A Method for Detecting the Real Location of Agency Website Based On Search Engine

Xiao-Hui CHOU[a], Liang GAN[b], Ai-Ping LI, Zhong-He HE[c], Feng-Yu YANG

Department of Computer, National University of Defense Technology, Changsha
[a]cxh_nudt@126.com, [b]gl.nudt@gmail.com, [c]riterhe@outlook.com

Abstract. This paper provides a method to detect the real location of agency website based on search engine. We will analyze and process the target agency website to obtain the server routing information, extract critical feature information from web content, combine with search engine to acquire web data, and calculate word frequency. Through named entity recognition, web text matching calculation and syntactic analysis, we can infer real location of the target agency website in the real world. The experimental results show that our approach is reliable, correct and effective.

1 Introduction

The information of a large number of agency websites in the network is not complete, people can not directly know the real location of these agencies through the website information. Some reasons may lead to this phenomenon. Some agencies may hide the location information on their website deliberately for the purpose of illegal operation, or carelessly left this information during building their website. Thus, it's difficult to accurately judge the real location of the agency. Moreover, many institutions website with duplicate website names and same geographical name, such as *Phoenix Town*, there are 16 Phoenix towns with the same name in China, which are distributed in different provinces. So people can not identify the website refers to which town government agencies if only through the website title such as *The People's Government of Phoenix Town*. Location information is an important information for a website. The lack of the information brings great difficulties for the Internet management department to supervise network, and users can not determine the safety and authenticity of the website. If we can find out the real location of these agency websites in the real world, it will help users make risk assessments [1] and help network supervision departments rectify the network, gather intelligence, investigate agencies and do other routine inspections.

2 Related work

At present, the main method is to search the location of agency website through the Yellow Pages information service platform which is an artificial information collection platform. But many network information manually collected is missing or incomplete. Manual collection and collation of web information not only requires labor cost, but also consumes longer time. Real-time maintenance updating also be an problem. Besides, users need to pay a certain amount of service costs[2]. The other method is to get IP address by domain name resolution, and search the server address corresponding to the IP to obtain the location information. But the method of obtaining the geographic information of the website is not accurate[3]. There are many research results on IP localization method. Song Jian proposed a method to locate IP address based on network topology measurement[4]. The method will do network detection of trace route according to the detected target extracted from the existing IP address database. After that, the raw data will be preprocessed, statistics and clustering. And then analyzing and correcting IP address information on the basis of the topological relations between the IP addresses which are obtained by the network measurement. Zhu Bin proposed a method that through the research and implementation of a resemble SVD algorithm and efficient query algorithm about IP to make the positioning results more accurate and efficient[5]. Wang Ting et al proposed a IP localization method based on association rule mining(IPGEL)[6]. Jia Weiwei proposed a method of IP location optimization based on Routing[7]. But these methods are aimed at locating the server address, and only by virtue of the geographical location of the server can not determine the real location of the agency.

3 System framework

The system first analyzes the agency website URL and resolves the domain name into IP address. Then locate server address according to IP. The system will judge whether the IP address and the location of the server is the only corresponding relationship. If the relationship is exist, we will get the conclusion that the location of the server is the real location of the agency. If through the domain name resolution we get more than one server to provide the same service, thus we can not determine the real location through the server address. Therefore we need to be further analysis of the website content to determine the real location of the agency of the website. The System design flow is shown in Figure 1.

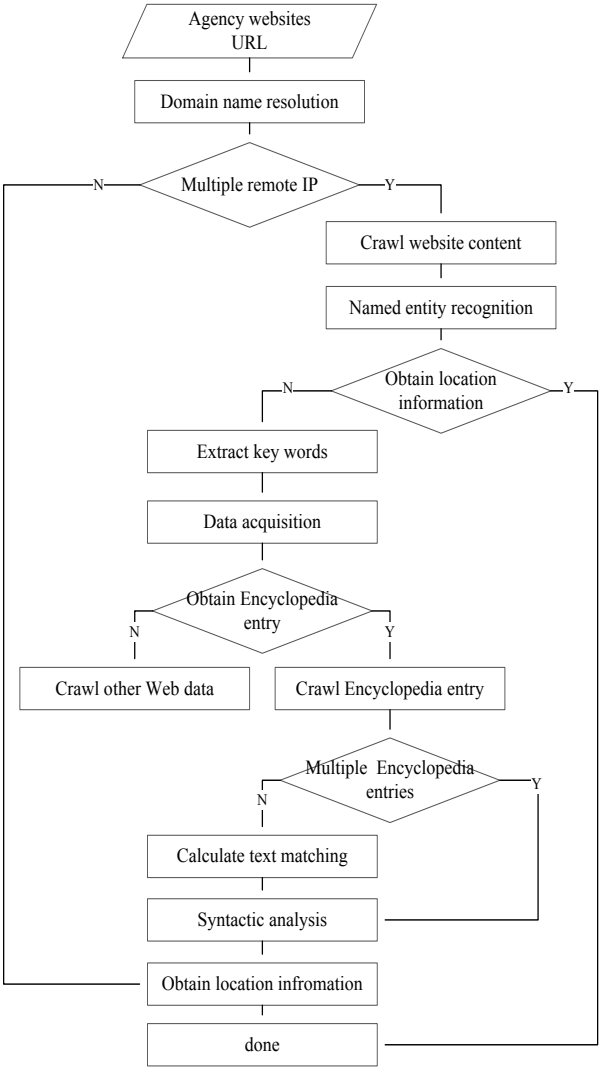

Figure 1. The system flow chart for detecting real location of agency website based on search engine.

Our method can be used to extract features from website content. The features will be used for reasoning about the organization's geographic information by analyzing the content of the target website. The main function module includes: server routing information analysis and processing, website structure and content crawling, data acquisition, named entity recognition computing, geographic information inference. The system structure is shown in Figure 2.

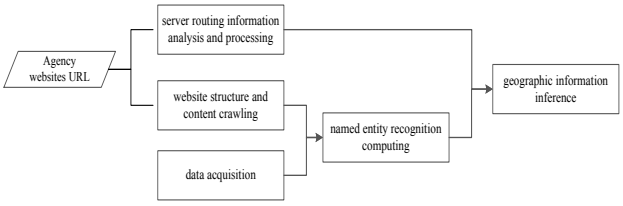

Figure 2. The system structure of real location of agency website detecting based on search engine.

4 Technologies

4.1. Data acquisition

The method uses search engine for data acquisition. Search engine is a system and a tool to return the relevant information according to the request of the user. Search engine includes three parts: information collection, information processing and user query. It collects information from the Internet or local database with a certain strategy and using a specific computer program. After organizing and arranging the information, search engine provides users with retrieval services, and will retrieve the relevant information to the user[8]. The combination of artificial intelligence, NLP, machine learning, big data and search engines has a great economic and academic value[9]. In our method, we acquire useful data returned by Baidu search engine, and analyze and process data in the next step.

4.2. Web crawling

The system uses the Jsoup framework to perform web crawling. Jsoup is a Java parser[10]. Through the target URL to crawl the corresponding web content and parse page. Jsoup can load the HTML document from string, the URL address, or local file, and generate a document object instance. It's very useful to crawl the information of website's record number or website structure according to the information provided by the website. Because record numbers are usually related to geographic location, so we may directly obtain the exact location of the agency.

4.3. Named entity recognition

Early named entity recognition work is mainly to identify the general proper nouns, including names, place names, organization names. Later research further subdivided the

type of these nouns. More and more entity types are proposed. Current research methods are using supervised learning, automatic construction rules or sequence tagging. Hidden Markov model is a statistical model which is widely used in the field of Natural Language Processing. It takes context information into account and describes a Markov process with hidden location parameters[11]. Cascaded hidden Markov model (cascaded HMM) is an improved version of the hidden Markov model. Its aim is to merge the named entity recognition, such as the name recognition, the place name recognition and the organization name recognition, into a relatively uniform theoretical model[12]. We extract the key words such as the geographic information and the characters from the web content by named entity recognition.

4.4. Syntactic analysis

Dependency Parsing (DP) reveals the syntactic structure by analyzing the dependence relation among the components in the language unit. Compared with the syntactic analysis of phrase structure, the dependency structure syntax analysis describes the syntactic structure which is the dependency between words and words[13]. Dependency parsing can be decomposed into two sub tasks: identification and classification, and can be represented by a dependency tree[14]. For example, The people's Government of Bohai town is located in the town of Bohai Street revitalization. Its dependency tree is show in figure 3.

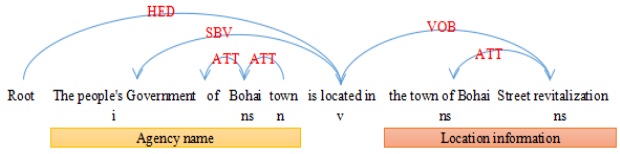

Figure 3. An example of dependency tree.

5 Experiment

The experiment uses Baidu as a search platform (http://www.baidu.com). In experiment, the initial query condition is submitted to the search engine. In the results page returned by Baidu, we select the top 30 results for analysis.

We selected three types of agency website including: government agencies, educational institutions, private enterprises. The number of test URL of each one is 100 and the result of experiment is show in figure 4.

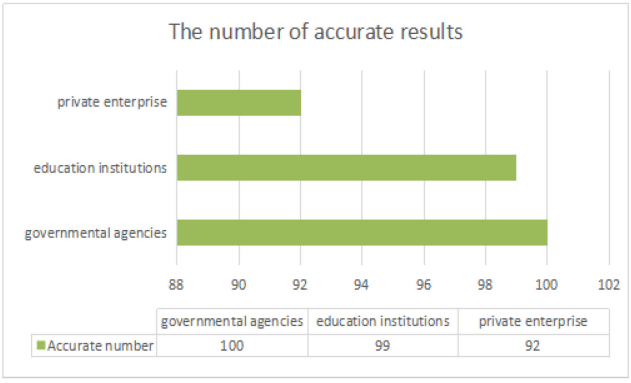

Figure 4. The number of accurate results.

The experimental results shows that the correct rate of the three types of agency websites were more than 90%. The output of detecting results are show in table 1.

Table 1. The Sample of recognition results

No.	Results		
	URL	Agency	Location
1	http://bhz.jingning.gov.cn/	The government of Bohai town	Heilongjiang province, Ningan city, Bohai Town, street revitalization
2	http://www.fenghuang.gov.cn/	The government of Phoenix town	Linzi District, Zibo City, Shandong Province
3	http://www.pku.edu.cn/	Beijing university	No. 5 the Summer Palace Road, Beijing, Haidian District
4	http://www.gdton.com/	Everbright Group	Guangdong Province, Dongguan City Dongcheng Jinghu blue County South four ring road

6 Conclusions

This paper provides a method based on the search engine to crawl and process the web data in real time. The system with multiple data acquisition analysis and processing module, can intelligently detect the real location of the website of agency. Compared with the manual sorting method, the method belongs to the automatic detection method and saves the time for manual judgment and processing. It has high chronergy because of collect and process data in real time and enhance the effectiveness of data. It covers a wider range of inquiries and has a better detection effect. But our method is mainly to deal with the type of agency websites. In the future, it can be extended to other types of websites. In the next work, we will pay more attention to improving algorithms to improve the detection rate.

Acknowledgment

The work is supported by National Basic Research and Development Program (No.2013CB329601, 2013CB329604) and National Natural Science Foundation of China (No.61502517, No.61472433, No.61372191, No.61572492).

References

1. Deng Haiping. To avoid the risk of lost contact with the private equity institutions [J]. Public financial advisor, 1 (2016)
2. The directory of national enterprises [EB/OL]. http://ml.fosang.com/post/1.html
3. Pan Xiaonan. Research and analysis of IP positioning technology [J]. Applied energy technology, 6: 47-49 (2015)
4. Song Jian. Design and implementation of IP address location system based on network topology measurement [D]. Beijing University of Posts and Telecommunications (2015)
5. Zhu Bin. Research and implementation of IP address based network entity geographic location technology [D]. Beijing University of Posts and Telecommunications (2015)
6. Wang Ting, Song Junde, Song Meina et al. A IP localization method based on association rules mining [J]. Journal of Southeast University (NATURAL SCIENCE EDITION), **45** 4: 657-662 (2015)
7. Jia Weiwei. IP positioning optimization based on routing tracking [D]. Zhengzhou University (2013)
8. Croft Donald, Metzler Trevor, Strohman. W.Bruce. Search Engine: information retrieval practice [M]. Beijing: Mechanical Industry Press (2009)
9. Liu Chao. Lightweight search engine based on Text Mining [D]. Southwestern University (2015)
10. Jsoup [EB/OL]. https://jsoup.org/
11. Chen Ji. A review of named entity recognition [J]. modern computer, 3: 24-26 (2016)
12. Yu Hongkui, Zhang Huaping, Liu Qun et al. Chinese Named Entity Recognition Based on cascading hidden Markov model [J]. Journal of communication, **27** 2: 87-94 (2006)
13. Wei Yong, Hu Danlu, Li Xiang, et al. Chinese place name recognition method considering syntactic features [J]. Journal of Surveying and Mapping Science and technology, 1 (2016)
14. Xiao Xin, Fan Shixi, Wang Xuan et al. Dependency parsing based on maximum entropy model [J]. Chinese Journal of information, **23** 2: 18-22 (2009)