International Journal of Security and Its Applications
Vol. 10, No. 3 (2016), pp.215-222
http://dx.doi.org/10.14257/ijsia.2016.10.3.20

Research on User Identification Algorithm based on Rewriting URL

Zhou Jiadi and Geng Hai

Zhejiang University of Finance & Economics Dongfang College
jiadizhou@sina.com

Abstract

With the gradual maturity of data mining technology, applying data mining technology into the log files of Web server has become an important research direction in data mining area. This paper focuses on user identification processing technology and proposed a different identification algorithm which should be selected according to different information. On the one hand, use IP address and user access time to identify different users in the logs, namely, heuristic rule-based user identification algorithm; on the other hand, when Cookie is not supported or not allowed by the client browser, add the sessionid field with unique identification behind the original chained address. Better user identification effect can be achieved through this method.

Keywords*: rewrite URL; heuristic rule; cookie; sessionid*

1. Introduction

With the rapid development of Internet, electronic commerce has become one of the most rapidly developing applications on the Internet, such as its low cost, convenient, safe, reliable, free from time and space. In the development of e-commerce, data is increasingly produced.In order to make better use of these data, scientists set up this technology of data mining. The use of Web mining[1] can effectively help enterprises to analyze the large amount of data from the Internet, extract effective information, and guide enterprises to adjust business strategy, to provide customers with dynamic personalized service.

Web mining technology is the combination of Internet technology and data mining technology, which, by mining the data on resource subnet, analyzes the unknown and valuable information hidden in Web. Its main mode is to acquire some patterns that reflect the hobbies and habits of users to a certain extent by analyzing the text in the Web sites accessed by users and clicking the links in the webpage. This analysis process involves many fields such as Web technology[2], ecommerce [3-4], informatics, statistics, artificial intelligence and computational linguistics. Web log mining is to mine and process Web logs to find out the models and habits of user access sites hidden in the bass log file data and analyze these access models and habits, and then optimize the topological structure in the structural adjustment of Web sites to have better interactions with these valuable models. Therefore, Web log mining can help Web site administrators to optimize the network page structure and provide individual service [5].

2 Web Log Mining

2.1 Definition of Web Log Mining

When users are accessing webpages on the internet or have various operations on the website [6], the log files of server will record these behaviors, which is called Web log. The rapid advance of internet technology and the appearance of mass storage device allow the network server to store mass Web log file data [7]. Web log mining is a kind of technology that analyzes the records in Web logs in order to find out the pattern that users

ISSN: 1738-9976 IJSIA

browse Web pages and infer client access behaviors [8].

Different clients have different hobbies of the same site, but there are similarities too, which can be reflected very well by the user access records recorded in server logs. Therefore, common hobbies of clients can be analyzed by the mining of server access logs. On the other hand, the same client may have different browse patterns, but in the long run, these browse patterns will show certain laws and trends, which are able to reflect the interest of client access. Besides, to mine the client browse information in Web logs can get important information such as page access limit, which is very important for administrators to improve the service quality of the pages.

2.2 Web Log Mining Process

Web log mining process includes data collection, data pre-process, and pattern discovery[9] and pattern analysis, as is shown in Figure 2-1.

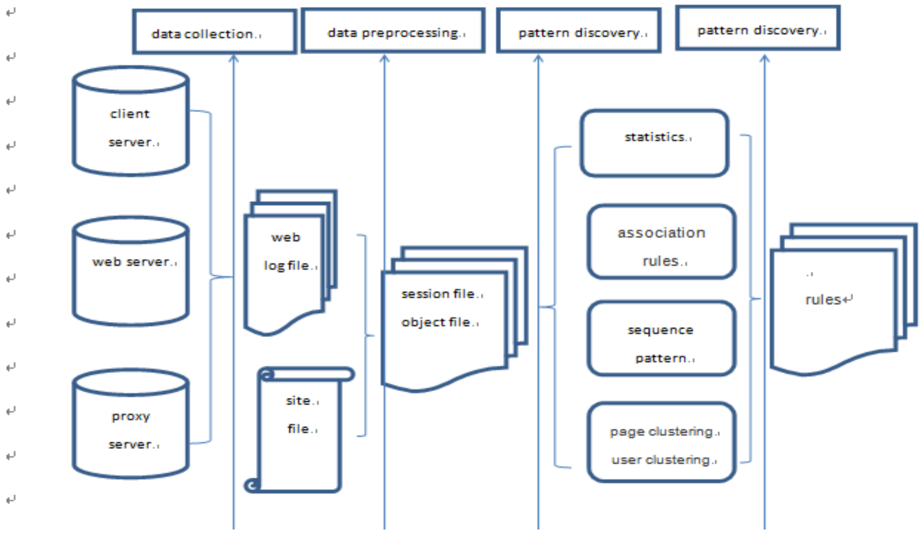

Figure 2-1. Web Log Mining Processing

2.3 Data format of Web logs

Web server will record the information of user access to the server in log files. The log files contain a log of information, mainly including client registration information, Cookies, client investigation information, topological structure and non-structured data of Web site and other related information. Web log files have two record formats: common log format (CLF) and expanded common log format (ECLF). Typical common log file data is as shown in Table 2-1.

All kinds of raw information such as errors during Web server's acceptance of processing request operation and running process is all recorded in the Web logs of website server. The running state of server can be accurately acquired by statistics, analysis and integration of log record data, in order to timey discover and eliminate the causes of error and understand the distribution of user access and provide more basis for better maintenance and administration of the system.

Table 2-1. Data Format of Server Log File

Domain	description
Date	user request time
Time	user request page of the specific time
C-Ip	client C-Ip address
C-Username	client user name
S-Ip	address of the S-Ip server
S-Port	server port number
S-Computername	server name
C-Method	user request method
C-Uri-Stem	user request page
S-status	the status of HTTP returns
S-Bytes	Number of bytes sent by the server
C-Bytes	Number of bytes received by the client
Cs(User-Agent)	service provider
Cs(Referer)	users browse the front page
Cookie(Cs(Cookie))	Cookie identifier
Cs-Host	server operating system

3 Pre-process of Web Log Data

3.1 Data Cleaning

Data cleaning refers to the process to clean the data items in the log files at the Web server end that are unrelated to the selected mining algorithm. According to different representation modes in Web log files and different user interest and mining tasks, different data cleaning strategies can be selected[10].

The following aspects should be taken into consideration for the data cleaning in Web log files:

(1) Data merge: the information recorded by Web server is stored in many files. Therefore, different log files in this stage on the Web server should be merged in order to mine the data used in one stage, and then the merged files should be analyzed and the content of them should be put into data files or database with certain format.

(2) Unrelated data Deletion: the server end can not only record the related information when clients access one webpage, but also can record the automatic download information of pictures and multi-media files related to this webpage. The implicitly requested information is not very useful for use to mine user access patterns, and therefore the unrelated data can be deleted by inspection of the suffix name of the requested domain.

(3) Proxy access processing: the logs of Web server have recorded the mass request information of search engine and other automatic proxies, which will influence our mining result and should be cleaned. One process mode is to examine the proxy domain of each item in the logs and clean many proxy and crawler items by means of string match; the other mode is to examine Robots.txt file and the proxy server will confirm whether some permit proxy access or not by examining this file.

(4) URL normalization: most Web servers tacitly regard default.html or index.html as the request to the catalogue and besides, the WWW before URL is not required and can be omitted sometimes.

Besides, during the process of data cleaning, we should also take other influence factors into consideration, and in practical application development, we should select the proper data cleaning method according to different sites and purposes.

3.2 User Identification

User: refers to the individual that accesses one or more servers via one browser. Local cache, proxy server and firewall make user identification complicated. The accuracy of common user identification [11] is increased by means of IP address, Cookie and user registration.

The user is the individual that accesses the server within a certain time. The relationship between user and log files is one-to-many, and one user may correspond with several log files. User identification is mainly based on the assumption below: each user has an independent IP address when they access the website. When a user finishes the browsing task, the same IP addresses will be assigned to other users correspondingly.

Concept involved during the process of user identification is described as follows:

Web log file R:

$$R =< \mathrm{r}1, r2,..., rk > ......k >= 0 \tag{3-1}$$

K is the total of log records in log file database.

Format of r in Web log record:

$$\mathrm{r} =< data, time, c\_ip, cs\_agent, cs\_uri, cs\_referer > \tag{3-2}$$

Each domain in R is the attribute of log records used during the process of user identification.

User:

$$U\mathrm{ser} =< UID, c\_ip, date\_s, date\_e, time\_s, time\_e, < rs,..., re >> \tag{3-3}$$

UID is the identification number of this user the only identification of this user, c_ip identifies the IP address of this user, date_s,time_s identifies the initiation time of this user access, date_e,time_e identifies the time when this user leaves the current access website, rs identifies the first piece of log record of this user access and re identifies the last piece of access record.

User identification is to find out the collection U of all users u corresponding in each piece of record r in log file R,

$$U = (U\mathrm{ser}1, User2,..., Userk)\ ,\ U\mathrm{ser}.c\_ip = ri.c\_ip\ . \tag{3-4}$$

3.3 User identification Algorithm based on Rewriting URL

Web server usually employs the following methods to track user state: (1) data of user state tracking can be read from the HIDDEN field added in HTML form; (2) rewrite the URL that is used for user state tracking data; (3) transmit the data used for user state tracking by Cookie; (4) use session mechanism. Session refers to the collection of continuous requests to one Web site by the user, and session mechanism is also a very common user tracking method. It can be realized by combining Cookie and URL rewrite. If the browser of the user doesn't support Cookie, the above Cookie-dependent session mechanism can't be achieved. Therefore, the implementation description to improve user identification algorithm by rewriting URL session tracking mechanism is as follows:

(1) Design the rewrite URL function. According to the implementation principle of rewrite URL session tracking mechanism, develop the Web that can identify current session, modify the application program of server end to realize URL rewrite function, and when the user browser doesn't support Cookie, there will also be session tabs of their access records and users can be tracked according to the tabs.

(2) Extract session id for user identification. Access the Web application of implementation of rewrite URL session tracking mechanism and each log record contains session id. If the access record of the user supports Cookie, the session id is at cs (Cookie) field, and if the access record of the user doesn't support Cookie, the session id is at Cs-uri-stem field.

(3) Set two time threshold values MaxTime and MinTime, and method to determine whether the users are the same one according to the two time threshold values is: if the time difference of the user to browse two webpages is larger than MaxTime, they are different users. If it is smaller than MiniTime, they are the same user.

Algorithm flow chart is shown as in Figure 3-1.Next is part of Algorithm program.

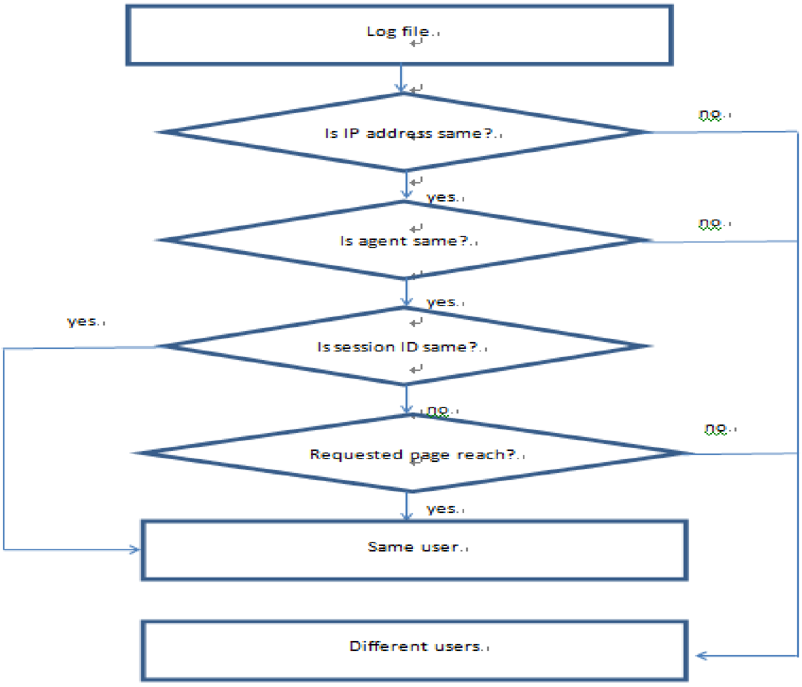

Figure 3-1. The Flow Chart of IASR Algorithm

```
      while(rs.next()){
      commonAgent = false;
      if(!IPList.contains(rs.getString("IP"))){
      IPList.add(rs.getString("IP"));
      User = newUser(rs, ++UserID);
      UserList.add(U ser);
      }
      else{
      for(int i=0; i<UserList.size(); i++){
      if(rs.getString("ip").equals(UserList.get(i
).getLog().getC_ip()){
      if(rs.getString("agent").equals(UserList.g
et(i).getLog().getUser_agent())){
      if(rs.getString("uri").equals(UserList.get(
i).getLog().getCs-uri-stem())){
      commonAgent = true;
      break;    }
      }}
      if( !commonAgent){
      User = newUser(rs, ++UserID);
      UserList.add(User);
      }
      else if( rs.getString("refer").equals("-")){
      long t1 =
df.parse(rs.getString("time")).getTime();
      long t2 = 0;
      for(int i=UserList.size()-l ; i>0; i--){
```

```
      t2=df.parse(UserList.get(i).getLog(
).getTime()).getTime();
      if(!rs.getString("uri").equals(UserL
ist.get(i).getLog().getCs_uri_stem())){
      User = newUser(rst,++UserID);
      UserList.add(User);
      break; }
      else {
      if(rs.getString("uri").equals(UserLi
st.get(i).getLog().getCs_uri_stem())&&(
tl-t2 >TIME))      {

      User = newUser(rs,++UserID);
      UserList.add(User);
      break; }
      else
      if(rs.getString("url").equals(UserLi
st.get(i).getLog().getCs_uri_stem())
      && (t1-t2 <= TIME))
      break; }}
      }
      }
      return UserList;
      }
```

Algorithm description is as follows:

When the access record of the user supports Cookie, the session id is cs (Cookie) field, and this article combines with log files and gives consideration to the access habits and amount of information of the page, adds the access time into the heuristic rules [12] as a factor. And the heuristic rules used are as follows:

a. Different IP addresses are different users;

b. If the IP addresses are the same, different browsers are different users;

c. If the IP addresses are the same and so is the browse information, judge according to user access sequence;

When the access record of the user doesn't support Cookie, insert session ID in algorithm to judge.

4. Experiment Result Analysis

This paper experimented on universal method of user identification and IASR user identification algorithm in Windows 7. The experiment included two steps: (1) randomly selected part of the log records from the mass log records of the platform server for data cleaning; (2) had user identification operations. The experiment extracted the log records on September 8, 2015, and after data cleaning, it cleaned the incomplete data records and field information unrelated to mining, and then extracted 50 pieces of log records from the log records after cleaning as the experiment data. The experiment result is as in Table 4-1.

Table 4-1. Experimental Result

method	Total records	Time(ms)	User number
IASR	50	12	24
Common method	50	25	31

Experiment result analysis: total records of experiment were 50 pieces, common method cost 25ms and identified 31 users, but IASR algorithm cost 12ms and identified 24 users. The speed of IASR user identification algorithm is faster than that of common method and the accuracy of user identification is also higher than that of common method.

Table 4-2. IASR Identification Results

User ID	Time	IP	Agent	Session	Uri	Referer
1	00:01:35	156.36.1.125	A			
2	00:04:24	189.221.22.38	B			
3	00:05:19	189.221.22.38	C	S1	/web/	
4	00:05:19	189.221.22.38	C	S1	/web/rt.html	

The front two lines were identified as different new users due to different IP address; the 3rd and 4th line, though with the same IP address, due to different client agent information, were distinguished as different users; the 3rd and 4th line contained 2 users and they had the same session ID, however, session ID was unique and one session could only represent one user, so the two users were actually the same user, but common method couldn't correctly and effectively identify the users that directly input URL information from the address bar of the browser.

5. Conclusions

User identification algorithm is a research hotspot in Web log mining field, and compared with session identification and pattern mining, the importance of user identification is obvious. The IASR user identification algorithm raised in this article is the expansion and improvement of the three heuristic rules, the characteristics and advantages of which are reflected in two aspects:

(1) The improvement of server end application program. URL rewrite involves the modification of server end application program, and under the condition that the server end Web application supports Cookie, employ URL rewrite user tracking method, and even the user browser forbids the user of Cookie, the server end can also correctly identify each user. This is the precondition of ISAR user identification algorithm, and only if each piece of log record contains session ID can the ISAR user identification operation be correctly implemented.

(2) User identification by introduction of session. By the improvement of server end application program, the log records all contain session information. ISAR user identification algorithm expands the three heuristic rules to four rules to judge. When two users have the same IP address and user proxy, it doesn't employ route analysis method, but has session judgment first, and if the session IDs of the two users are the same, judge them as the same user, and if not, have route analysis.

There are still some problems unsolved in this article though improvement has been made, for example, circumstance that one user uses several browsers to access the site and incomplete log records caused by local cache.

Reference

[1] K. Gulati and N. Narender, "Query Recommendation in Hidden Web Search Engine using Web Log Mining Techniques", International Journal of Computer Applications, vol. 102, no. 13, **(2014)**, pp. 6-9.

[2] Y. Yan, "A Practice Guide of Prdicting Resource Consumption in a Web Server", Review of Computer Engingeering Studies, vol. 2, no. 3, **(2015)**, pp. 1-8

[3] W. Fang, "Web data mining application discussion in e-commerce platform", Technology innovation and application, vol. 10, **(2014)**, pp. 44-44.

[4] Y. Yue, J. H. Xiao and S. Y. Luo, "An Overview of Literrature on E-Commerce Customer Loaalty", Review of Computer Engineering Studies, vol. 2, no. 4, **(2015)**, pp. 1-6

[5] Y. Wang, "Web mining-based digital library individual service system research", Information science, vol. 4, no. 19**(2014)**.

[6] R. Li and H. Zhu, "Optimization of session identification algorithm in Web log mining pre-process", Computer knowledge and technology: academic exchange, vol. 5, no. 11, **(2009)**, pp. 8616-8618.

[7] M. S. Chen, J. S. Park and P. S. Yu, "Efficient data mining for path traversal patterns", Knowledge and Data Engineering, IEEE Transactions, vol. 10, no. 2, **(1998)**, pp. 209-221.

[8] D. P. S. U. Maheswari, "A New Clustering and Preprocessing for Web Log Mining", World Congress on Computing and Communication Technologies, **(2014)**, pp. 25-29.

[9] N. M. Khairudin, A. Mustapha and M. H. Ahmad, "Effect of temporal relationships in associative rule mining for web log data", The Scientific World Journal, **(2014)**.

[10] Q. Song and J. Shen, "Efficient and pleuripotent mining algorithm in Web logs", Computer research and development, vol. 38, no. 3, **(2001)**, pp. 328-333.

[11] H. Li, X. Meng and Y. Wu, "Anonymous user identification algorithm research in access mining", **(2014)**.

[12] Z. Li, "Web log mining and association rule-based individual recommendation system model research", Southwest University, **(2014)**.