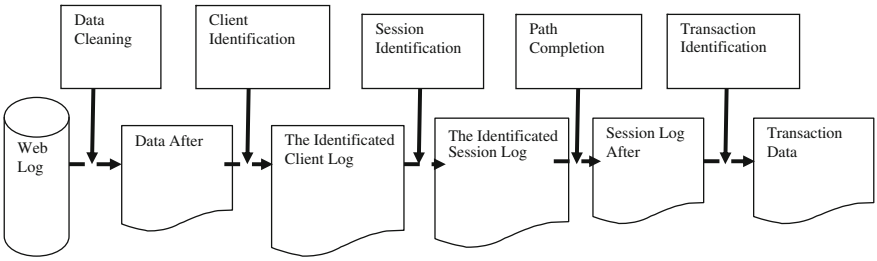

Fig. 27.1 Mining process of web use

In a word, Data Preprocessing can greatly improve the quality of Data Mining and shorten the time needed in practical Data Mining. Web Usage Mining, whose object is mainly web log, is even more affected by Data Preprocessing, for web log, different from the traditional well-structured data base or data from Data Warehouse, is semi-structured. In addition, the incomplete data in web log lead by all kinds of reasons, and the purpose of Web Usage Mining (Liu 2007a; Zhang 2006; Liu 2007b), which is unlike that of transaction data mining, require log files to be preprocessed before mining, converting the log files to format easy to mining and laying foundation for improving the accuracy and effectiveness of final pattern mining. As the Fig. 27.1 shows that it is a complete process of Web Usage Mining. And the paper pays attention to the module of Data Preprocessing, of which several steps, Data Cleaning, Identify Clients, Session Identification and Path Completion are included, as shown in Fig. 27.1.

27.2 Data Preprocessing

27.2.1 Data Cleaning

The task of Data Cleaning is to delete data unrelated to mining, such as pictures of GIF, JPEG and jpg format. These pictures are in WebPages in large numbers, when clients visit WebPages, pictures and cartoons exist in log files as independent records. For most mining task, they can be ignored (of course, we have to reconsider if the websites is special for pictures). Though deleing these records contributes nothing to improve mining effect (Wang 2000; Liu 2003; Tang 2002; Xu 2003; Ji 2009), it can decrease the data to be processed afterwards, improve processing rate and reduce the effect of invalid data upon mining process. The experimental data adopted is a week's logs files (2012\3\1–2012\3\7) from http://my.sdyu.edu.cn/, in all 132 M bytes. Before Data Cleaning, there are 1,274,051 records. After Data Cleaning as the above method, 336,741 records are left. Thus we can see that this step can greatly reduce the data to be processed later, and improve the processing rate.

27.2.2 Clients Identification

Clients' identification is to identify from logs files which clients visit the website and each client visits what web pages.

The clients registered are easy to be identified. However, many clients unregistered surf the internet by web proxy server or several clients share one computer, and the existing of firewall and a client using different browsers, all these add difficulties to client identification. Certainly we can confirm the visiting clients by Cookies, but taken personal privacy into consideration, Cookies are forbidden by many clients' browsers.

How to distinguish many visiting clients using the same computers? Some ideas are to detect whether we can directly visit the page from the page we have visited last time by topology map in websites (Pirolli 1996), if we can't, it is probably that many clients use one computer, and then we can identify clients by Client Session Automatic Generation Technique based on Navigation Patterns (Cooley 1997b).

Considered that the two methods above can only partly solve the problem of clients' uncertainty, the paper proposes an algorithm—Client and Session Identification Algorithm (CSIA). This algorithm takes comprehensive consideration, and combines Client_IP, topology diagram in websites, the browsers version and Referer_Page to identify individual clients, with good accuracy and expansibility. Meanwhile, the algorithm—CSIA also can complete the task of Session Identification, the specific content of the algorithm will be given in Session Identification part, and the paper will first proposes the definition of client as followed:

Definition 1 Clientsi = <Client_ID, Client_IP, Client_URL, Client_Time, Client_RefPage, Client_Agent > , 0 < i < n, n represents the total numbers of clients, Client_ID is the identification of clients identification and Client_IP, Client_URL, Client_Time, Client_RefPage, Client_Agent respectively stands for the IP address of clients, the pages visiting, the time of visiting pages, the pages visited and operating system clients used and the version of browsers, from which the unique client is confirmed.

27.2.3 Session Identification

The time scan of a log file is different, from 1 day, one week–one month etc. In such a long time, a client can not visit a website only for one time, so how to distinguish client's access record left by one—time visit or many times visits? And this involves the problem of Session Identification. A session is time serials of URLs in the process that the client visits a webpage. The relative definition is given as followed:

Definition 2 The definition of $Sessions_i$: $Sessions_i$ = <ClientID, S_j, [$refpage_{j1}$, $refpage_{j2}$,……$refpage_{jk}$] > , 0 < i < n, n represents the total numbers of sessions, ClientID is the identification identified in the process of clients identification,

S_j refers to the j session of the client, $refpage_{jk}$ is the assemble of Refpages-the visited pages in session.

The most simple and most commonly used way to distinguish two different sessions of one client is to set a timeout, usually 30 min. If the request time interval for two web pages exceeds the threshold preset, then the client is thought to begin a new session. Taking 30 min as the timeout has been tested by much practical application, and it is simple and easy to carry out, so the paper adopts this method.

Definition 3 Cube is the saving mode of clients and sessions identified by the algorithm—CSIA. The definition is that Cube = <ClientID, S_j, Client_IP, [$refpage_{j1}(t_{j1})$, $refpage_{j2}(t_{j2})$,……, $refpage_{jk}(t_{jk})$], among them, ClientID, S_j, $refpage_{jk}$ have the same definition as that in first one, and different from definition 2, in definition 3, Client_IP and t_{jk} are added, of which t_{jk} is the page-visiting time.

The thought of algorithm:

A large number of continuous record fragments exist in log files, and each record fragment is from the same Client_IP. If we inspect the record fragments item by item through isExistedIP (Cube) and isSameClient ($Clients_i$), then the running efficiency is dramatically lowered. Therefore the thought of algorithm in this paper is as followed: first judge whether the current record has the same Client_IP as the last one, if they have the same Client_IP, then they are supposed as the same client; or judge whether the client has been identified, the specific judgments are shown in the algorithm below.

This refers to a hypothesis that if the several continuously appeared records is from the same IP address, and then supposes that the continuous records are from the same client. The hypothesis is based on the following two points:

(1) The acceptable client identification accuracy: In order to check the accuracy of client identification under the above hypothesis, we exact 200 record fragments as mentioned above from log files (each continuous fragment is from different IP address). As the hypothesis shows, they are considered as from 200 different clients, however, they are identified as 202 different clients by isExistedIP (Cube) and isSameClient ($Clients_i$) respectively. So it is clear that the above hypothesis has little effect on the accuracy of client identification, and it is acceptable.
(2) The efficiency of algorithm: Adopting the above hypothesis can avoid checking large numbers of records item by item, thus greatly improving the running efficiency of program.

Algorithm: CSIA
Input: log files; TimeSpan
Output: ClientID, Session
(for the specific Algorithm occupies too much space, here proposes the framework in JAVA language limited to space constrains)

```
String [ , ,] Cube;
// Define three-dimensional array, to store Client_ID, the client visiting Urls sequence (time sequence) and the visiting time for each web page and the total number of each session visit.
for (Record r∈Log)    // Record Record shows the records in web log, r represents an individual record;
 {
  if (r.IP<>Last-r.IP)
// The IP address of the current record is not the same as the last one, and Last-r. IP idicates the last record;
  {
   if ( isExistedIP(Cube)= =false )    // the new Ip address, clients identified donít have the IPaddress;
    {
// Store Client_ID, Client_IP, r.Url, r.Time of current client information in to Three-Dimensional Array;
     Client_ID++; //client identification add one;
      SaveCube ( i, j, k, Client_ID, r.IP, r.Url, r.Time );
    }
  else if (isSameClient ( Clients_i, r)==ture )
  // the current client and indentified client in Clientsi is the same visitor;
  {
     if (r.Time- Cube [i, j, 1]<TimeSpan)
    // If The timespan of two times visiting of web page is less than the timeout, then it is regarded as the same session;
    {
        j=j+1;
    Cube [i, j, 0]= r.Url;
    Cube [i, j, 1]= r.Time;
     }
     else  // if overtime,then begin a new session;
     {
     Cube[i, 0, 1]=j-2;  //j means the total numbers of web pages that client visited last session ;
```

```
                i=i+1;
                k++;
                j=0;  //for the new seesion begins , j is set as 0;
               SaveCube ( i, j, k, Client_ID, r.IP, r.Url, r.Time );
                }
        }
   else  //not the same visitor ;
                {
           Client_ID=Client_ID+1;
           i=i+1;
           SaveCube ( i, j, k, Client_ID, r.IP, r.Url, r.Time );
             }
}
```

else *// the IP address of the current record is not the same as that of the last one ,then suppose that they should be the same visitor；*

```
        {      j=j+1;
          Cube[i, j, 0]= r.Url;
          Cube[i, j, 1]= r.Time;
            }
          }
```

```
    public void  isExistedIP ( Cube, r.IP )
    {
```

// Search for the $Client\_List_i$ *that have been identified definition 3, if the IP address exists,then return to ture, or false；*

```
      }
    public void  isSameClient ( Clients_i, r)
    {
```

// take comprehensive consideration onto $Clients_i$ *definition one, to identify whether the current client is the same client as that identified in client-list, if is ,then return to ture, and the exisiting number of session k, or false*

```
    }
```

```
public void  SaveCube ( i, j, k, Client_ID, r.IP, r.Url, r.Time )
{
    // Put Client_ID，r.Url，r.Time store Client_ID，r.Url，r.Time into
three-dimensional array;
        Cube [i, j, 0]= Client_ID+ì －î + k;  // Represents the k session of the
same client;
    Cube [i, j＋1, 0]= r.IP;
        Cube [i, j＋2, 0]= r.Url;
    Cube [i, j＋2, 1]= r.Time;
```

The advantages and disadvantages of algorithm:

(1) High accuracy: Overcome the disadvantage of low identification accuracy caused by the traditional way adopting only IP address to identify visitors.
(2) High efficiency: Realize the client identification and session identification; overcome the low efficiency caused by identifying client and session respectively.
(3) Good data store formats: Store the client and session identified into dynamic three-dimensional array construed by CSIA, avoiding the waste of store space, as shown in Fig. 27.2: client-session axis means client and session identified, among them, n is the total number of clients, k is the k session of some client, and different sessions are shown by different Client ID, such as 1-1, 2-1, 3-1, …, n-k, the same client may have several different sessions, such as the client 3 with two sessions 3-1 and 3-2 in Fig. 27.2. IPn is the ip address of the n client. IP-Urls sequence axis shows the sequence of Urls visited by client in an individual session, and the sequence is arranged by time, for example, the Urls of web pages represented by A, B, C, D, E, F… time axis stores the time that client visit some page, for 2012-3-1 10:21:36 is the time that client 1 visits the page A, this contributes to the sequence pattern mining after data preprocessing. The Fig. 27.2 represents client 1 visits seven pages in one individual session. In this way, it is convenient to store the results of client identification and session identification and related useful information, thus laying foundation for further improving the efficiency and accuracy of pattern mining.
(4) Disadvantage: in judging whether the current client and the client identified is the same visitor, many factors must be considered, thus causing the decrease of the operating rate of algorithm. But the analysis of web log file mining is not real-time, so operating rate is not the top factor considered, and it is worthy of weighing identification accuracy and operating rate.

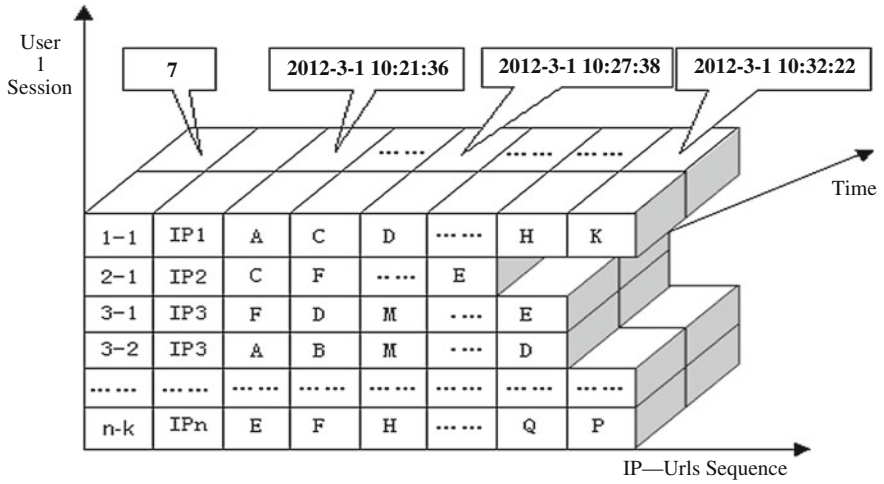

Fig. 27.2 A three- dimensional array data storage

27.2.4 Path Completion

Another important step of pre-processing is Path Completion. The number of URLs those clients browse recorded in web log is less than the actual number those clients browse, because of the page resident in local cache or agent cache in Web Browser, the important information transmitted by Post technique unrecorded in web log and the use of back button in browsing, which also lead to the incomplete of URLs sequence related to client browsing in log files. Therefore to remedy the effect of the question upon pattern mining, path completing is needed.

To avoid this question of the incomplete of access path caused by customers' buffer, we can adopt the HTTP/1.1 agreements that forbid buffering. Though the current website trends to provide dynamic information service, and buffering can not be considered for websites adopted dynamic technique, static HTML page is also applied in website practice, so the treatment of buffering is of great importance for session identification. Meanwhile, only one method can not solve the problem of incomplete access path caused by back button, so some strategies need to take. One strategy is to analyze with the help of topological graph in websites, if the pages visited by uses include the link pointed to the current page, then suppose that request should be sent by the present page (Cooley et al. 1999), so the omitted URLs can be added to the existed path. Though the method can not guarantee the one hundred percent accuracy, it can yet be regarded as an accessible way for it can achieve good effect indicated by experiment.

27.3 Results

Object by a week's log files (2012\3\1–2012\3\7) from http://my.sdyu.edu.cn/, in all 132 M bytes, the log files with 336,741 records remaining after data cleaning, with Athlon (tm) XP 1700+ as test-bed and internal memory 256 M, the test identifies 5,625 clients and 18,536 sessions by CSIA. If only identifies by IP address then only 5,270 be indentified, with $5{,}625-5{,}270 = 355$ clients being ignored, thus we can see the high client identification accuracy of CSIA. From the view of marketing and customer service, if the 355 clients are clearly indentified, and provide personalized service for each of them, then it may be probable that quite a significant proportion of clients turn to the loyal customers, bring more interests to company.

27.4 Conclusion

The paper pays attention to the data preprocessing module in web usage mining, focusing on the specific realization of preprocessing, and proposes the algorithm CSIA that can better identify client and session simultaneously, with higher identification accuracy than the common identification algorithm. And the paper hopes that the work done can help web usage mining researchers.

At present the study on Web Usage Mining at aboard is abundant, and some universities and research institute at home also have studied on this aspect, but the influential one is not that much. Considering the business value, wide application prospect and the large developing space of its related technique of Web Usage Mining, much more research strength will be put into this area. The focus of research will have more tendencies on the visualization of pattern analysis and analysis results, and man–machine interaction aspect based on the continuous study on pattern discovery.

References

Cooley R (1997a) Web mining: information and pattern discovery on the world wide web. In: 9th international conference on tools with artificial intelligence (ICTAI'97), New-port Beach, USA, 1997, pp 558–567

Cooley R (1997b) Grouping web page references into transactions for mining world wide web browsing patterns. In: Proceeding Of the IEEE knowledge and data engineering exchange workshop (KDEX-97)

Cooley R, Mobasher B, Srivastava J (1999) Data preparation for mining world wide web browsing patterns. J Knowl Inf Syst 1:5–32

Ji Y (2009) Application cases of data mining technology. China Machine Press, Beijing

Liu Y (2003) Research on content mining technology based on web. Harbin Institute of Technology, Harbin

Liu W (2007a) Design for web usage mining model. Appl Res Comput 24(3):184–186
Liu L (2007b) The Preprocessing of web usage mining. Comput Sci 5:200–204
Pirolli P (1996) Silk from a sow sear: extracting usable structures from the web. In: Proceeding of 1996 conference on human factors in computing systems (CHI-96), Vancouver, British Columbia, Canada
Shao F (2009) Principle and algorithm of data mining. Science Press, Beijing, pp 379–380
Tang Q (2002) The text mining based on web. Comput Eng Appl 21:198–201
Wang J (2000) Research of web text mining. J Comput Res Dev 37(5):513–520
Wu Q (2002) Client identification in the processing of web log mining. Comput Sci 29(4):64–66
Xu M (2003) Study on text mining on web. Basic Autom (5):44–46
Zhang W (2006) Clustering web client based on interest similarity. Shandong Univ (Nat Sci) 41(6):54–57
Zhao W (2003) Research on data processing technology in web log mining. Comput Appl 23(5):62–64