The Strategy of Personal Customization and Method of Collecting Professional Dynamic Information

Bohan Li1,2, Yongxiang Cai1,*, Shuying Deng1 and Zongyi He1
1School of Geosciences, Yangtze University, Wuhan, China; 2State Key Laboratory of Geo-Information Engineering, Xi’an, China; *Corresponding author email: 100738@yangtzeu.edu.cn

Abstract. The need for personal customization in professional dynamic information collection requires crawling information that is published within specified time durations, extracting accurate and interest-focused information, and assuring the completeness of the information. By combining the technology of jsoup and Lucene as well as the extensibility research on Heritrix, this paper proposes a set of personal customization strategies and methods for professional dynamic information collection: analyze the composition of the URL link to filter out information beyond a specified time window according to different URL types; set up different cleaning templates to extract the attribute items for different styles of websites; establish user personalization blacklist to filter out the information with poor relevance to user interests; apply correction function to improve the robustness of the crawler and guarantee the completeness of the information. Using the collection part of the Dynamic Information Collecting System of Oil and Gas Resources project as a case verification, the results show that the strategies and methods can be used to achieve the three objectives: timeliness; accuracy and interest-focus; and completeness. The proposed strategies and methods may be applied to the construction of an industrial dynamic information collection system.

Keywords: Dynamic information collecting; Personal customization; Web crawler; Deep interest-focus; Information completeness.

1. Introduction
With the rapidly increasing amount of Web-based information, it is increasingly vital to know how to extract precise information from numerous and complicated messages to fulfill users’ demands. "Web information overload" and "resource disorientation" have become almost universal issues for Internet users. To improve the accuracy and the interest-focus of messages and extracted information, compared to the traditional web crawler and crawling strategies, the focused web crawler provides improved performance in response to attempts to limit it under certain conditions. Therefore, it is widely used to collect dynamic industrial information through controlling the crawling scope and the URL that is put in the pending array.

Q.Z.Liu et al. [1] used Heritrix and jsoup to design a network commodity information extraction system. This system can extract Web information, but it cannot manage the crawling to achieve the interest-focused condition. In addition, compared with news pages, the commodity information's metadata are too simple, and the extracted contents do not form a template. YHe et al. [2] added the judging condition of a URL to Heritrix and analyzed the collected content by using Tika to realize the specific domain corpus. Filtering methods are commonly used, but they can only screen out the website URLs that contain the fixed keyword. M.Zhang et al. [3] based their work on the Heritrix framework. They used the IP address of a website to crawl only a certain area of a web page, and
expanded the Heritrix crawler’s Post-processing chain to conduct the limited information collection. This process has stronger practicability when crawling information. To some extent, these studies have helped to reduce the network information overload and improved the accuracy and interest-focus of extracting messages and information. Nevertheless, providing users with a personally customized information collection service is still a widely researched subject [4]. In this paper, we proposed a set of personal customization strategies and methods for professional dynamic information collection by combining the technology of jsoup and Lucene, and we performed extensive research on Heritrix that was focused on meeting users’ demands. Overall, shallow learning method is more effective in small samples [11], where it has faster processing speed. However, shallow learning method usually extracts low-level features only, which lacks adequate information to process facial expression recognition; The original feature information is also partially lost during extraction, since shallow learning method relies on pre-designed features. Moreover, pre-designed features are susceptible to disturbance from uncontrollable factors (illumination, posture, change in occlusion, etc.), resulting in low robustness under natural conditions. These flaws make shallow learning method difficult to generalize the recognition of new face images, creating challenge to apply this method to facial expression recognition.

2. The Key Issues and the Corresponding Strategies in Personal Customization

The aim of users' personal customization is to crawl web information that is published within a certain time scope [5] and offer users the high-quality news in which they are interested. Generally, the users' personal customization requirements include three aspects: information accuracy, preciseness, and completeness. During the professional dynamic information collection process, personal customization is required to fetch information and extract content from the appointed websites and sections with the published time in the precise time duration and make sure that the results are significantly interest-focused and complete. The relevant resolution strategies and methods that are proposed incorporate the following three aspects.

2.1. Time Definition

One of the major aspects in the significantly interest-focused information is the focus on the time scope, which means that the publication time of the collected information should be in the user defined time window.

After analyzing the composition of the URL links, it is noticed that some websites include dates in the content page’s URL like one of the CNPC’s links -http://news.cnpc.com.cn/system/2016/01/23/001576931.shtml. For these URLs, we expand the Post-processing chain of the Heritrix crawler and filter out the URLs beyond the specified time period. This strategy could offer users the information within the specified time window that they set.

For websites without times in the content page’s URL, the time information must be extracted from its content page, and then the disqualified information is abandoned.

2.2. Precise Content Extraction

Another significant aspect in the demand for significantly interest-focused information is the focus on contents. Precise extraction is intended to extract the precise attributes, such as the title, date, text and so on and to form structured information that is closely related to users' interests.

First, most web pages are coded using HTML. Because different websites have different layout styles and the same attribute item in different websites may be programmed using different HTML elements, it becomes harder to extract precise attribute items [7]. Second, the irrelevant messages could still exist even when the information is crawled from the appointed websites and specified sections. Therefore,
how to eliminate that irrelevant information becomes one of the first concerns [8]. Our precise content extraction strategy is as follows.

We analyze the HTML of each website and determine the extraction methods for attribute items, including the title, time, source and so on. Then, we combine them and set up the website cleaning template since different websites have different styles. This will form different corresponding cleaning templates to ensure the accuracy of the extracted content during extraction.

Besides, setting users' personal blacklists is proposed. They can filter out the information with poor relevance to users' interests.

### 2.3. The Integrity of Information Crawling

The completeness of the information refers to making sure that all desired information in the target website and section, which includes the information that is closely related to users' interests and within specified time period, can be efficiently crawled.

In Heritrix, the crawling time of a website can be set. The renewal cycle of each website is different. To make sure that all the desirable information is crawled and to improve the crawling efficiency, it is necessary to figure out the reasonable crawling time of each website. In addition, sometimes, the crawling process may be unexpectedly halted due to various reasons, such as network suspension, the procedure accidentally stopping, and the instability of the target website. Therefore, the automatic detection and correction mechanism needs to be applied to successfully finish the crawling task and ensure the completeness of the obtained information. The information integrity crawling strategy contains the following.

The appropriate crawling time of every website is determined by several exploratory experiments.
1) Set a crawling duration and then crawl the web pages for the same website.
2) Compare the information items between the original website and the crawled local information, and examine whether there is any missed information.
3) Adjust the crawling time to be longer or shorter, repeat step 1 and step 2 several times, and determine the appropriate crawling time parameter for the website that can make sure that no information is lost and the crawling time is as short as possible.

During the crawling process, we set up the automatic detection and correction mechanism to improve the steadiness of the crawling and ensure that the crawler could automatically complete the crawling task.

### 3. Implementation Methods of Partial Strategies

#### 3.1. The Structure of Heritrix

The Crawl Controller controls Heritrix's beginning and ending of the whole fetching task. First, users give it an initial URL-torrent file, and then, it obtains the new URLs from the Frontier. Then, the controller passes the data onto the processor chains through a series of processor management steps. Then, it puts the qualified URLs into the array, and continues the crawling procedures. Each URL occupies an independent thread [9].

The processor chains of Heritrix include the following five chains: the Prefetch processing chain, the Fetch processing chain, the Extractor processing chain, the Write/index processing chain and the Postprocessing chain. The Frontier Scheduler is one of the classes in the Postprocessing chain. It can rejoin the links from the Extractor and perform the next process [10]. To satisfy users’ precise demands for fetching information such as the demands of the appointed websites, sections, or specified time period, we expand the crawling design, and expand the Frontier Scheduler in the processing chain of specific crawling to fetch the specific information for specific websites.

#### 3.2. Crawling Information within a Specified Period

According to the previous time accuracy strategies, the implementation process of crawling information within a specified period is shown in figure 1. First, we analyze the composition of the URL link to decide whether to filter web pages by the time information in the URL or by the extracted time information from the web page. In general, there are two composition types of URLs: one is the
main domain + date information + news number, like http://news.cnpc.com.cn/system/2016/01/23/001576931.shtml; and the other is the main domain + news number, like http://www.china5e.com/news/news-936579-1.html. Second, we crawl and filter out information according to different URL types. The main steps for crawling the information within the user defined time window are as follows.

1) The Heritrix crawler automatically gets the time period specified by the user.
2) Determine if the current website's URLs contain date information. If the answer is yes, go to step 3; otherwise, go to step 6.
3) Filter out those URLs that are outside the specified time window and crawl the remaining web pages. Then, clean them to extract the precise content.
4) Determine whether all the web pages of the site are completed crawled. If the answer is no, go to step 3; otherwise, go to next step.
5) Determine whether all the sites are completed crawled. If the answer is no, go to step 2; otherwise, end the crawling.
6) Crawl the webpage, and extract the time attributes through the cleaning template.
7) Determine whether the time information is outside the specified period. If the answer is no, continue to extract other attribute items such as the title and content using the cleaning template; otherwise, go to next step.
8) Determine whether all web pages of the site are completed crawled. If the answer is no, go to step 6; otherwise, go to step 5.

3.3. Content Extraction Method
To serve users better, the information that is fetched by expanding the interface of Heritrix needs to be further processed and stored in a local database according to the index. Therefore, we need to extract attributes such as the titles, times, sources, and contents from the crawled web pages.

Jsoup provides an API to extract the titles, resources, dates and contexts. For an attribute item that is associated with an attribute label of a class, the org.jsoup.nodes.Element.getElementsByClass(String className) method can be used to fetch this kind of information. For an attribute item that is associated with an attribute label of a title, jsoup offers the org.jsoup.nodes.Document.title() method. If the label has a specific attribute, the org.jsoup.nodes.Element.getElementsByAttributeValue(String key, String value) method can be used. After extracting the attribute items, we can use the Regular Expression to match the expected content.

After getting the extraction methods for attribute items including titles, times, sources, and so on, we combine them and set up the website cleaning template [11-14]. One can refer to reference [11] for the
specific implementations of the different cleaning templates corresponding to different styles of websites. Then, the extracted content's accuracy can be guaranteed.

3.4. Establish User Personalization Blacklist
To increase the degree of users' interest-focus, the strategy of setting users' personal blacklists has been put forward. It can filter out the information with poor relevance to a user's interests. The users' personal blacklist can be put in the corresponding XML file by mainly using the keywords that users are not interested in but are in the crawled information. The blacklist can be used to filter out the information whose title contains the undesired keywords. We also provide users with the functions of adding and modifying the blacklist words.

3.5. Correction Function to Improve the Robustness of the Crawler
The crawling process makes full use of the correction function to improve the robustness of the crawling and ensure that the crawler could automatically complete the task. Before crawling, the controller needs to read the configuration files for the website in a folder. The folders are named using crawling numbers, including the configuration information from the order.xml file and the torrent seed information from the seeds.txt file. If the last crawling was successfully completed, the crawled information will be saved in a folder named Mirror, and a new folder named crawling number+1 with configuration files will be created. If the crawler is accidentally shut down the last time that it ran, a new folder with configuration files cannot be created. Checking the configuration files and the Mirror folder in the highest numbered folder before starting to crawl a website is a good method to avoid crawling interruptions. If there is a Mirror folder in it, it means that there was an accident that occurred in the last crawling and it did not create a new folder. Therefore, a new folder named maximum number+1 needs to be automatically created, and the two configuration files, order.xml and seeds.txt, need to be written in the folder.

The implementation process of the correction function to improve the robustness of the Crawler

**Figure 2.** The implementation process of the correction function to improve the robustness of the Crawler
When the crawling is finished, checking the size of the Mirror folder is a good way to ensure the completeness of the crawled information. If the size of the Mirror folder is 0, which means that the crawling task was accidentally suspended, it needs to automatically crawl the information again.

Using the interface of Heritrix, we rewrite and implement the two functions copyFileForHeritrix(String strSourceFilepath, String strTargetFilepath) and startTask(String taskpath) to improve the robustness of crawling and ensure that the crawler could automatically complete the task that was described above.

4. Experimental Results

We apply the personal customization strategies and methods for professional dynamic information collection that were described above to the Oil and Gas Dynamic Acquisition System of Ministry of Land and Resources project. According to the users' requests, the main target websites include those of the CNPC, Sinopec, China National Offshore, China Energy, and the China Petroleum and Chemical Industry Federation. According to the time accuracy strategies and implementation methods that were previously described, the URLs of each website are analyzed as in Table 1. The URLs of the first three websites contain date information, but the fourth website's URLs do not. The last website's URLs only have the time information of the year and month. If the user-defined time is accurate with respect to the month, we can filter the information by checking the URLs from the first three and the last websites. As for the fourth site, we need to download the page and extract time information from it to filter out the ineligible information. However, if the user-defined time is accurate with respect to the day, only the first three websites can filter information by checking the URLs. As for the last two websites, we need to download the pages, extract the time information from them and then filter out the ineligible information.

Table 1. Composition information of website URLs

| WebSite                       | Year information | Month information | Day information | URL example                                                                 |
|------------------------------|------------------|-------------------|-----------------|-----------------------------------------------------------------------------|
| CNPC                         | √                | √                 | √               | http://news.cnpc.com.cn/system/2016/01/23/001576931.shtml                   |
| Sinopec                      | √                | √                 | √               | http://www.sinopecnews.com.cn/news/content/2016-01/22/content_1584779.shtml |
| CNOOC                        | √                | √                 | √               | http://www.enoc.com.cn/art/2016/1/20/art_261_2266941.html                  |
| China Energy                 | ×                | ×                 | ×               | http://www.china5e.com/news/news-936579-1.html                             |
| China Petroleum and Chemical Industry Federation | √ | √ | × | http://www.cpcia.org.cn/html/19/20163/153504.html |

In terms of the strategy of setting users' personal blacklists, we find some keywords from the information that have little relevance to users' needs for the target websites and sections and generate a blacklist. For example, we set “Polyethylene”, “compound fertilizer”, and “Plastic” in the personal blacklist. If the title contains any word in the blacklist, this message will be filtered out before being stored in Database.

We carry out some experiments to determine the crawling time for each target website. Table 2 shows the results from the application examples. Websites have different update cycles. Some are updated once a week and others are updated daily. We set different crawling durations for each website, conduct the experiment in each update cycle, and compare the updated news items that were crawled from the website to the actual updated items that were counted manually from the website to check if
any item failed to be crawled, which helps to get the appropriate crawling duration parameters for each website. The results are shown in Table 3. The number of news items that are crawled with the different crawling times turns out to be the same in each group, and the new information that is published in the update period of the original website can be crawled to local the database with few missed items. Those missed news are found to be filtered by the blacklist as our tracking analysis. In other words, when we set the crawling time as 3–5 minutes, all the qualified information in the update cycle in the original website can be crawled to the local database.

Table 2. Results of crawling by setting different crawling durations

| Group | Testing time | Target website | Target section | Crawling Time (min) | Crawled information items | Updated information items | Old information items | Missed items |
|-------|--------------|----------------|----------------|--------------------|--------------------------|--------------------------|----------------------|--------------|
| 1     | 1/13/2016    | CNPC           | Weekly         | 5                  | 3                        | 18                       | 5                    | 13           |
|       |              |                |                |                    |                          |                          |                      |              |
| 2     | 1/23/2016    | CNOOC          | Daily          | 9                  | 3                        | 38                       | 9                    | 29           |
|       |              |                |                |                    |                          |                          |                      |              |
| 3     | 1/14/2016    | China Petroleum and Chemical Industry Federation | Weekly         | 7                  | 3                        | 7                        | 7                    | 0            |
| 4     | 1/23/2016    | CNOOC          | Weekly         | 7                  | 3                        | 6                        | 6                    | 0            |
|       |              |                |                |                    |                          |                          |                      |              |
| 5     | 1/14/2016    | China Petroleum and Chemical Industry Federation | Weekly         | 8                  | 3                        | 8                        | 8                    | 0            |
| 6     | 1/23/2016    | CNOOC          | Weekly         | 4                  | 3                        | 9                        | 4                    | 0            |
|       |              |                |                |                    |                          |                          |                      |              |
| 7     | 1/14/2016    | China Petroleum and Chemical Industry Federation | Weekly focus   | 13                 | 3                        | 222                       | 12                   | 210          |
| 8     | 1/22/2016    | China Energy   | Weekly focus   | 15                 | 5                        | 206                       | 14                   | 191          |
| 9     | 1/23/2016    | CNOOC          | Weekly         | 11                 | 3                        | 213                       | 11                   | 210          |

To comprehensively verify those personal customization strategies and methods in professional dynamic information collection, we set the time interval update in the corresponding programs, as shown in Table 3. All the previous proposed strategies and methods are applied to the Dynamic Information Collecting System of Oil and Gas Resources project.

Taking the CNOOC website as an example, the user defined time window is set to from Feb. 23th, 2016 to Feb. 29th, 2016. In accordance with the previous experimental results, the crawling time of each website is set to 3 minutes. We track the missed news item and find the following news title:

The continued weakness in demand led to a slump in polyethylene prices in Europe.
The results show that the personal customization strategies and methods can be used to achieve the three objectives that were previously described: crawling information within a specified time window, extracting accurate and interest-focused information, and assuring messages' completeness. The proposed strategies and methods are universal and it is feasible to apply them to different industry’s dynamic information collection.

**Table 3.** Results of crawling by setting the time interval in corresponding programs

| Group | Testing time | Target website | Update period | Information items in original website | Target section | Crawling time (min) | Crawled information items | Updated information items | Old information items | Missed items |
|-------|--------------|----------------|---------------|---------------------------------------|----------------|---------------------|--------------------------|--------------------------|-----------------------|--------------|
| 1     | 2/29/2016    | CNPC           | Weekly        | 32                                     | Exploration & Development | 3               | 32                       | 32                       | 0                       | 0            |
| 2     | 2/29/2016    | CNOOC          | Weekly        | 17                                     | Exploration & Development | 3               | 17                       | 17                       | 0                       | 0            |
| 3     | 2/29/2016    | Weekly         | Foreign coope ration | 8                                     | 3               | 7                   | 7                        | 0                       | 0                       | 1            |
| 4     | 2/29/2016    | China Petroleum and Chemical Industry Federation | Daily | 16                                     | Weekly focus | 3               | 16                       | 16                       | 0                       | 0            |
| 5     | 2/29/2016    | China Energy   | Weekly        | 24                                     | Weekly focus | 3               | 24                       | 24                       | 0                       | 0            |

5. Conclusions

Based on the users demands for significantly interest-focused and complete information, we proposed a set of personal customization strategies and methods to achieve the three objectives: crawling information within a specified time window, extracting accurate and interest-focused information, and assuring messages' completeness. The strategies and methods are applied in the Dynamic Information Collecting System of Oil and Gas Resources project. The results demonstrate the effectiveness of the proposed strategies and methods, which can be applied to the construction of an industrial dynamic information collection system.

Acknowledgments

The research is supported by the State Key Laboratory of Geo-Information Engineering of China
(Grant No. SKLGIE2017-M-4-6) and the National Natural Science Foundation of China(GrantNo.41701537).

References

[1] Liu.Q.Z, Yu.Z.L, Design and Implementation of Web Information Extraction System Based on Heritrix and jsoup, Journal of Shandong Normal University(Natural Science) 2015,30(2),16-19.

[2] YHE.G, PAN.Q, Design and Implementation of Specific Data Acquisition System Based on Heritrix and Focused Crawler, Telecom Market 2013,6,127-132.

[3] Zhang.M, Sun.M, Design And Implementation Of Qualified Spider Based On Heritrix, Computer Applications and Software 2013,30(4),33-35.

[4] Wang.H.M, Research And Implement Of Web Information Intelligence Collection And Personalized Service System, He Bei University of Technology,2012.

[5] Yao.B.Y, The Design and Implementation of Personnalized News Fetching and Polymerization System, University of Chinese Academy of Sciences(College of Engineering and Information Technology),2014.

[6] Zhu.M, Luo.S.X, Research of a Focused Crawler to Specific Topic Based on Heritrix, Computer Technology And Development 2012,2(2),65-68.

[7] Mei.L, Research On Object Extraction Based On The Object-Level Vertical Search In Specific Field, University Of Electronic Science And Technology Of CHINA,2015.

[8] Liang Yongchun, JIAO Wenqiang , TIAN Liqin, Design and Implementation of Network Public Opinion Monitoring Based on Big Data News Website Text Mining, Journal of North China Institute of Science and Technology 2018 15(4),82-92.

[9] Liu.G.J, Xia.J.L, Research and Application of Web Crawler Based on Heritrix, Software Guide 2013, 12(5),123-125.

[10] Zhou.J.C, Hu.H.P, Yue.H, Design and Implementation of Lucene-based Full-text Retrieval System, Computer Engineering and Science 2015, 37(2),252-256.

[11] Gao.W.F, The Design and Implementation of the Heritrix-based Topic Web Crawlers, Journal Of Nanning Polytechnic 2011, 16(1),97-100.

[12] Li.K, Cai.Y.X, Gan.J.L, Wang.J.Y, Yang.D, Hu.S.Y, Key Problems and Their Solutions for Industrial Dynamic Information Collection System, Journal of Surveying and Mapping 2016,3,61-166.

[13] Xu.C.F, Wang.Y.C, Zhai.H.Y, Automatic Extraction System of Webpage Information, Journal of Changchun University of Science and Technology 2015,2,151-154.

[14] An Yi, Research on Analysis of Emotional Tendency Based on Tieba Text, Southwest Jiaotong University Master Degree Thesis,2018.