한국정보전자통신기술학회논문지(jkiiect)'19-6 Vol.12 No.3
논문 19-12-03-231 http://dx.doi.org/10.17661/jkiiect.2019.12.3.231

Research on Data Acquisition Strategy and Its Application in Web Usage Mining

Cong-Lin Ran*, Suck-Tae Joung**

웹 사용 마이닝에서의 데이터 수집 전략과 그 응용에 관한 연구

염종림*, 정석태**

Abstract Web Usage Mining (WUM) is one part of Web mining and also the application of data mining technique. Web mining technology is used to identify and analyze user's access patterns by using web server log data generated by web users when users access web site. So first of all, it is important that the data should be acquired in a reasonable way before applying data mining techniques to discover user access patterns from web log. The main task of data acquisition is to efficiently obtain users' detailed click behavior in the process of users' visiting Web site. This paper mainly focuses on data acquisition stage before the first stage of web usage mining data process with activities like data acquisition strategy and field extraction algorithm. Field extraction algorithm performs the process of separating fields from the single line of the log files, and they are also well used in practical application for a large amount of user data.

요 약 웹 사용 마이닝 (WUM)은 웹 마이닝과 데이터 마이닝 기술의 응용 중의 하나다. 웹 마이닝 기술은 사용자가 웹 사이트에 액세스 할 때 웹 사용자가 생성 한 웹 서버 로그 데이터를 사용하여 사용자의 액세스 패턴을 식별하고 분석하는 데 사용된다. 따라서 우선 데이터 마이닝 기술을 적용하여 웹 로그에서 사용자 액세스 패턴을 발견하기 전에 합리적인 방법으로 데이터를 수집해야 한다. 데이터 수집의 중요한 일은 사용자의 웹 사이트 방문 과정에서 사용자의 자세한 클릭 동작을 효율적으로 얻는 것이다. 이 논문은 주로 데이터 수집 전략 및 필드 추출 알고리즘과 같은 웹 사용 마이닝 데이터 프로세스의 첫 단계 이전의 데이터 수집 단계에 중점을 둔다. 필드 추출 알고리즘은 로그 파일에서 필드를 분리하는 프로세스를 수행하며 대용량의 사용자 데이터에 대한 실제 응용에도 사용된다.

Key Words : Data Acquisition Strategy, Data Processing Flow, Field Extraction Algorithm, User Log, Web Usage Mining

1. Introduction

With the rapid development of the Internet, information technology has been widely used in the service industry, such as education, e-commerce, entertainment. Especially with the development of 5G communication technology and the popularization of multi-intelligent terminals, the number of Internet users in the world reached 4.346 billion, and the penetration rate of the Internet reached 65.8% by March 2019[1].

This research was supported by Education Department of Jiangxi Province in 2017~2019(No. JXJG-17-17-15) and the National Social Science Foundation of China (NO.18XXW011).
*Department of Information Technology Center, Jiujiang University, China
**Corresponding Author : Department of Computer and Software Engineering, Wonkwang University, Korea(stjoung@wku.ac.kr)
Received May 6, 2019 Revised June 10, 2019 Accepted June 11, 2019

Faced with the huge amount of information on the Internet, it's how to get valuable information from it conveniently and accurately, and find information, which users may be interested in, can be actively pushed to users. In order to meet these requirements, the key is to acquire massive and high-latitude log information during users' visit to Web site. After analyzing the characteristics of users' usage, and then push the information to users targeted[2].

Therefore, the acquisition of massive Web Logs and the valuable information mining have become an important research of Web usage mining.

The paper discusses the goals and technology of web usage mining and then designs four data acquisition strategies and one algorithm for field extraction. An application of the strategy is given with programming as well as the methods of web log data storage and analysis of web log data preprocessing that are required for data acquisition.

This paper is organized as follows. Section 2 and 3 discuss some related work and an overview of web usage mining. In section 4 four collection strategies are proposed. Section 5 presents an application of the strategy and field extract algorithm by programming and conclusion is given in section 6.

2. Related work

Web Usage Mining is the application of data mining techniques to discover usage patterns from Web data. Massive user log data implies Web users' interests and other valuable information. Facing with the huge amount of information and the increasing complexity of information in user log recording, people have to seek new technologies and methods for effective acquisition of massive data, its high-dimensional analysis and user access pattern recognition.

In 1996, scholars Chen M S [3], Mannila H [4] and Yan T [5] put forward a new idea of using data mining methods in the field of Web, and initially formed Web data mining, that is, Web mining. It can be seen that Web mining is based on the theoretical research of data mining. Chen MS proposes a data mining method using path association pattern in Web environment. He proposes to transform access log data into maximum forward reference model to filter out effective backward reference, and finally uses frequent association pattern and related algorithm to mine user's reference access path set. Mannila H proposes that user's behavior be regarded as an event, and mining these events to predict the next event, that is, user's next behavior.

The research of web usage mining tools is also the focus of web usage mining research. Recently, some more advanced tools for pattern mining and analysis have emerged. This kind of pattern discovery tools use artificial intelligence, data mining, and information science knowledge to mine knowledge from collected data, such as Web Miner [6] system to discover association rules and sequential patterns from server access logs. Pirolli [7] combines access path pattern, web page type and site topology information to classify pages for user access.

Under the background of large data, the field of artificial intelligence has been paid more and more attention. The emergence of large data processing and analysis platforms. Some statistical analysis tools such as Python and Weka[8], have also been widely used.

3. Web usage mining

According to the classification of data categories used in the mining process, web usage mining is one part of web mining[9]. Web usage mining is to use data mining technology to analyze and process the log data generated in the process of site users accessing web servers, so as to discover the access patterns and interests of web users. This information is potentially useful and understandable unknown information and knowledge for site construction, which can be used to analyze site visits, assist site management and decision support, etc.

Web usage mining is divided into four stages, such as data preprocessing stage, session recognition stage, pattern discovery stage and pattern analysis stage. The process of web log mining is shown in Fig. 1. Each stage adopts different mining techniques according to the tasks to be completed.

Before the data preprocessing stage of web usage mining begins, It's specially important to provide more valuable user log data for data preprocessing as far as possible by using data acquisition strategies.These data are mainly unstructured data and information is disordered. Usually, web log data is stored in a user-specified directory in notepad format.In order to correctly understand the implicit meaning of user's network behavior data, it is necessary to process the acquired log data into data sets that can be used for analysis. Therefore, acquisition and preprocessing of log data are crucial, such as processing missing values, deleting unique attributes of data, feature coding, data standardization, regularization, data dimension reduction, etc.

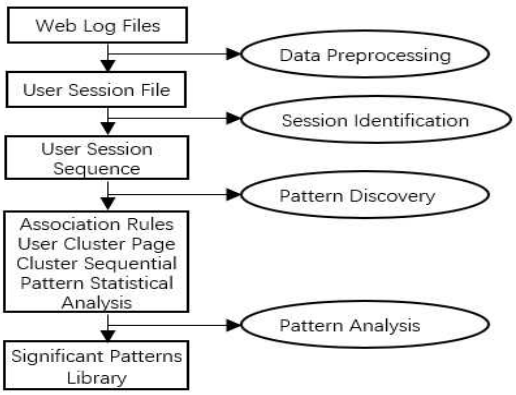

Fig. 1. The process of web log mining

4. Web log data acquisition strategy

Web mining process is similar to data mining process, but the difference is usually only data collection. In traditional data mining, these data are often collected and stored in data warehouse. For Web mining, data collection is a difficult task. For Web mining, data collection is an arduous task, especially in the process of Web structure mining, a large number of Web pages need to be crawled. The data of Web usage mining mainly comes from the behavior data collection of usersvisiting Web sites. The data acquisition process is as follows in Fig. 2. Then, the data processing is archived and stored in the database through different

processing technologies, and the data is prepared for the data preprocessing stage.

Fig. 2. The process of data acquisition

The Web logs data mainly include global UUID, access date, access time, IP address of the server that generates log entries, operations that the client tries to perform, server resources accessed by the client, queries that the client tries to execute, port numbers connected by the client, authenticated user names of the access server, client IP addresses that send requests for server resources. Information such as operating system, browser, operation status code (such as 200, 500, etc.), sub-state, operation status expressed in terms of windows, clicks.

The process of web log data acquisition is a key link in Web usage mining, and it is also a process that every Internet operator must deal with. After collecting data through proper data mining, it will bring qualitative improvement to the overall website operation ability and website optimization, and can truly achieve data analysis and data operation. Usually, Web site analysis data acquisition mainly includes four acquisition strategies: automatic log acquisition, ODBC technology, JavaScript tagging and packet sniffer technology.

4.1 Automatic data acquisition strategy from web server logs

As can be seen from the Fig. 3, the collection of web site analysis data begins when a visitor enters a URL and sends an HTTP request to the web site server. After receiving the request, the web site server will add a record to its Log file, which includes the remote host name (or IP address), login name, login full name, the date of sending the request, the time of sending the request, the details of the request (including the method, address and protocol of the request), the status of the request returned, and the size of the request document. The web server then returns the page to the visitor's browser for display. This automatically acquired log data is stored in the Web application server in notepad format, and then imported into the log storage database through data processing tools for standby.

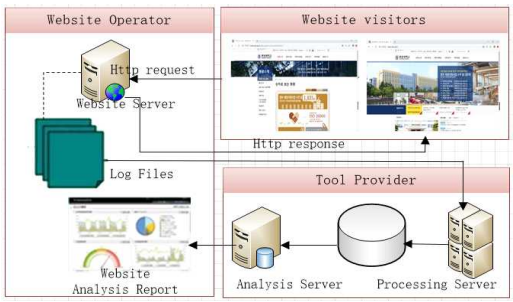

Fig. 3. Data acquisition procedure through web logs

4.2 ODBC log data acquisition strategy

Microsoft Access or SQL Server can be selected to create a database that be used to store log data and then to create tables automatically or manually in the database. Here we choose Internet Information Services (IIS) 7.0 and Microsoft SQL Server 2010 Enterprise Edition to illustrate the implementation of this strategy by using the

way of automatically creating tables in the database to acquire log records. The steps are shown in Fig.4.

Execution Steps
(1). Login to the running SQL Server (User account must have database management access rights)
(2). Open the SQL Server Query Analyzer
(3). On the file menu, click Open to locate in the directory % Windir%\System\32Inetsrv
(4). Save the file after replacing the first line of Logtemp.sql script file in the previous folder with InternetLog inetlog
(5). Choose the database of the InternetLog table to be created. It is generally recommended to rebuild an IISLogData database without choosing the default primary database
(6). Click on the query, and then click on execution to complete

Fig. 4. Execution steps of ODBC log data acquisition Strategy

Another way is to create tables manually. After a system Data Source Name (DSN) is create, Database tables are created with using the file logtemp.sql which the file is in the directory % Windir%\System32\Inetsrv. Then, to open the attributes of the designated site in IIS, select "Enable Logging" in the web Site tab, and select "ODBC Logging" in the "Active Log Format" menu.Last,configure its attributes correctly, click OK, and finally restart the web site.

However, we do not recommend using this ODBC mode to configure IIS log data acquisition strategy. Because IIS servers are very busy, Meanwhile sending log data to the SQL Server database the system resource overhead required by the Web site is very large. That may lead to site discontinuity and inaccessibility, even server downtime, and denial of web site services.

4.3 Javascript buried point data acquisition strategy

This is a common data acquisition method. This method uses the extensible interface defined by Google analysis, and then writes Javascript code to implement custom events and indicators to track data acquisition. The process of buried point data acquisition is shown in Fig. 5.

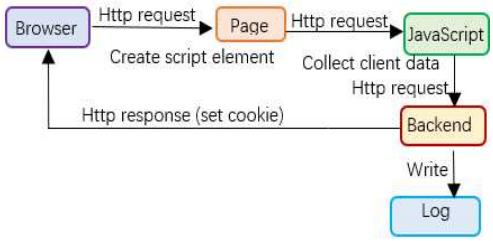

Fig. 5. The process of buried point data acquisition

From the process of this strategy, it can be seen that the user's behavior triggers an Http request from the browser to the statistic page, such as opening the web page. When the page is opened, buried JavaScript fragments in the page are executed. Log acquisition users need to develop JavaScript code added to the web page. This code usually creates a script tag dynamically and points src to a separate js file. At this time, the separate js file (JavaScript in Fig. 5) will be requested and executed by the browser. This js file is the real data collection script. After data acquisition is completed, js requests a back-end data collection script (Backend in Fig. 5). The script file is a dynamic script program disguised as a picture, which may be written by php, Python or another server language. The JS file passes the collected

data to the back-end script through HTTP parameters. The back-end script parses the parameters and records the visits in a fixed format. Question logs may also be used to plant cookies for tracing in http responses. The data collected by this data acquisition strategy is comprehensive, timely and of higher value. In the experimental part, we also adopt this strategy.

4.4 Packet sniffer strategy

The data acquisition process through the packet sniffer strategy is shown in Fig. 6. Packet sniffers can be hardware or software. For example, SNIFF, which acquires user data by monitoring network card ports. If you develop your own software to collect data, you need to be familiar with TCP/IP protocol. If it's hardware, you need to buy the product. This method is easy to deploy and flexible. It takes time and economic cost. The collected data includes not only user's usage log but also other user's network behavior records. The data is comprehensive, but it will occupy a certain amount of network bandwidth, which may affect the speed of users' access to web sites.

As can be seen from the Fig.6, requests made by web site visitors will go through the packet sniffer before they reach the web site server, and then the packet sniffer will send requests to the web site server. The data collected by the packet sniffer is stored in the database after being processed by the tool manufacturer's processing server. Subsequently, web site operators can analyze the report system to view these data.

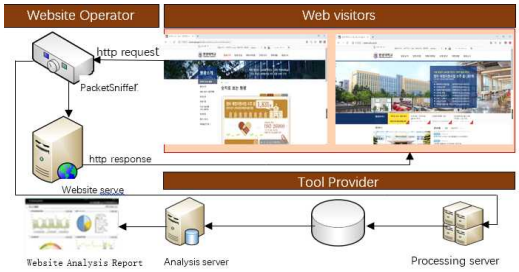

Fig. 6. Data acquisition procedure through packet sniffer

In this section, we propose four data acquisition strategies for user log data acquisition of web server. The automatic data acquisition strategy is a common way to acquire user logs on Web servers. When the website users browse the website, the log records are automatically generated and saved in the web server log directory. The load of web server is very small. However, the log records grow rapidly and occupy a large amount of disk space. Sometimes, it's unsuccessful to store the key log files. And it will be difficult to find a record when encountering problems. The ODBC log data acquisition strategy, which is the simplest way, only needs to configure IIS options and build a database to store logs. Log acquisition information will be automatically saved in the database according to the preset requirements. Using database retrieval function, log records can be quickly searched and filtered. The disadvantage is that the load overhead of the web server is large, which will affect the request and response of IIS. The javascript buried point data acquisition strategy, which acquires the most comprehensive user log records, and can flexibly define the content needed for the acquisition by programming, such as

capturing the user's mouse and keyboard behavior in time and saving it. It can track user interest in the site in real time. We applied this data acquisition strategy in the next section. The packet sniffer strategy can acquire users' network log behavior and web server's access behavior. It needs to purchase hardware support. The cost is relatively high. The hardware is deployed at the network layer and also occupies network bandwidth resources, and network access congestion may occur. Therefore, this log acquisition strategy is rarely used in applications.

5. Application of web log data acquisition strategy

We apply log data acquisition strategy to practical projects. Here we decide to choose a web site of university as the object of diary data acquisition. The university has nearly 50,000 students from all over the world. The Web site is deployed on 64-bit virtual machines in the standard data center room. It consists of one application server, two database servers, three Web site publishing servers and one log server. All operating systems are Windows 2008, and the database is Microsoft SQL Server 2010 Enterprise Edition. This Web site contains more than 80 sub-sites. All Web site pages are stored on the publisher in the form of static pages. The number of users visiting the site is relatively large. The users visiting the site come from all over the world and from different categories, especially during the enrollment period, the number has increased dramatically.

5.1 Log Data Acquisition

In the application, in order to reduce the load of the site and not affect the performance of the site service, we adopt JavaScript buried point data collection strategy to obtain user data.

The first step is buried point setting. A buried point in the page need to be set. That is, insert a piece of JavaScript code fragment. The code snippet of the buried point is shown in Fig. 7.

```
1  <script type="text/javascript">
2      var _maq = _maq || [];
3      _maq.push(['_setAccount', 'zaomianbao']);
4      (function () {
5          var ma = document.createElement('script');
6          ma.type = 'text/javascript';
7          ma.async = true;
8          ma.src = 'http://xxx.xxx.xxx.xxx/ma.js';
9          var s = document.getElementsByTagName('script')[0];
10         s.parentNode.insertBefore(ma, s);
11     })();
12 </script>
```

Fig. 7. Code Snippet of the Buried Point

_ MAQ is a global array for placing various configurations, each of which is formatted as maq. push (['Action','param1','param2',...] The main purpose of the latter anonymous function () is to introduce an external JS file (ma.js) by creating a script through the document.createElement method and pointing the src to the corresponding ma.js according to the protocol (http or https), and finally inserting the element into the DOM tree of the page. For the sake of security, the value of src in this code segment is replaced by X. The expression of ma.async = true is to call external JS file asynchronously, i.e. it does not block the browser's parsing and executes asynchronously after the external JS download is completed.

The second step is front-end data collection. We have written a data collection script (ma. js), which will be executed when requested. The script usually does several things.

(1). Collect information through the built-in JavaScript object in browser, such as page title (document. title), referrer (document. referrer), cookie (document. cookie) and so on.

(2). Parse the _maq array and collect configuration information. This may include user-defined event tracking, business data, etc.

(3). The data collected in the above two steps are parsed and spliced in a predefined format (get request parameters).

(4). Request a back-end script and put the information in the http request parameter to carry to the back-end script.

Part of the JS code of the script file ma.js is given in Fig. 8

```
(function () {
var params = {};
      if(document) {
            params.domain = document.domain || '';
params.url = document.URL || '';
params.title = document.title || '';
params.referrer = document.referrer || '';
              }
------
      var args = '';
for(var i in params) {
            if(args != '') {
                  args += '&';
            }
            args += i + '=' + encodeURIComponent(params[i]);
      }
var img = new Image(1, 1);
    img.src = 'http://xxx.xxx.xxx.xxx/log.gif?' + args;
}
```

Fig. 8. JS Code of the Script File ma.js

The entire script is placed in anonymous functions to ensure that the global environment is not polluted. Whereas log.gif seemingly requests static resources, it is actually an end script in the Nginx.

The third step is to process scripts at the back end.Log.gif is a back-end script, a script masquerading as a gif picture. Backend scripts generally need to do the following things:

(1). Resolve http request parameters to get information.

(2). Obtain some information from the Web server that the client can't get, such as the visitor's ip.

(3). Write the information into log format.

(4). Generate an empty gif format image o f 1x 1 as the response content and set the contenttype of the response header to image/gif.

(5). Set-cookie in the response header to set some cookie information needed.

To collect data by embedding JavaScript code in the page, this method can customize the user's behavior data and obtain user's access behavior flexibly (such as hovering position of mouse, clicking on page components, which search items, total session time, etc.). Then through ajax request to the background log, the information collected in this way will be more comprehensive, for the web site. The optimization and management provide more comprehensive and accurate decision-making basis.

5.2 Log data field extraction algorithm

The acquired log data contains various fields which need to be separate out for the processing. The different information contained in the log file is extracted into the corresponding fields in the database. The server used different characters which work as separators. The most used separator

character is ',' or 'space ' character. The pseudo-code of The FieldExtract algorithm is given below.

The Pseudo-code of the FieldExtract Algorithm

```
Input: Log Files
Output: DB
Begin
1. Open a DB connection (ODBC is commonly used)
2. Create a table to store log data
3. Open Log Files
4. Read all fields contain in Log Files
5. Separate out the Attributes in the string Logs
6. Extract all fields and Add into the Log Table (LT)
7. Close a DB connection and Log Files
End
```

About the execution process of the FieldExtract algorithm, after the database is deployed on the web server, a log record table is created according to the log acquisition requirements, and a database ODBC connection is established in the log acquisition program. The log records needed for batch acquisition are stored in log tables by MSSQL language and regular expressions, and disconnected after completion. This key log record extraction algorithm, which is commonly used in log record acquisition strategies, differs in the types of databases and the delimiters of records in different formats of log records, as well as the SQL statements and regular expressions in corresponding programs.

Usually, the types of web log formats include Apache log format composed of common log format (CLF) and extended log format (ECLF) and W3C log format of IIS. The two normalized format log separators are different, and the programming implementation of log segmentation algorithm is also different. The Field Extract Algorithm is a log file cutting method which adopts different partitioning methods according to different log parameters. In this paper, as an application case, the W3C log of IIS is extracted in the log acquisition strategy.

The Field extraction Algorithm is used to cut web log files in standard format, and is also the basis of other data filtering algorithms. Algorithms can be implemented in different programming languages, and parallel execution programs can be developed for different performance servers.The performance and efficiency of the algorithm are also different. The research on the performance and efficiency of different algorithms will be particularly reflected in other research papers on related algorithms.

5.3 Web log data storage

After the log data is collected and stored in the server, the data is stored in Hadoop's HDFS according to the amount of data. In practical applications, logs are usually generated by multiple servers, including Nginx generated logs, as well as custom formats generated by Log4j in programs. The usual data storage architecture is shown in Fig. 9. We choose this method to store the extracted log data. The three-year log data is nearly 40G, which provides a fast response distributed way for the next preprocessing and cleaning data.

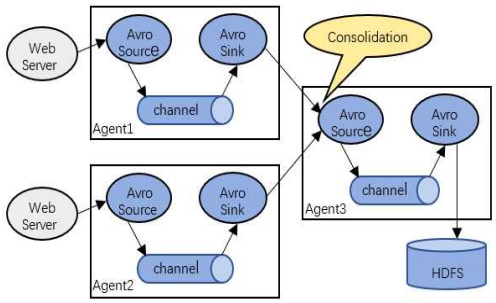

Fig. 9. Data Storage Architecture

5.4 Analysis of web log data preprocessing

We analyze nginx logs by using MapReduce. The default log data format example in Nginx is given, such as,Nginx default log data format, such as:202.68.132.160 - - [18/Sep/2018:09:49:57 +0000] "GET /images/index.jpg HTTP/1.1" 200 19839 "http://www.xxxxx.cn/A00n" "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.3(KHTML, like Gecko) Chrome/29.0.1547.66 Safari/537.36", And we can use MapReduce to analyze nginx log analysis [10].

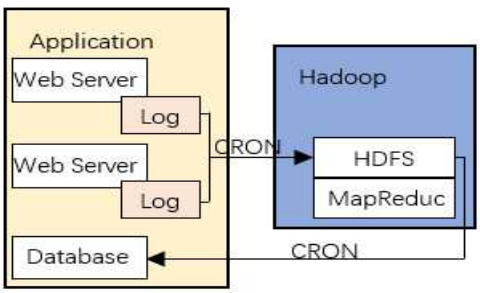

Fig. 10. Data Analysis Model

MapReduce can be used directly for log analysis of the acquired log data. Its analysis model is shown in Fig. 10. Calculated in Hadoop, it can be imported into relational database to display regularly.

6. Conclusion

Data acquisition is an important task of WUM application. Therefore, data must be acquired with appropriate strategies before extracting and cleaning the web logs from web site and then parsing the data and putting it in a relational database or a data warehouse. The data preparation process is often the most time consuming. This paper designs four data acquisition strategies and an algorithm for field extraction and then they have been applied through programming in practice. Practice has proved that using JavaScript technology to customize the strategy of collecting data by burying points, when acquiring a large number of web user logs, this strategy has the least impact on server performance, and the Web site does not have interrupted access. It's easy to obtain user data in time and realize data field extraction. It provides an important data support for discovering user patterns and log visualization from Web logs by using data mining technology in the next step.

REFERENCES

[1] World Internet Users and 2019 Population Stats, https://www.internetworldstats.com/stats.htm 2019.03.

[2] Intelligent Information push-pull Technology, https://baike.baidu.com/item/%E6%99%BA%E8%83%BD%E4%BF%A1%E6%81%AF%E6%8E%A8%E6%8B%89%E6%8A%80%E6%9C%AF/8266146, 2019.03.

[3] M. S. Chen, J. S. Park, K. S. Hong, P. S. Yu, "Efficient Data Mining for Path Traversal Patterns", Proc. of the IEEE International Conference on Knowledge and Data

Engineering, pp. 209-220, March, 1998.
[4] H. Mannila, H. Toivonen, A. I. Verkamo, "Discovery of Frequent Episodes in Event Sequences", Proc. of the IEEE International Conference on Data Mining and Knowledge Discovery, pp. 259-289, 1997.
[5] T. W. Yan, M. Jacobsen, H. G. Molina, U. Dayal, "From User Access Patterns to Dynamic Hypertext Linking", Proc. of 5th Internationl World Wide Web Conference, 1996.
[6] X. F. Xu, "Key Classification Mining Algorithms for Massive Data", pp. 35-51, 2010.
[7] M. L. Liu, X. F. Li, T. Snu, "Survey of Data Mining Technology Standards", Computer Science. Vol. 35, pp. 8-10, 2008.
[8] S. P. Singh, Meenu, "Analysis of Web Site Using Web Log Expert Tool Based on Web Data Mining", Proc. of International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 63-68, March, 2017.
[9] Web mining, https://en.wikipedia.org/wiki/Web_mining, 2019.03.
[10] J. Zhang, Z. Q. Shi, "User Characteristics Analysis Based On Web Log Mining", Proc. of 7th International Conference on BioMedical Engineering and Informatics, pp. 863-866, Oct.,2014.

Author Biography

Cong-Lin Ran **[Member]**

- July. 2004 ~ Sept. 2018 :Jiujiang University, China, Lecturer
- July. 2010 : Huazhong University of Science and Technology, China, MS
- Sept. 2018 ~ current : Department of Computer and Software Engineering, Wonkwang University, Korea, Ph.D. student

〈Research Interests〉 Big Data Processing & Machine Learning

Suck-Tae Joung **[Member]**

- March. 1996 : Computer Engineering of Tsukuba Univ., Japan, MS
- July. 2000 : Computer Engineering of Tsukuba Univ., Japan, PhD
- Feb. 2001 ~ current : Wonkwang Univ., Dept. of Computer and Software Engineering, Professor

〈Research Interests〉 Big Data Processing&Machine Learning, Visual System