Developing A Customized Web Mining System with PHP Language: A Case of Kaohsiung Land Administration Website Data

Chuan-Fu Chuang, Shiuann-Shuoh Chen

Department of Business Administration, National Central University, Taiwan
iamcf.chuang@gmail.com, kenchen@mgt.ncu.edu.tw

Abstract

With the growth of WWW, the unstructured data of web is full of useable information. But these available data are not easy to collect, access, and handle of large scale. Each company, it crawls the available data of site-specific Official or e-Commerce websites and storage huge number of data to progress the analysis of big data. But, here is a difficult question that needs to construct a data mining platform on Web server for companies. In this paper, we use PHP Language to develop a customized web mining system, taking a case of Kaohsiung Land Administration Website Data, defining a novel framework of web mining system, which can reduce cost of constructed platform and enhance efficient of Web crawler.

Keywords: WWW, Data mining platform, PHP, Customized web mining system, Web crawler

1 Introduction

WWW has become the largest public source of real-time information in globe, which is transforming the strategy mode of Enterprises [1]. The web crawling is to filter and collect various full of unstructured data from HTML documents, according to a requirement subject-matter, storage the huge number of information contents into the Excel, DB, or other forms. Because of the huge amount of web information, the performance of data mining is mostly affected by storage framing and the storage database [2]. For this, the web crawling requires great capacity of storage space, corresponding hardware costs and advantageous web crawler algorithms. But, here is a difficult question that needs to construct a data mining platform on Web server for companies.

PHP language is continuing change and continuing growth, it has the function of self-regulation and conservation of organizational stability, that very suite to develop the customized projects system on the web application [3]. In this paper, we use PHP Language to develop a customized web mining system, taking a case of Kaohsiung Land Administration Website Data, defining a novel framework of web mining system. The framework can reduces cost of constructed platform and enhances efficient of web crawler. Thus we select also the famous Excel format to improve the executing efficiency of data mining, resulting in the cost down of software and hardware when increasing the data storage efficient.

In the study, the WMSF has to consider the hybrid clouds design to improve the usability and performance of Web Systems, which is also a problem of Greenness IT [4]. We present the solution of WMSF for crawling the contents of land administration web, and compare the number of Web crawler records, Memory limit and Max execution time of them.

The others of this paper are organized as follows: Section 3, we give a detail description of WMSF and their junction properties. The CWMS profiles of the case are discussed in Section 4 and we discuss the data format of SCVs in section 4.2. Additionally, the further research on the Design and Implement of WMSF, information contents retrieval in the web crawling programing is proposed in Section 5. Finally, the experimental study outcomes of WMSF are given in Section 6.

2 Related Work

Microsoft Excel is a spreadsheet software, it has the intuitive interface, excellent computing and charting tools. In addition to suitable for the data-mining effort, which can also to integrate and dump data from big data into the Excel format. There are many study of academics to use Microsoft Excel where used for data analysis, including Adverse drug events, Aware analysis of antibiotic resistance of general public, Nano-toxicity testing analysis of engineered nanomaterial, Geographic Information explorative analysis, and so on [5-8]. For example, the web crawler component collects the text data of contents posted on the web, and provides the datasets to be used for analysis by storing the collected text data in the databases [9].

*Corresponding Author: Shiuann-Shuoh Chen; E-mail: kenchen@mgt.ncu.edu.tw
DOI: 10.3966/160792642019102006009

The combining of web crawler technology and the text analysis can be developed to an agreeable policy tool for understanding the text information of the node-resource on the WWW. Thus, the web crawler can automatically mining into the link node of HTML documents and index the hidden information in the document node based on keyword definitions [10]. But, in the focused crawling process, the web crawler can set the node and content of filtering webpage, which can also progress the data storage of crawled documents when the depth of crawl completed [11].

The functional architecture of Integrated Framework for keyword-based text data collection and analysis consisting of the web crawler, user interface, data analyzer, and database components as shown in Figure 1 [9].

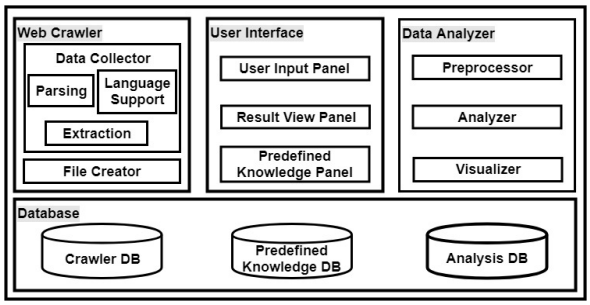

Figure 1. Functional architecture of integrated framework

In the first of functional architecture of Integrated Framework, the web crawler component consists of the data collector and file creator function blocks. The data collector performs parsing, language support, and extraction. Parsing involves searching the webs of articles and crawling the text data of articles on specific webs. Thus, it establishes a uniform resource locator (URL) to search the webs of articles based on the input keyword defined by the user and counts the number of webs of articles using the corresponding URL.

In addition, during parsing, it creates a URL for the webs to crawl the articles based on the number of searched webs of articles. Then, it requests a hypertext mark-up language file of all the articles contained in the web using the corresponding URL. Language support involves encoding in the UTF-8 type to prevent data loss during web crawling.

Extraction involves extracting the text data of the articles from the HTML file. The file creator creates a data file that records the extracted text data and stores the file in the crawling database of the database component. Namely, the crawling database stores the text data of articles collected from the web through the web crawler component [9].

Oh et al. [2] pointed out previous studies are operated by collecting data and by maintaining sessions between servers to facilitate script execution by clients. However, this method has a limitation in that many keywords have to be managed in order to collect from the deep web. So they proposed a solution which is the process of Script execution and address extraction:

(1) Execute the following script from the script queue.

(2) Store new addresses into the collection Page URL stack.

(3) Possible not to visit according to visit depth setting.

(4) Presumed to reach the last page

(5) Load the next address from collection Page URL stack.

(6) Conduct the next script in the script queue.

(7) Judged to arrive on the last page.

(8) Load the following address from collection Page URL stack.

The automatic collection process of deep web records is repeated in the process. Page movement is marked on hypertext reference (HREF) attributes in the anchor tags or behind the on click attributes in an anchor tag. If it is operated as a script, the HREF attribute does not begin with “http://” and has function forms of JavaScript.

Since collected scripts can be executed only in the original documents, users need to manage information where it was collected, and in the way it should be conducted, in order to execute the scripts. The collection Page URL is stored in the collection Page URL stack. The web crawler moves to the stored page and extract scripts. Extracted scripts are preserved in a script queue. They are processed in input order. When pages are extracted from the collection Page URL stack, they are embedded in a browser, and the scripts are executed [2]. The comparison between Oh et al. study and this study are shown in Table 1.

Table 1. The comparison between Oh et al. study and this study

Types	Oh et al. study	This study	Description
Web crawler language	C#	PHP	PHP is a free general-purpose scripting language.
URL algorithm	◯	◯	Automatically collecting data and quickly access
Focused crawler	◯	◯	Primarily collects selected important webpages or themes to collect each webpage data
Storage types	×	Excel	Large scale data storage and stability
Web Mining System Framework	×	◯	Fast creating web mining platform with low cost
Customized Web Mining	×	◯	Unstructured data crawling and improved efficiency
Enter keywords	×	×	Focused crawler has replaced Enter keywords

3 Definition of Web Mining System Framework

Web Mining is a process that obtains information from the source webs through the URL link relationships between Webs and extracts to the entire resource Webs [4]. This process is mainly done by the web crawler called Web Mining System Framework (WMSF) which usually consists of Source Web A, Resource Web B, Data Collector and File Creator, as shown in Figure 2.

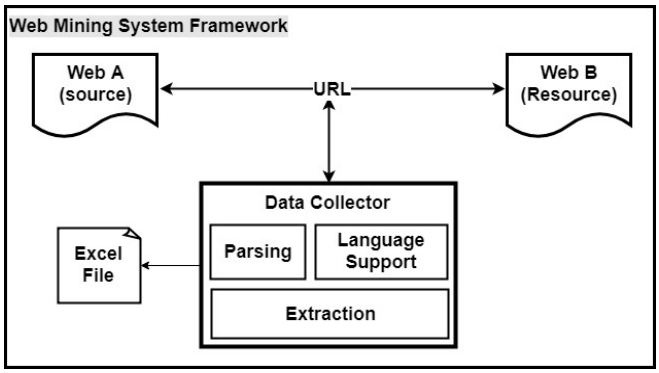

Figure 2. The most economical web mining architecture

The instance of data collector performs parsing, language support, and extraction. Parsing involves searching the webpages of Kaohsiung Land Administration (KLA) and crawling the text data of published records on specific webpages. For this, it specifies a uniform resource locator (URL) to search the webpages of KLA based on the universal resource identifier (URI) defined by the user and computes the number of webpages of records using the corresponding URL.

In addition, during parsing, it creates a node value for the searched data of records to crawl again the new records based on the searched node of records. Then, it requests a hypertext markup language (HTML) file of all the data of records contained in the webpage of KLA using the corresponding URL. Language support involves encoding in the PHP and UTF-8 type to prevent data loss during web crawling. Extraction involves extracting the text data of each node from the HTML file of KLA, and stores directly the data in the Excel components (i.e., in the study replaced a file creator with a MS Excel file). Such as the code of A1 cell is as following:

$objPHPExcel->getActiveSheet()->
setCellValue ('A1', 'Heir name');

The web mining crawls the Resource Web B from the Source Web A, extracts and saves the unstructured data of Web B to specified Excel file. As crawling a web page usually takes seconds time to wait internet communication, the max execution time will be enabled to parallel adjust the corresponding memory size so that improves the web crawler efficiently.

4 Customized Web Mining System

The Customized Web Mining System (CWMS) filters and collects the public resources of Kaohsiung Land Administration Web (KLAW) data in Internet (URL : https://landp.kcg.gov.tw/), which provides a search platform of Web crawler, including mining the data of inherited registration and buildings etc.

4.1 CWMS Profiles

The CWMS uses the Apache server v2.4.18 and PHP v7.0.8 to develop a web crawler. First, loads the component library of Excel, adds the Function of 'PHPExcel' and sets the coordinates of each value. Second, the certain extension Objects of 'file_get_ contents ($get_url)' and 'DOMDocument()'in CWMS are written for 'getElementsByTagName (HTML Node)', run Web mining to filter and collect the unstructured data; Third, use 'foreach (Object A -> Object B)' to gather the useful content we need from the 'getElementsByTagName (HTML Sub-node)' and store to the specified corresponding variables (SCVs); Then, build the grammar vocabulary of PHP and the creating each cell of Object, store the SCVs to the specified cell of Object; at last, create a display webpage of from start to end times with search finished page, including the records quantity of Storage. This system can executes and updates the KLAW information regularly to guarantee the real-time and accuracy of search Excel file.

4.2 Data Format of SCVs

The data in the KLAW is saved to SCVs in the uniform format and then stored to each cell of Excel, such as the case of 34,131 records in the Excel. The format is shown in Figure 3.

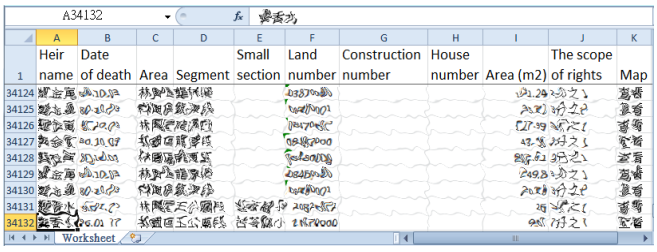

Figure 3. The data format

5 Design and Implement of CWMS

The developing environment of CWMS is the OS Windows 7 (32bit), the Apache server v2.4.18, the PHP v7.0.8 and the RAM 2GB. The Hardware configuration is shown in Table 2.

Table 2. The Hardware configuration

OS	RAM	System type	HDD
Windows 7	2GB	32Bit	500GB

The codes to create a programing file are as following:

1. *Create some code of 'phpExcelClasses' and pre-set first worksheet by the component library of Excel, add an URL of resource Web with variable data type:*

```
require_once "phpExcelClasses/PHPExcel.php";// including the component library
require_once "phpExcelClasses/PHPExcel/IOFactory.php";
$objPHPExcel = new PHPExcel();
$objPHPExcel->setActiveSheetIndex(0);// first worksheet
```

2. *To the document of DOM type, first create a document by Function object 'DOMDocument()' as loaded HTML document:*

```
$dom=new DOMDocument();
@$dom->loadHTML ($get_html); // loaded HTML document
```

3. *DOM Object-oriented create a code by Function object 'getElementsByTagName()', Read the 'table' element of loaded HTML document, write the information of node to variable data type $tables:*

```
$tables=$dom->getElementsByTagName ('table');// get the node contents of 'table'
```

4. *Add a code by Function object 'getElementsByTagName()', Read the 'td' element, write the information of node to variable data type $tds:*

```
$tds=$table->getElementsByTagName ('td'); // get the node contents of 'td'
```

5. *Create a code by Function object 'foreach()', Read each floor of the DOM node '$tables' replies circularly, repeat the code in step 4, add the content of DOM node $td to $tds:*

```
foreach ($tables as $table); // get circularly the node contents of '$table'
```

6. *Create a code by Function object 'foreach()', Read each floor of the DOM node '$tds' replies circularly, then add Switch loop, Compare the value of the expression with the value of case in the structure, while add the node Value to variable data type $val1 to $val11:*

```
foreach ($tds as $td); // get circularly the node contents of '$td'
switch ($i){;// get the value of case
    case 1:
    $val1=trim ($td->nodeValue);// get the content of node 1
    break;
    …
    case 11:
    $val11=trim ($td->nodeValue); // get the content of node 11
    break;
}//switch end
```

7. *Create a code by Function object 'for()' in the Switch loop, Add the '$objPHPExcel' grammar of PHP, then add to corresponding 'setCellValue' cells of 'getActiveSheet' from variable $val1 to $val11:*

```
for ( $j=$count+2 ; $j<=$count+2 ; $j++ ){
    $objPHPExcel->getActiveSheet()->setCellValue ('A'.$j,$val1);
    $objPHPExcel->getActiveSheet()->setCellValue ('B'.$j,$val2);
    …
    $objPHPExcel->getActiveSheet()->setCellValue ('J'.$j,$val10);
    $objPHPExcel->getActiveSheet()->setCellValue ('K'.$j,$val11);
    }
```

First is the "$val1" and "$val2" of SCVs, next is "$val3", which can mean as the filter node of extracts, then the object of "$objPHPExcel" can the storage node content to the cell of Excel.

6 Experimental Study of CWMS

In the following experiments, the CWMS adopts the incremental update algorithm concept of reusability of memory to setup the load-adjusted-profile-oriented [12]. The Max execution time, Memory limit and the quantity of Web crawler records used in this study have been adjusted to get a running time sequences of low-to-lot data of Web mined. In addition, all of the experiments were conducted using the records-based-flag reset mechanism. The experiment was conducted by running Web crawler records quantity 3,000, 9,000, 18,000 and 34,131 for each experimental type. The experimental data and outcome is shown in Table 3, and final experiment finished displaying Web outcome as shown in Figure 4.

Table 3. The experimental data and outcome

Experimental Types	Experimental Parameters			
Web crawler records	3,000	9,000	18,000	34,131
Memory limit	256M		512M	
Max execution time	72,000		96,000	
Running time sequences	288s	958s	1994s	4389s

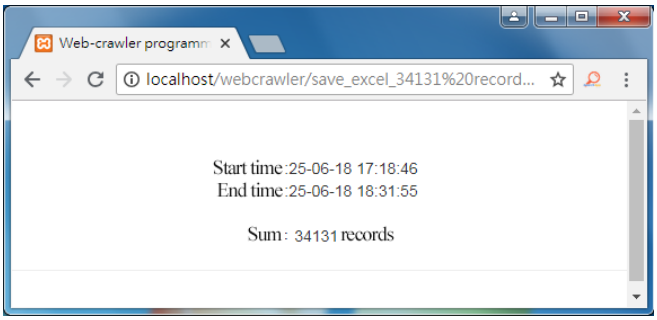

Figure 4. The Web display outcome of finished experiment

We got the time of the 18,000th and 34,131th records, and calculated the time of the 27,000th record as 3,300 seconds by the TREND function of Excel. For this, we have an arithmetic progression of 9,000th, 18,000th and 27,000th. Then we calculated the time of 36,000th, 99,000th and 990,000th records according to the linear trend prediction formula 'y = 16171x – 20258' as 44,426s, 157,623s and 1,758,552s. The Linear trend prediction diagram of Web mining records and times is shown in Figure 5.

Records	Seconds	Hours	Day
9,000 R	958	0.266	0.011
18,000 R	1,994	0.554	0.023
27,000 R	33,300	9.250	0.385
36,000 R	44,426	12.341	0.514
99,000 R	157,623	43.784	1.824
990,000 R	1,758,552	488.487	20.354

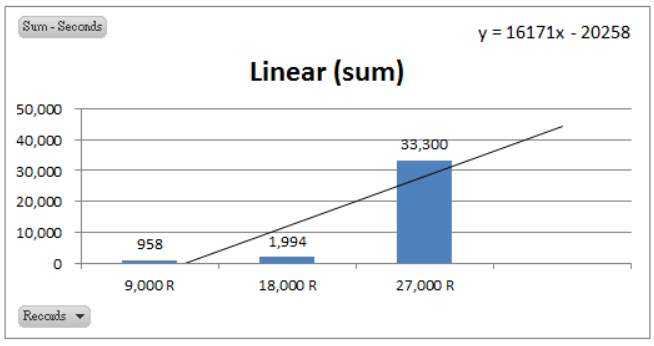

Figure 5. Linear trend prediction diagrams of Web mining records and times

We got the outcome of each running prediction value of CWMS by Comparison of Experimental Results in Figure 5 including the forecast values of Running Time Sequences (RTS) in the future, which can estimating man-hours to understand delivery and calculate the remuneration of worker, so the method of CWMS is feasibly.

7 Conclusions

In this paper, we present a novel framework of web mining system, that effectively exploits PHP Language to develop a customized web mining system which improved the unstructured data crawling efficiency, the automatically collecting data and access speed, the large scale data storage stability, the fast creating web mining platform with low cost.

The web crawler is one of the most important applications of the WWW, whose select of database directly affects the performance of storage and search [10]. Given that the website quality and effectiveness issue should be well considered for developing the website information technology platforms [13]. For example, the quality and self-efficacy of knowledge management systems direct influencing the benefit of enterprises [14]. Thus, the customized web mining system (CWMS) has considered the quality and self-efficacy, and gives a solution of experimental data and outcome. For this, we have come up with a novel architectural design required for implementing such a Web mining system.

The data storage structure and web mining system are designed and Implemented and the data format of SCVs, the most economical configuration of Hardware and the experimental data and outcome are listed in the data structure, design and experimental study of CWMS. Thus, we give the linear trend prediction diagram of Web mining records and times, got also the effective values of experimental result, and verified the feasible of Web Mining System Framework.

The web mining system that we have developed the capability to carry out resource-node-based mining of the data by the source web which can even hosted on a cloud environment with high performance computing infrastructure [15]. In the cloud, the expansions in the memory limit and the enhancements in the max execution time are expected to promote the tasks of various data mining more effectively based on the WMSF. For example, the WMSF can to collect the big data of Internet of Things (IoT), while can according to a requirement node-analysis, storage the information contents into the DB, Excel file or other forms. Moreover, the strategy mode of enterprises adopted gradually the big data of CWMS to analyses trending for future perfection of the data environment and profitability [16], which may be another issue to study.

References

[1] D. Blazquez, J. Domenech, Web Data Mining For Monitoring Business Export Orientation, *Technological and Economic Development of Economy*, Vol. 24, No. 2, pp. 406-428, March, 2018.

[2] H. J. Oh, D. H. Won, C. Kim, S. H. Park, Y. Kim, Design and Implementation of Crawling Algorithm to Collect Deep Web

Information for Web Archiving, *Data Technologies and Applications*, Vol. 52, No. 2, pp. 266-277, June, 2018.

[3] T. Amanatidis, A. Chatzigeorgiou, Studying the Evolution of PHP Web Applications, *Information and Software Technology*, Vol. 72, pp. 48-67, April, 2016.

[4] C. T. Lu, C. S. E. Yeh, Y. C. Wang, C. S. Yang, Hybrid Clouds for Web Systems: Usability and Performance, *Journal of Internet Technology*, Vol. 19, No. 1, pp. 187-195, January, 2018.

[5] S. Iftikhar, M. R. Sarwar, A. Saqib, M. Sarfraz, Causality and Preventability Assessment of Adverse Drug Reactions and Adverse Drug Events of Antibiotics among Hospitalized Patients: A Multicenter, Cross-Sectional Study in Lahore, Pakistan, *PLoS ONE*, Vol. 13, No. 6, pp. 1-18, June, 2018.

[6] T. Mason, C. Trochez, R. Thomas, M. Babar, I. Hesso, R. Kayyali, Knowledge and Awareness of The General Public and Perception of Pharmacists about Antibiotic Resistance, *BMC Public Health*, Vol. 18, No. 711, pp. 1-10, June, 2018.

[7] I. S. Sohal1, K. S. O'Fallon, P. Gaines, P. Demokritou, D. Bello1, Ingested Engineered Nanomaterials: State of Science in Nanotoxicity Testing and Future Research Needs, *Particle and Fibre Toxicology*, Vol. 15, No. 1, pp. 1-31, July, 2018.

[8] L. Savini, S. Tora, A. D. Lorenzo, D. Cioci, F. Monaco, A. Polci, M. Orsini, P. Calistri, A. Conte, A Web Geographic Information System to Share Data and Explorative Analysis Tools: The Application to West Nile Disease in the Mediterranean Basin, *PLOS ONE*, Vol. 13, No. 6, pp. 1-14, June, 2018.

[9] M. Cha, J.-H. Kwon, S.-B. Lee, J. Park, S. Youm, E.-J. Kim, Integrated Framework for Keyword-based Text Data Collection and Analysis, *Sensors and Materials*, Vol. 30, No. 3, pp. 439-445, January, 2018.

[10] H. Hu, Y. Ge, D. Hou, Using Web Crawler Technology for Geo-Events Analysis: A Case Study of the Huangyan Island Incident, *Sustainability*, Vol. 6, No. 4, pp. 1896-1912, April, 2014.

[11] S. Thenmalar, T. V. Geetha., The Modified Concept Based Focused Crawling Using Ontology, *Journal of Web Engineering*, Vol. 13, No. 5, pp. 525-538, November, 2014.

[12] P. Y. Hsu, S. T. Hsieh, Y. C. Chuang, Effective Memory Reusability Based on User Distributions in a Cloud Architecture to Support Manufacturing Ubiquitous Computing, *International Journal of Computer Integrated Manufacturing*, Vol. 30, No. 4-5, pp. 459-471, March, 2017.

[13] W. H. Tsai, W. C. Chou, J. D. Leu, An Effectiveness Evaluation Model for the Web-based Marketing of the Airline Industry, *Expert Systems with Applications*, Vol. 38, No. 12. pp. 15499-15516, November-December, 2011.

[14] S. S. Chen, Y. W. Chuang, P. Y. Chen, Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of KMS Quality, KMS Self-Efficacy, and Organizational Climate, *Knowledge-based Systems*, Vol. 31, pp. 106-118, July, 2012.

[15] V. Thapar, O. P. Gupta, PI^3 Performance Model of Software as a Service (SaaS) Cloud Environment, *International Journal of Advanced Research in Computer Science*, Vol. 8, No. 3, pp. 926-937, March/April, 2017.

[16] C. W. Shen, P. Y. Hsu, Y. T. Peng, The Impact of Data Environment and Profitability on Business Intelligence Adoption, *Lecture Notes in Artificial Intelligence*, Vol. 7197, pp. 185-193, 2012.

Biographies

Chuan-Fu Chuang received his Ph.D. in Business Administration from National Central University, Taiwan in 2019. His research interests include E-business and Big Data. He has been counseling with E.C. Department, COM-WEB since 2012, where he is a Full IT Consultant and currently the Adviser of iPay520 Consolidator.

Shiuann-Shuoh Chen received his Ph.D. degree from Vanderbilt University, USA in 2000. He has been with B.A. Department, NCU since 2006, where he is a Full Assistant Professor and currently the Director of ERP & Big Data Center. His research interests include E-Business System, Business Data Analytics and Innovation Management.