A Cloud-Based Cybermetrics Ranking System

Ssu-An Lo and Chiun-Chieh Hsu

Abstract—The management of university web sites is becoming more critical than before due to the rapid growth of the population dependent on the world wide web as the most important information source. Webometrics Ranking of World Universities (WR) proposed by Centre for Scientific Information and Documentation (CINDOC-CSIC), which ranks the university web sites, has obtained much attention recently. In this paper, we proposed RankCloud as a foreknowledge system for Webometrics Rankings. It adopts web mining to gather the WR indices from the WWW automatically in flexible periods, and uses cloud computing to share useful information in real time for the managers of university web sites. If the web performance of an institution is below the expected position, university authorities should reconsider their web policy, promoting substantial increases of the volume and quality of their electronic publications. Besides, the web site managers may adopt website optimization to promote their WR rankings.

Keywords—Web Mining, Cloud Computing, Website Optimization, Search Engine.

I. INTRODUCTION

JUST as a commercial company needs a well-designed web site to be recognized by customers, a university also needs high-quality and rich-content web pages to present its superiority in academic performance. Webometrics is concerned with measuring aspects of the web: web sites, web pages, hyperlinks, electronic files, electronic papers, technical reports, and web search engine results. The leader of Center for Scientific Information and Documentation (CSIC), Isidro F. Aguillo, by whom Webometrics Rankings of World Universities (WR) [1] is proposed, made statements about academic information on Internet as “Information provided could be useful to increase the commitment to electronic publication, guiding the Web policies and therefore improving Web presence and visibility”. Hence, WR was designed to emphasize the sharing of web resources provided by world universities. By the WR platform, universities in developing countries can show their importance in the same manner as the ones in developed counties. In the world, a large portion of universities consider web site management for WR a critical opportunity to promote their reputation and to appeal future students [9]. Although some people criticize about the ranking value of WR, the indices of WR are commonly recognized as the “e-management” degree of a university. Nevertheless, as far as the authors can concern, most Taiwanese universities start to seek solutions to promote their local and worldwide ranking positions in WR, including some of the Taiwanese top universities. However, CSIC only announces their ranking results in static period (January and July); the ambitious universities can only wait to be informed of their indices scores twice a year.

In this paper, we apply web mining and cloud computing techniques to design a foreknowledge system RankCloud in order to help universities to improve their web site management and hence their WR index scores and ranking positions. RankCloud provides information of WR indices in flexible period by gathering the web contents of the four main search engines the same way as WR does. By the visualized results provided by RankCloud, managers of university web sites may pursue solutions to strengthen their weakness; also they can verify the effectiveness of their efforts immediately, without waiting to be informed in January and July. Finally, the benefits of RankCloud go to all the users on the web, because the web sites will be managed with more care and documented with more sufficient and up-to-date academic data.

The contribution of this paper is mainly in two aspects. Firstly, we adopt web mining and cloud computing techniques to construct RankCloud as a foreknowledge system for Webometrics Rankings. RankCloud adopts web mining to gather the WR indices from the WWW automatically in flexible periods, and uses cloud computing to share useful information in real time for the managers of university web sites. In addition, we also propose heuristic observations such as website traffic, the amount of indexed pages, the amount of back-links, and the prominence of web pages within the keyword search results, which are all considered important by the search engines. Secondly, in view of these criteria, we propose website optimization techniques to help website managers to fulfill these requirements in order to be accepted by web users and the search engines with positive assessments. We have applied the proposed schemes in several university websites and found that, by following the schemes to adjust their website management, the universities indeed promote their positions in the world university ranked lists.

The rest of the paper is organized as follows. Section 2 gives an overview of Webometrics ranking (WR) and the related applications of information processing. Section 3 introduces the proposed system RankCloud. In section 4, the advantages and some experiments are given. Finally, section 5 concludes the paper.

II. BACKGROUND AND RELATED WORK OF INFORMATION PROCESSING

The concept of the proposed system RankCloud follows the steps as: knowing the rules of Webometrics Ranking, retrieving the ranking data in webpages, sharing the collected and value-added data to the clients, providing website optimization
suggestion, and helping the clients to improve their WR ranking positions. Therefore, in this section, we briefly introduce these topics as background in order to give a clearer picture about what service RankCloud provides.

A. Webometrics Ranking

Knowing the essence of WR can help the clients of RankCloud figure out the direction to labor. Webometrics orders about 120,000 universities in the world according to four indicators concerning the academic data shared on university websites, which are size (S, amount of web pages), visibility (V, amount of incoming links), rich files (RF, amount of rich files), and scholar (Sc, amount of scholar documents). The definitions of the four indices of WR are shown in Table 1 [1].

<table>
<thead>
<tr>
<th>Index No.</th>
<th>Index</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>The amount of pages indexed by Google, Yahoo, and Bing</td>
</tr>
<tr>
<td>2</td>
<td>V</td>
<td>The amount of inverse links indexed by Yahoo</td>
</tr>
<tr>
<td>3</td>
<td>RF</td>
<td>The amount of files in pdf, doc, ppt, and pps types indexed by Google, Yahoo, and Bing</td>
</tr>
<tr>
<td>4</td>
<td>Sc</td>
<td>The amount of scholar documents indexed by Google Scholar</td>
</tr>
</tbody>
</table>

The values of the indices can be crawled by giving queries to the corresponding search engines, and then calculate the total score of a website by substituting the values into the formula below:

\[
WR = 0.2 \times \text{Score}(S) + 0.5 \times \text{Score}(V) + 0.15 \times \text{Score}(RF) + 0.15 \times \text{Score}(Sc)
\]  

RankCloud collects information for its clients from the three search engines Google, Yahoo, and Bing; as well as calculates the WR scores (indicator values) of the universities automatically, by using formula and queries. RankCloud utilizes the computation and storage resource of Azure platform such that the data collecting and analysis processes are parallelized and the total time needed to update all the index values of all world universities is within one week. Therefore, RankCloud provides continual ranking information for website managers to check the effectiveness of their website improving actions such that they do not need to wait for the announcements of CINDOC-CISC in six-month period.

B. Web Mining

RankCloud adopts web mining techniques to extract useful information from the search engines, in order to present the results which the website managers are eager to obtain. There are basically three fundamental types of web mining: web content mining, web usage mining, and web structure mining, as depicted in Figure 2 [3]. Web content mining is to facilitate page classification and web information retrieval. The common learning techniques in this kind of mining include decision tree, support vector machine (SVM), Naive Bayes, and K-nearest neighbor classification. This kind of mining also helps automated data extraction from the WWW [5].

Web usage Mining recorded in user log files is helpful for discovering the behavior, the intention, or the interest of individual users. The results of web usage mining form the basis for personalization and recommendation of e-commerce and e-services.

Web structure mining, also known as page template analysis, produces DOM (document object model) tree [6] by the tags embedded in page source codes. The tags are nodes in a DOM tree, and the inclusion relation between each pair of tags is a link from a father node to its child node. Since the pages within a site are interconnected by hyperlinks, mining the web structures can facilitate the improvement of link quality; for example, help to remove improper links linking to pages not existing or rearrange linking structure for convenient browsing [12].

C. Cloud Computing

The services delivered by cloud computing can be classified as (1) Infrastructure as a Service (IaaS), (2) Platform as a Service (PaaS), (3) Software as a service (SaaS). IaaS is on the lowest level of the cloud service stack which provides storage, bandwidth, and computation power as resources. Key examples include Amazon EC2 (Elastic Compute Cloud) [7] and Sun Grid. PaaS provides system management and application development environment such as Google App Engine and Microsoft Azure. The development and deployment on PaaS must follow the rules and limits released by the platform providers. SaaS is on the top level of the cloud service stack and provides commercial or personal services such as Gmail and Microsoft Office Live [8].

There are more and more researches focused on the application and architecture of cloud computing. Some researches discuss scientific computing using the cloud computing. It becomes commonly accepted that cloud computing is a promising technology and paradigm for scientific applications which need powerful computation capacity and non-sustained workload. Cloud computing is also applicable to education [10], enterprise operations [9], and e-commerce [2]. Figure 3 shows the instance of combining the cloud services provided by Google Apps (Gmail, Google Calendar, Google Docs, and Google Sites) to facilitate the knowledge management, marketing strategy, and research & development management of small and medium-sized enterprises [11]. Another example combines Google AppEngine and Google Web Toolkit to develop a Web POS (point of sale) system to speed up the data processing and to achieve data sharing, independent maintenance, and real-time customer services.
RankCloud, which provides WR ranking services on the cloud, has the following features:

1. Access security: since the service can be accessed ubiquitously, RankCloud provides cloud service security mechanisms (CSSM) including IP domain locking, IP access monitoring, user authentication, and connection forced logout.

2. Privacy: the data in RankCloud is processed according to a strict privacy mechanism, e.g., detailed index values can only be accessed by their owners, and top-level values of other university are limited to be used only for comparisons.

3. Data stability: we propose an outlier value detection mechanism (OVDM), by which RankCloud will retrieve the related data from the web twice if the value of an item differs from itself between two retrievals. The items under detection are well defined in proper granularity in order not to recheck them too often.

4. Service Usability: RankCloud is implemented on Microsoft Azure cloud platform, where only few hours were reported out of service in the last few years. In addition, the load of users’ requests is properly parallelized by this platform so that users will not feel speeded down even when the load is heavy.

5. Display Optimization Factor: Page display speed of a website relevant to the users’ first impression, therefore, quick response is essential for a popular website. Materials which take much transportation time should be carefully designed as arranged on deeper level of pages.

D. Website Optimization

Website optimization (WSO) is usually referred to as means by which website managers or webmasters can adopt in order to pursue for improving the visibility of their websites. The basic idea is to follow the indexing rules of search engines such that the amount of indexed URLs (uniform resource locator) of a website can be increased, and the ranked positions of the site in the searched result lists can be improved (nearer to top positions). However, the process of website optimization is complex and time consuming. Website managers need to find out the factors which influence the ranking algorithms of the search engines, to adjust the related parts, and then to check the effectiveness of the adjustments. In addition, the optimization process needs to be handled with care because the search engines may penalize websites which are discovered using black hat methods, which are methods involve deception and disapproved of by the search engines. The penalties may include reducing their rankings or eliminating their listings from the databases altogether.

As mentioned above, the Internet technology drives world universities to find solutions to enhance the visibility of their websites, and website optimization is one of the possible routes. There are six main optimization indicators:

1. Content Optimization Factor: website designers should focus on assuring the content original and attractive in order to allure users regularly revisit their websites. To reach this goal, webmasters and administrative authorities should be generous in sharing genuine valuable materials on their pages, the “word of mouth” effect will naturally occur.

2. Layout Optimization Factor: The fundamental requirement of page configuration is that the web pages should be able to look normal by different browsers, in a variety of resolutions, and so on. Furthermore, the layout optimization of a web page is to arrange visual elements properly such that the users can easily acquire information from the page with joy.

3. Objective Optimization Factor: refers to make the website to reach the planned objective, for example, academic sites to be able to share academic materials through the pages in order to enhance enrollment and donations.

4. User Interface Optimization Factor: Webmasters should carefully consider the design of user interface in order to keep users stay and revisit their sites, thus enlarge the flow and impact of the websites. Basically, user interface optimization includes the smooth operation on a single page and the fluid transition between multiple pages.

5. Display Optimization Factor: Page display speed of a website relevant to the users’ first impression, therefore, quick response is essential for a popular website. Materials which take much transportation time should be carefully designed as arranged on deeper level of pages.

6. Search Engine Optimization Factor: The term Search engine optimization (SEO) comes from the interaction between websites and the search engines. The main purpose of SEO is to adjust the pages on a website to meet the requirements of index rules of search engines such that for specific keywords, the website can be placed in the higher position of the search result lists. Thus increasing the visibility of the site is the ultimate goal of SEO.

III. THE PROPOSED METHOD

A. The Operation of RankCloud

The system diagram of RankCloud is shown in Figure 1. The proposed method RankCloud is developed on Microsoft Windows Azure platform, and the user interface is ASP.net. We use Visual Basic as programming language and select Microsoft SQL Azure as database system. Hence RankCloud is accessible anywhere by any intelligent device such as personal computer, notebook, and cellular phone, only if they can connect to Internet and have browsers.

This engine uses the proposed technique called auto data fill-in techniques to infill the proper scores to WR index tables. The data extraction engine within RankCloud commits the queries to search engines Google, Yahoo and Bing as described in Table 2, then obtains the index scores and integrates the values into the corresponding fields in the WR database tables. This engine uses the proposed technique called auto data fill-in to infill the proper scores to WR indices reports.

We simply fill in the queries to search engine, and the extraction engine will take advantage of URL of the search engine result page in order to automatically produce the source code through RankCloud. Then, the extraction engine tries to
find the beginning and ending of the html tag which takes charge of the search engine result. Finally, the extraction engine may automatically collect the search engine results in the future.

![Cloud-based service system RankCloud](image)

The implementation and operation process of RankCloud follows the six steps: analyze the templates of website pages, produce the rules of page templates, reconstruct the schema of database tables, automatically fill meaningful materials into the database tables, WR Indices scores calculation and analysis, and share WR Indices scores to universities by cloud computing. As the step captions reveal, RankCloud uses page structure mining techniques to perform web page template analysis. With the results, RankCloud produces template rules into its rule base. Then RankCloud executes reverse engineering to define the fields of the database tables. The most important and critical step is the automatic data fill-in process, which extracts meaningful material from the source websites and fill them into the tables. Then, RankCloud calculates and analyzes WR Indices scores to report tabular or graphical outputs to its clients. As the final stage, RankCloud share WR Indices scores to universities by cloud computing.

B. The Proposed RankCloud Framework

We propose three-layered architecture which embedding all the modules of RankCloud services. These layers are user interface layer, logical application layer, and data storage layer:

(1) User Interface Layer: As stated in the previous subsection, the functions of RankCloud interface provide the users valuable information and friendly environment.

(2) Logical Application Layer: In this layer, RankCloud is composed of three components: login authentication component, data collection and computing component, and export component.

(2.1) Login authentication component: This component not only creates user accounts to assure legal usages, but also restrict the access rights only for the authenticated IP domains.

(2.2) Data collection and computing component: This component executes the data collection and computing functions.

(2.3) Export component: This component produces the customized reports as described from previous subsection B. Users can learn useful information from the comparisons and trend diagrams.

(3) Data Storage Layer: RankCloud uses SQL Azure as cloud database for its superiority in data access speed and low development cost. The data storage management of SQL Azure mirrors all data items in the same datacenter and also stores another backup copy in a datacenter other than the local one. This strategy decreases the possibility of data lose and eases the recovery process.

We deploy RankCloud on Azure platform in order to store the tremendous and growing data volume on the cloud, with the benefit which is not necessary to expand our own datacenter to satisfy the need of increasing data amount. For applications which provide information services like RankCloud, cloud storage is an ideal solution because we do not need to deal with the data backup and recovery issues. As our own experience, before we mounted RankCloud onto the cloud, we must keep monitoring the volume of the crawled data and add new disks before the current ones are full.

IV. EXPERIMENTS AND RESULTS

In this section, we detail out the experimental design for evaluating the performance of our proposed RankCloud. Up till now, more than 30 universities among totally 173 universities in Taiwan are using RankCloud to monitor their ranking positions and to acquire diagnoses and advices for website optimization. Hence from the aspect of practical application, these universities have actually approved of the service quality of RankCloud. Furthermore, to verify the performance of RankCloud experimentally, we have conducted experiments to evaluate the efficiency and effectiveness of the proposed system, in order to prove that RankCloud indeed is capable of helping universities to promote their rankings.

We have evaluated the efficiency of RankCloud in collecting the ranking data and computing the indices scores in the current system configure. RankCloud takes about 120 minutes for all the 173 universities in Taiwan, about 30 days for the 6480 universities in Asia, and 60 days for the 20300 universities in the world. Although this speed is much better than the CSIC announcement (twice a year), we believe that the speed can be easily promoted by utilizing more compute and storage power of the cloud utilities.

The current system average usage is about 1000 requests per day, and the average response time in within 3 seconds. We also believe that since the service is cloud based, the service quality will remain stable even when the number of clients extends from 30 to thousands, when the service of RankCloud is adopted by the Chinese universities.

People who need the WR information of their own university can only manually submit queries to the four search engines by the query formats as described in Table 2. The local server column shows the situation before the WR foreknowledge system was moved to the cloud. In this case, although the data collecting and result analysis are automated, their capabilities are still limited by the computation power and the storage space of the local server. In addition, system performance are degraded
when the work load from users becoming heavier. Hence the quality of service will also be lowered. 

*RankCloud* collects the WR index scores of all the world universities with no interruption. By the support of cloud platform of its storage power, we do not need to expand our disk volume all the time to keep up the increase speed of the crawled data. In this manner, *RankCloud* can also provides another function of drawing the trend diagram by the history data of all the universities that users intend to compare with. The length of the history is current time backward to when *RankCloud* was started. Another benefit cloud platform brings is that the availability of *RankCloud* services is lifted by the elastic computation power. The incoming requests from users are responded by forwarding them to the cloud computation units.

We adopt an experimental method to empirically study the effectiveness of applying the proposed schemes to the Taiwanese universities, which are in the top five hundred positions in the WR global ranked list. The observed universities are divided into experimental and control groups, which comprise ten and fifty-five universities respectively. Table 3 shows the results of the two groups by comparing the two announcements in July 2010 and January 2011. We can conclude that for the four indices ranks as well as the total rank, universities in the experimental group indeed made more progress than those in the control group. In other words, the proposed website optimization schemes positively assist the webmasters to improve their WR rankings.

<table>
<thead>
<tr>
<th>Indices Objectives</th>
<th>Amount Total Rank</th>
<th>S Rank</th>
<th>V Rank</th>
<th>RF Rank</th>
<th>SC Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group (Without applying the proposed schemes)</td>
<td>55</td>
<td>300</td>
<td>297</td>
<td>294</td>
<td>222</td>
</tr>
<tr>
<td>Experimental Group (Have applied the proposed schemes)</td>
<td>10</td>
<td>419</td>
<td>351</td>
<td>492</td>
<td>279</td>
</tr>
</tbody>
</table>

V. CONCLUSION

Pursuing higher WR rankings should not be the ultimate goal for a university. It should better be regarded as an important index about the management quality of a website or even a university. *RankCloud* is constructed for the purpose of providing automatic early information for university website managers, and giving them more details and deeper insight of the WR ranking indicators. We employ web mining and cloud computing techniques to construct the foreknowledge system *RankCloud*, which collects the WR indicator values and fills in the values into proper database tables for further display automatically. *RankCloud* facilitates website managers to observe the indicators from multiple points of view and to adopt website optimization to improve the indicator scores and ranking positions. The whole WWW population will finally benefit from the WR ranking improvements of university websites, because in the mean time more high quality academic materials are shared and spread through ubiquitous media – the Internet.

REFERENCES


Ssu-An Lo received the B.E. and M.E. degrees in Information Management from Chinese Culture University (PCCU), Taipei, Taiwan, in 2003 and 2005, respectively. He is a Ph.D. Candidate in Information Management from National Taiwan University of Science and Technology (NTUST) now. His current research interests include information retrieval, web mining, search engine, website optimization, webometrics, internet marketing and Web 2.0.

Chiun-Chieh Hsu is currently a full professor in the department of Information Management at National Taiwan University of Science and Technology. He received his B.E., M.E., and Ph.D. degrees all in Electrical Engineering from National Taiwan University in 1983, 1987, and 1990, respectively. He worked as a software and firmware design engineer in Acer Computer Company from 1983 to 1985. His current research interests include parallel and distributed processing, network, information retrieval, and graph theory.