2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics

The Application of Web Crawler in City Image Research

Jiangying Xu
School of Information Science and Engineering
University of Jinan
Jinan, China
e-mail: xjyujn@qq.com

Lixin Du
School of Information Science and Engineering
University of Jinan
Jinan, China
Corresponding author, e-mail: du.lixin@163.com

Chunsun Duan
School of Information Science and Engineering
University of Jinan
Jinan, China
e-mail: ise_duancs@ujn.edu.com

Mingyue Li
School of Information Science and Engineering
University of Jinan
Jinan, China
e-mail: lmyujn@qq.com

Abstract—**Grasping the city image accurately is always a challenge for city manager. The main work of our study is to capture the web pages of a specific city using web crawler technology, process and analyze the captured data to get keywords representing the city. We propose a distributed crawler architecture and collect city-related data based on the architecture. Subsequently, several example specific cities are crawled successfully. The crawled data are analyzed by using data preprocessing, data analysis, and other methods. The results show that the presented approach will be very useful for city manager.**

Keywords-city image; web crawler; data preprocessing

I. INTRODUCTION

City image is the general public's impression of a city, including political image, economic image, humanistic image and geographical image [1]. Better city image is important for future of city development. A good city image can not only enhance the sense of honor, belonging, and responsibility of city residents, but also enhance the attractiveness of the city. For a city manager, a good city image can also greatly enhance the city's overall competitiveness. With the development of the Internet and the rise of new media, the influence of traditional paper media and television is declining. However, with the dramatic increase in the number of speeches and videos on online media platforms, the influence of Internet public opinion on city image is growing. By analyzing the network big data related to a specific city, we can obtain keywords representing that city, thereby helping city managers better grasp the image of the city.

Web crawler is a program that crawls web page information from the Internet according to a specific web search strategy automatically [2-4]. It is widely used in Internet search engines and applications that require big data scraping. Currently, the mainstream search engines are Google, Bing, Yahoo, Baidu, etc. Using a search engine, we can collect and retrieve web page information for a given keyword easily. At the same time, some well-known Chinese Q & A and social networking sites, such as Zhihu and Douban, have lots of city-related information. Collecting relevant web pages and discussion information of a specific city through a search engine is a simple and efficient way for the study of city's image.

II. RELATED WORKS

There has some research work on city image in recent years. Al-ghamdi et al. discussed the image of a city in light of new technology evolution and develop a framework to study image of the city in the information age [5]. Riza M et al. aimed to discuss the influence of iconic architecture through creating identifiable images on quality of life [6]. Adamus-Matuszyńska et al. presented a novel approach for supporting decisions made by local authorities when preparing and implementing a city image improvement strategy [7]. In terms of city image monitoring, Chen et al. designed and implemented a city network image detection system. By collecting public information and using a series of steps such as text segmentation, part-of-speech tagging, and syntax analysis, the target city's network image can be evaluated accurately [8]. Wong et al. used text-mining techniques to investigate the evolution of the content of TripAdvisor online reviews about Macau between 2005 and 2013 [9]. Ali et al. proposed fuzzy ontology-based sentiment analysis and semantic web rule language rule-based decision-making to monitor transportation activities, and to make a city- feature polarity map for travelers [10]. Huang et al. proposed the concept of city fashion image, and compared and analyzed the fashion images of 37 cities based on big data of online news [11].

In web crawler research area, Shrivastava presented a systematic study of the web crawler, include characteristics of web crawler, policies used by web crawlers, general architecture of web crawler and so on [4]. Patil et al. designed an appropriate system for efficiently deep web harvesting, a two-stage enhanced web crawler framework is proposed for efficiently harvesting web interfaces [12].

978-1-7281-6024-5/20/$31.00 ©2020 IEEE

Sobrinho et al. introduced a process that can address elementary data from public user profile in Instagram. These data can be scraped and loaded into a database by the web crawler [13]. Fan et al. proposed a distributed crawler system. The results show that method is more efficient and stable than the single-machine web crawler system [15]. Bahrami et al. introduced a scalable web crawler that employed cloud computing technology which is implemented in Windows Azure Cloud Platform [14]. Amudha et al. described about different types of web crawler and the policies used in the web crawlers. They think the modification and extension techniques in web crawling are the topics of future research [16].

Summarizing the above, the main research works of city image are focused on city image detection, city image communication, city image building, etc. Some researchers have begun to use data mining, natural language processing and other technologies to extract information related to the image of the city for network image detection. The results of city image research using web crawlers are scarce.

III. Model and Architecture

A. *Problem Model*

The research problem in our study can be formalized as follows.

- Let c be a designated city.
- $K = \{k_1, k_2, \ldots, k_n\}$, K is the keyword set for city c. Each k_i is a keyword for searching.
- D is a collection of data such as web pages and messages on the Internet.
- $D_i = f(k_i, D) = \{d_0, d_1, \ldots, d_m\}$, $i \in [1, n]$, $j \in [1, m]$, $D_i \subseteq D$. $\forall d_j$, k_i exits in d_j. D_i is a searching result web pages set by keywords k_i. f is a function whose function is to take k_i as input and get the result webpage set D_i related to k_i.
- $D' = D_0 \cup D_1 \cup \ldots \cup D_n = \bigcup_{i=1}^{n} f(k_i, D)$, D' is the set of all searching result web pages for city c.
- $K' = g(D') = \{k'_1, k'_2, \ldots, k'_s\}$, g is a function that is responsible for deduplication, word segmentation, and clustering of data in the search page set D', and finally obtains a keyword set that can represent a specific city c.

In our research, determining how to solve f and g are our goals and K' is the end results.

B. *System Architecture*

As shown in Figure 1, the system framework is composed of four modules, namely data collection, data preprocessing, data analysis and data visualization.

1) Data Collection Module: The data collection module is responsible for collecting data on the website. Our crawler collect data from Google search engine, Baidu search engine, Zhihu Q & A community, etc. Each search engine has its own characteristics, and variant webpage structure. Therefore, it is necessary to design a corresponding crawling strategy for a specific website.

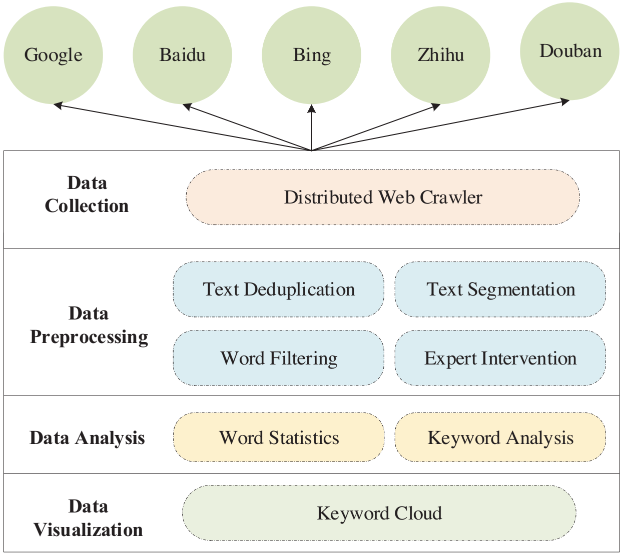


Figure 1. System Architecture.

The web crawlers must solve several problems: (1) there are a large number of web pages to be crawled; (2) many tasks are decomposed for each crawler executing; (3) the target websites always have anti-crawling mechanisms. Therefore, a single web crawler cannot achieve fully satisfactory results. In order to improve the crawler's efficiency and accuracy, we propose a distributed crawler architecture as shown in Figure 2.

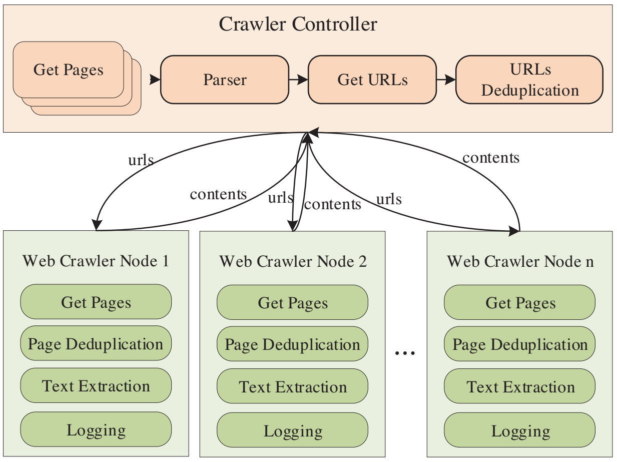


Figure 2. Distributed Web Crawler Architecture.

The processing flow of the crawler controller is as follows. First, receive the keywords entered by the user and the specified number of pages to be queried; Second, start multiple threads to search for relevant content by keywords from Google search engine, Baidu search engine, Zhihu, etc., and get results from Get Pages Module; Third, call the Parser Module for page parsing; Fourth, obtain the target uniform resource locator (URL) queue through the get URLs module and the URLs deduplication module; The end, the controller distribute URLs to multiple machines as tasks to execute.

The crawler controller is responsible for scheduling all crawler nodes and distributing the links to be crawled in order. At the same time, it receives web page content from each machine and stores these contents in a database. Suppose we have a list of 1,000 links waiting to be crawled, and there are ten crawler nodes. The role of the crawler controller is to allocate thousand tasks to ten crawler nodes according to some mechanism, and receive information from each crawler nodes. The web crawler algorithm of each crawler nodes is as follows.

Algorithm1: Web Crawler Algorithm.

step 1. Receive URLs and store them in the local URL queue.

step 2. Determine if the URL queue is empty. If it is empty, the algorithm ends, otherwise go to step 3.

step 3. Take a URL from the URL queue, wait a few seconds, and then visit the page.

step 4. Determine if the page is returned successfully. If it returns successful, go to step 5. Otherwise, record the error log and go to step 2.

step 5. Deduplication of pages using SimHash [17] algorithm. If the webpage does not repeat, go to step 6, otherwise go to step 2.

step 6. Extracting text content using a line block distribution function, and send the text content, source website, title and other information to the crawler controller, then go to step 2.

2) Data Preprocessing Module: The data preprocessing module is responsible for preprocessing the collected data. The processing flow includes text deduplication, text segmentation, word filtering, and expert intervention. First of all, compare the records stored in the database. If there are multiple records whose text content, source website, and other information are exactly the same, it is determined that there is duplicate text. We will delete the redundant duplicate records and only retain one record that can be queried; The next, use the word segmentation tool to divide the content in the database into several words, and write them to the specified file in an appended form; Moreover, the results of the segmentation are filtered. In our study, we use HIT 's stop list and custom list to remove stop words and some meaningless words in the text, such as “2019”, “city” and “today”. Finally, use manual intervention to remove meaningless words again to ensure the accuracy of the results.

3) Data Analysis Module: The data analysis module is responsible for analyzing the preprocessed data. It can be divided into two parts: word statistics and keyword analysis. The word statistics part counts the frequency of all words in the word segmentation result file, and writes the words to the new file according to the word frequency from high to low. The keyword analysis part applies the Latent Dirichlet Allocation [18] model to classify all keywords by topic, and write the classification results into a new file.

4) Data Visualization Module: The data visualization module is responsible for data visualization. In our research, we use a keyword cloud for visual display, which is formed by reading word frequency files.

IV. Experimental Evaluation

A. Experiment Setup

In order to verify efficiency of method described above, we implemented the above crawler program in python language. We deploy the crawler controller on the server and use 5 machines as the web crawler nodes. The development environment is Anaconda, operation system is Windows 10. Each machine has 64GB RAM with an Intel i5-8400 CPU.

We take Hangzhou, Jinan and Qingdao as examples to experiment with the proposed method. In order to avoid the anti-crawling mechanism of the website and not hinder he normal access of the target website, the crawler’s wait time is set to two seconds.

B. Result and Evaluation

Taking Google search results as an example, Table I shows the top ten words and words frequencies in the original statistical results of the three cities. This result has not removed stop words and meaningless words.

TABLE I. Original Statistical Results

City Rank	Hangzhou		Jinan		Qingdao	
	keyword	*count*	*keyword*	*count*	*keyword*	*count*
Top 1	Hangzhou	1205	Jinan	2214	Qingdao	2534
Top 2	China	858	China	1459	China	2157
Top 3	hangzhou	587	Shandong	852	Was	967
Top 4	Hotel	495	2019	623	Tsingtao	846
Top 5	2019	350	jinan	506	2019	761
Top 6	Com	276	10	477	Chinese	728
Top 7	City	275	Was	450	qingdao	694
Top 8	Chinese	273	Hotel	398	10	618
Top 9	More	226	00	374	City	552
Top 10	10	222	city	349	Beer	533

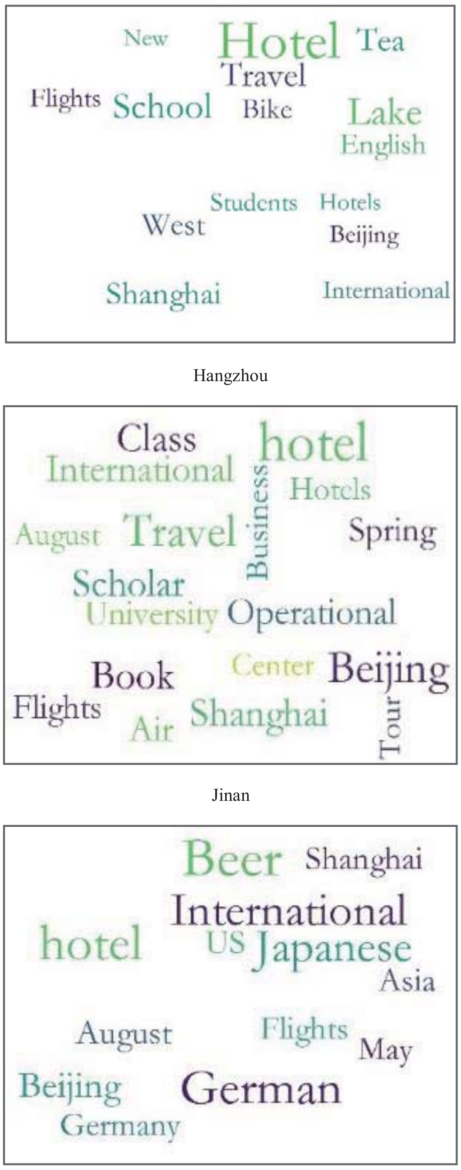
It can be seen from Table I that in the word frequency statistics without word filtering, there are a lot of meaningless words in the top 10 words, such as "Hangzhou", "10" and "0". The results after word filtering and artificial filtering are shown in Table II.

TABLE II. Final Statistical Results

City Rank	Hangzhou		Jinan		Qingdao	
	keyword	*count*	*keyword*	*count*	*keyword*	*count*
Top 1	Hotel	495	hotel	398	Beer	533
Top 2	Lake	213	Travel	251	hotel	463
Top 3	School	186	Beijing	238	German	425
Top 4	Travel	157	Class	187	International	372
Top 5	Tea	150	Book	186	Japanese	358
Top 6	Shanghai	140	Shanghai	182	Beijing	222
Top 7	West	134	Scholar	178	August	184

City Rank	Hangzhou		Jinan		Qingdao	
	keyword	*count*	*keyword*	*count*	*keyword*	*count*
Top 8	English	113	International	168	US	181
Top 9	Bike	111	Operational	159	Asia	177
Top 10	Flights	96	Air	155	Shanghai	170

Figure 3 shows the keyword cloud results for Hangzhou, Jinan, and Qingdao. The size of keywords in the keyword cloud indicates the most frequency of words.



Hangzhou

Jinan

Qingdao

Figure 3. Keyword Cloud of Three Cities.

From the results in Table II and Figure 3, we can conclude that the keywords of the three cities are significantly different. The keywords related to Hangzhou are mainly "Lake", "Travel", "Tea", etc. The public's image of Hangzhou is concentrated in tea, attractions, and tourism; The keywords related to Jinan are mainly "hotel", "travel" and other words. The public's image of Jinan focuses on tourism, education and air; The keywords related to Qingdao are mainly "Beer", "hotel", "International", etc. The public's image of Qingdao focuses on beer, tourism, and internationalization. These results are in line with our opinion, and the results of Jinan have been used in the evaluation of the image of Jinan by the Jinan city manager.

V. Conclusions and Future Work

In our study, we use the web crawler technology to capture and analyze city-related information in online media platforms. The processing flow of the program includes data collection, data preprocessing, data analysis, and data visualization. From experiment results, our method can capture the relevant information of the city well and obtain the city keyword cloud by analyzing this information. This result is helpful for city manager to grasp the image of the city. Our research results have been adopted by Jinan. In future work, we will build a source map of words to analyze the source of each word, and consider a more intuitive form of data visualization.

Acknowledgment

This work is funded by the 2019 Big Data Social Service Special Project of University of Jinan (No. XKY1927 and No. XKY1932), and the Science and Technology Program of University of Jinan (No. XKY1811).

References

[1] Z. Li, and Y.Wang, “Brief Analysis of City Image,” Art Panorama, 2018, pp. 130-131. doi: 10.3969/j.issn.1002-2953.2018.03.047.

[2] X. Zhang, and M. Xian, “Optimization of distributed crawler under Hadoop,” *MATEC Web of Conferences*, Vol. 22, p. 02029. EDP Sciences, 2015, doi: 10.1051/matecconf/20152202029.

[3] V. Shkapenyuk, and T. Suel, “Design and implementation of a high-performance distributed web crawler,” *Proceedings 18th International Conference on Data Engineering*, IEEE, Feb. 2002, pp. 357-368. doi: 10.1109/ICDE.2002.994750.

[4] V. Shrivastava, “A Methodical Study of Web Crawler,” *Vandana Shrivastava Journal of Engineering Research and Application*, Vol. 8, Issue 11, Nov, 2018, pp. 01-08, doi: 10.9790/9622-0811010108.

[5] S. Al-ghamdi, and F. Al-Harigi, “Rethinking image of the City in the Information Age,” *Procedia Computer Science*, 65, 2015, pp. 734-743, doi: 10.1016/j.procs.2015.09.018.

[6] M. Riza, N. Doratli, and M. Fasli, “City branding and identity,” *Procedia-Social and Behavioral Sciences*, 35, 2012, pp. 293-300, doi: 10.1016/j.sbspro.2012.02.091.

[7] A. Adamus-Matuszyńska, J. Michnik, G. Polok, “A Systemic Approach to City Image Building,” The Case of Katowice City. *Sustainability*, 11, 4470, 2019, doi: 10.3390/su11164470.

[8] J. Chen, Z Zhen, G. Li, “Design and Construction of City Network Image Monitoring System,” *Journal of the China Society for Scientific and Technical Information*, Mar, 2019, 38, pp. 299-309, doi: 10.3772/j.issn.1000-0135.2019.03.008.

[9] C. Wong, and S. Qi, “Tracking the evolution of a destination's image by text-mining online reviews-the case of Macau,” *Tourism management perspectives*, 23, pp. 19-29, doi: 10.1016/j.tmp.2017.03.009.

[10] F. Ali, D. Kwak, P. Khan, S. Islam, K. Kim, K. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," *Transportation Research Part C: Emerging Technologies*, 77, 2017, pp. 33-48, doi: 10.1016/j.trc.2017.01.014.

[11] H. Huang, and L. Peng, "A Study of the Chinese Urban Fashion Image Based on the Big Data of Internet News," *Journal of Xiamen University* (Arts & Social Sciences), 2019, doi: 10.3969/j.issn.0438-0460.2019.04.014.

[12] Y. Patil, and S. Patil, "Implementation of Enhanced Web Crawler for Deep-Web Interfaces," *International Research Journal of Engineering and Technology (IRJET)*, Aug, 2016, 03, pp. 2088-2092.

[13] J. Sobrinho, G. Júnior, and C. Vinhal, "Web Crawler for Social Network User Data Prediction Using Soft Computing Methods," *International Journal of Computer Science & Information Technology (IJCSIT)*. Vol. 11, April, 2019, pp. 79-88, doi: 10.5121/ijcsit.2019.11207.

[14] M. Bahrami, M. Singhal, and Z. Zhuang, "A cloud-based web crawler architecture," *In 2015 18th International Conference on Intelligence in Next Generation Networks*, IEEE, February, 2015, pp. 216-223, doi: 10.1109/ICIN.2015.7073834.

[15] Y. Fan, "Design and Implementation of Distributed Crawler System Based on Scrapy," *In IOP Conference Series: Earth and Environmental Science*, IOP Publishing. Vol. 108, No. 4, p. 042086, January, 2018, doi: 10.1088/1755-1315/108/4/042086.

[16] S. Amudha, and M. Phil, "Web crawler for mining web data," *International Research Journal of Engineering and Technology (IRJET)*, Feb, 2017, 04, pp. 128-136.

[17] C Sadowski, and G. Levin, "Simhash: Hash-based similarity detection," Technical report, Google, 2007.

[18] D. Blei, A. Ng, and M. Jordan. "Latent dirichlet allocation," Journal of machine Learning research 3, Jan, 20, pp. 993-1022.