An Algorithm of Street-level Landmark Obtaining Based on Yellow Pages

Te Ma, Fenlin Liu, Xiangyang Luo, Meijuan Yin, Ruixiang Li

State Key Laboratory of Mathematical Engineering and Advanced Computing,
Zhengzhou Science and Technology Institute, China
tema_123@163.com, fenlinliu@vip.sina.com, luoxy_ieu@sina.com, raindot_ymj@163.com, xxxlrxxx@163.com

Abstract

Street-level landmarks are the important foundation for achieving the high-precision geolocation of target IPs. Considering that yellow pages contain a large number of Web and Email domain names corresponding to institutions; the content is stable; and the format is fixed, this paper proposes a street-level landmark obtaining algorithm based on yellow pages. The domain names of institutions in yellow pages are extracted by using regular expression, and the corresponding IPs are parsed. Landmarks are screened according to whether an IP attribution is consistent with the cities where all possible corresponding institutions are located. By using the SLG geolocation algorithm, the landmarks with a geolocation error falling within the evaluation threshold are rated as reliable landmarks. The experimental results show that the proposed algorithm can effectively correct the mis-deletion and mis-evaluation of some landmarks by the existing typical landmark obtaining algorithm: based on 10 Chinese yellow pages (about 2 million institutions in 5 cities) and 3 American yellow pages (about 1 million institutions in 3 cities), a total of 55,960 reliable street-level landmarks for Web and Email servers are obtained. Among the 346,753 Web server IP evaluations, 48,361 landmarks are revised and 40,753 reliable street-level landmarks are augmented compared with the Web-Based landmark obtaining algorithm.

Keywords: IP geolocation, Yellow pages, Street-level landmarks, Candidate landmarks, Landmark evaluation

1 Introduction

The purpose of network entity geolocation is to determine the geographical location of a network entity. Since each network entity in the Internet often corresponds to a unique IP address, it is thus often referred to as IP geolocation [1]. At present, in terms of business application, IP geolocation technologies can provide services for targeted advertising and network data localization; regarding cyber crime tracking, it can provide technical support for the network forensics of illegal acts such as online fraud; and in addition, there are numerous practical applications for IP geolocation, such as making the deployment strategies of network infrastructures and helping with the performance, service continuity, regulation and other aspects of cloud computing [2-4]. The existing IP localization algorithms mainly include: (1) geolocation algorithms based on database query, such as MaxMind[1], IP2Location[2], and IP138[3], etc.; (2) landmark-Based geolocation algorithms, such as Octant [5], and SLG (Street-Level Geolocation) [6], etc.; (3) geolocation algorithms based on delay measurement and topology analysis, such as CBG (Constrained-Based Geolocation) [7], and TBG (Topology-Based Geolocation) [8], etc.; and (4) geolocation algorithms based on additional technical support such as GPS (Global Positioning System), and Wifi [9-10], etc. Although the existing algorithms have a certain degree of geolocation ability, for the actual network IP targets, the geolocation ability thereof is often constrained by many factors. Especially for high-precision geolocation algorithms, the quantity and reliability of landmarks often have a great impact on geolocation results [11-13]. As a consequence, the obtaining and evaluation of reliable landmarks have important theoretical and practical significance.

Landmark refers to the network entity with a known geographical location and a stable IP identity. As the important basic data of IP geolocation, landmarks are similar to the reference points in GPS geolocation and Beidou geolocation [14]. Landmarks can be divided into city-level landmarks and street-level landmarks according to geographic granularity. City-level landmark obtaining algorithms mainly include: a network forum-based landmark obtaining algorithm [15], which extracts from a network forum users' IP addresses and the geographical area information involved in the forum, and uses these IPs as candidate landmarks. The geographical locations of the last hop routing IP and its previous hop routing IP of these candidate landmark IPs in the database are queried.

*Corresponding Author: Xiangyang Luo; E-mail: luoxy_ieu@sina.com
DOI: 10.3966/160792642019092005009

When the query result is in the same city as the candidate landmark IP, the candidate landmark IP is evaluated as a reliable landmark. However, the use of this algorithm is subject to the privacy protection restriction of the network forum. City-level landmark obtaining algorithms also include: the DNS-based parsing Structon algorithm [16]. The proposed algorithm takes the Web server IP corresponding to the domain name consistently parsed by multiple DNS servers as the landmark. But the reliability of landmarks obtained based on this algorithm is affected by technologies such as the CDN network and shared hosting. Street-level landmark obtaining algorithms mainly include: a landmark collection algorithm based on users' GPS locations [17]. This algorithm corresponds the first public IP address accessing a user to the user's GPS location so as to obtain a street-level landmark. Obviously, the reliability depends on the distance between the actual location of the public IP and the user's GPS location. Street-level landmark obtaining algorithms also include: a Web-Based landmark obtaining algorithm [6]. The proposed algorithm corresponds the web server IP of the domain name corresponding to an institution to the geographical location of the institution, then obtains a large number of candidate landmarks, and screens the landmarks according to the one-to-one correspondence between an IP and a domain name. This algorithm is the most typical landmark obtaining algorithm currently. Nevertheless, the applications and branches of CDN networks, shared hosting and hosted service technologies abound in actual networks. Among candidate landmarks, there are a large number of IPs and domain names that are not one-to-one corresponding to each other. These candidate landmarks are excluded based on the Web-Based landmark obtaining algorithm, and some reliable landmarks may be deleted by mistake.

Yellow pages are a database collecting the public information of institutions. It is characterized by a stable content and fixed format, and contains a large number of institution names, the Web domain names and email domain names of the institutions. These data are easy to extract. Obviously, the Web and Email server IPs are the important data sources for street-level landmarks. Different from the Web-Based landmark obtaining algorithm, which simply uses the geographical locations of institutions, branches and shared hosting as the geographical locations of candidate landmarks, this paper firstly reduces the size of candidate landmarks by using the rule that whether the IP attribution is consistent with the city where the corresponding institution is located, thereby lowering algorithm overhead for further landmark evaluation. Finally, the SLG geolocation algorithm is used to evaluate the candidate landmarks, which further improves the reliability of landmarks.

The rest of the paper is organized as follows: Section 2 describes the basis for taking yellow pages as landmark obtaining data sources; Section 3 elaborates the steps of algorithm data preprocessing, candidate landmark obtaining and street-level landmark evaluation; Section 4 analyzes the feasibility and the performance of the proposed algorithm; Section 5 verifies the effectiveness of the proposed algorithm through experiments, and compares with the existing typical landmark obtaining algorithm.

2 Problem Formulation

Yellow pages contain a large number of information about government, enterprises, schools, public service units. These institutions often disclose their own Web domain names and email domain names. The positions of various types of data in the yellow pages are relatively fixed, which facilitates batch extraction by using regular expression, as shown in the following figures.

Figure 1 and Figure 2 respectively capture the yellow page contents of China University Information Inquiry Peking University (National University Information Inquiry. http://gkcx.eol.cn/schoolhtm/) and the US YP yellow page (YP.com. http://www.ypstate.com/) Eden Technologies, Inc. It can be seen that there are plenty of specific information about the institutions in the yellow pages, including the institution names, Web domain names and Email domain names of the institutions in red.

Table 1 shows the number of the institution names, Web domain names and email domain names extracted from the China University Information Inquiry and the US YP yellow page. It can be seen that the number of the obtained domain names is far more than the number of institutions, indicating that yellow pages contain a large number of institutions, branches, and the Web domain names and email domain names of institutions.

In the original stage of network construction, in order to facilitate the management and maintenance of basic network service facilities, most institutions often set basic network service facilities such as Web servers and email servers inside their own institutions. Precisely based on this feature, this paper takes the geographical location of an institution as the actual area where the server IP of the domain name corresponding to the institution is located, obtains candidate landmarks through the established screening rule, uses the SLG geolocation algorithm to evaluate the candidate landmarks, and finally obtains reliable street-level landmarks.

Figure 1. China university information inquiry "Peking University"

Figure 2. US YP Yellow page "Eden Technologies, Inc."

Table 1. The number of the institution names and domain names extracted from China university information inquiry and US YP yellow page

Yellow page name	Quantity of institutions	Quantity of Web domain names	Quantity of Email domain names
China University Information Inquiry	96573	270412	198752
US YP Yellow Page	845321	2451309	1252513

To facilitate the elaboration on subsequent algorithms, this paper defines the following symbols:

Ci represents a city, and different subscript *i* represents different cities.

Ni is used to represent the geographical location of an institution, and different subscript *i* represents different institutions.

A candidate landmark is represented by (*IP*, *Ni*), that is, an IP address is corresponding to a geographical location of an institution.

3 Proposed Algorithms

The framework of the street-level landmark obtaining algorithm based on yellow pages is shown in Figure 3.

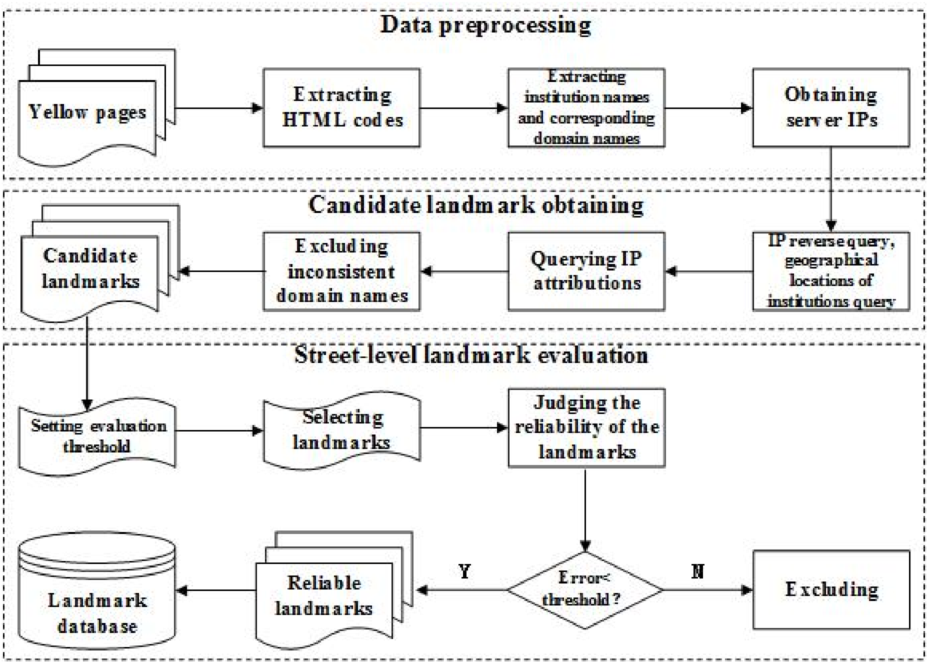


Figure 3. Framework of the street-level landmark

The proposed algorithm includes data preprocessing, candidate landmark obtaining, street-level landmark evaluation, etc., which are described in detail below.

In the data preprocessing stage, institution names in a yellow page and the corresponding Web and Email domain names are extracted; and parsing is conducted via a DNS server to obtain the Web server and Email server IPs corresponding to the domain names.

In Figure 3, the steps are specifically as follows:

Step 1. Selecting yellow pages. Selecting the yellow pages with relatively fixed data formats from the Internet. Such yellow pages facilitate the extraction of data by using regular expressions.

Step 2. Extracting HTML codes. After selecting yellow pages, obtaining all the URLs in the yellow pages by web crawler program, and extracting all the HTML codes of the yellow pages.

Step 3. Extracting institution names and corresponding domain names. In the same yellow page, data with different attributes correspond to different labels, while the position of data with the same attribute is relatively fixed. After finding the label of the corresponding data, the required data are extracted by using different regular expressions, which are specifically as follows:

An institution name is generally composed of numbers, Chinese characters, English uppercase and lowercase letters, spaces, etc. The regular expression is as follows:

$$[0\text{-}9\backslash s\backslash u4e00\text{-}\backslash u9fa5a\text{-}zA\text{-}Z]+ \qquad \textbf{(1)}$$

"*0-9*" match numeric characters; "*s*" matches spaces; "*\u4E00-\u9fa5*" match Chinese characters; "*a-zA-Z*" match English uppercase and lowercase letters; "\" is used to distinguish the types of characters; and "+" matches the above content that has appeared multiple times

A Web domain name usually consists of a multi-level domain name. The domain name is a combination of English uppercase and lowercase letters, numbers, and special symbols. The regular expression is as follows:

$$[a\text{-}zA\text{-}Z0\text{-}9\_\text{-}]+(\backslash.[a\text{-}zA\text{-}Z0\text{-}9\_\text{-}]+)+ \qquad \textbf{(2)}$$

the matching rules of "*a-zA-Z*", "*0-9*", and "+" are the same as the formula (1); "_" matches the underline; "—" matches strikethrough; and "\." matches the "." character.

An Email domain name usually consists of a username and a multi-level domain name, and also has a corresponding identifier. The regular expression is as follows:

$$[a\text{-}zA\text{-}Z0\text{-}9\_\text{-}]+@[a\text{-}zA\text{-}Z0\text{-}9\_\text{-}]+(\backslash.[a\text{-}zA\text{-}Z0\text{-}9\_\text{-}]+)+ \qquad \textbf{(3)}$$

the matching rules of "*a-zA-Z*", "*0-9*", and "+" are the same as the formula (1); the matching rule of "\." is the same as the formula (2); and the "@" matches the identifier of the Email domain name.

Step 4. Obtaining server IPs. Requesting the DNS server, using the nslookup command to parse the Web server corresponding to the domain name as well as the Email server corresponding to the Email domain name, and obtaining the corresponding IP addresses.

In the candidate landmark obtaining stage, all the domain names corresponding to the IPs are obtained through IP reverse query. The domain names are screened according to whether an IP attribution is consistent with the cities where all possible corresponding institutions are located. The candidate landmarks are obtained. Steps are specifically as follows:

Step 1. IP reverse query, and querying the geographical locations of institutions. Obtaining a certain IP address through IP reverse query, and obtaining the geographical locations of all possible institutions corresponding to IP_1 if IP_1 is corresponding to all domain names.

Step 2. Querying IP attributions. Using IP geolocation database to query the attribution of IP address IP_1 and denoting it as C_k.

Step 3. Excluding inconsistent domain names. Screening the domain names according to the rule that whether an IP attribution is consistent with the cities where all possible corresponding institutions are located, and excluding the domain names whose IP attributions are not consistent with the cities where the institutions are located: for IP_1, denoting the possible candidate landmark set as $A=\{(IP_1,N_1),(IP_1,N_2)...(IP_1,N_n)\}$, and if there is a geographical location set $\{N_{i1},N_{i2},...,N_{in}\}\subseteq\{N_1,N_2,...,N_n\}$, and $N_{ij}\notin C_k$, then excluding (IP_1,N_{ij}), j =1, and 2...n from A, and finally obtaining the candidate landmarks.

In the street-level landmark evaluation stage, a landmark evaluation threshold is set, and the existing accurate landmarks are selected. The geolocation algorithm is used to evaluate the candidate landmarks, and the landmarks whose geolocation error (the actual physical distance between the candidate landmark and the existing accurate landmark) falls within the threshold range are rated as reliable street-level landmarks. The steps are as follows:

Step 1. Setting an evaluation threshold and selecting landmarks. Setting a landmark evaluation threshold for the candidate landmarks according to geolocation accuracy requirements; and taking the IPs with known geographical locations as the existing accurate landmarks, namely, the reference points of the geolocation algorithm.

Step 2. Judging the reliability of the landmarks. The geolocation algorithm is used to geolocate the candidate landmarks to obtain geolocation results.

Step 3. Obtaining street-level landmarks. Taking the candidate landmarks whose geolocation error falls within the existing accurate landmark evaluation threshold range as reliable street-level landmarks (when for the same IP, there are multiple candidate landmarks satisfying the threshold, the candidate landmark with the smallest geolocation error is rated as a reliable landmark), and storing them in the landmark database.

4 Analysis of Proposed Algorithm

Candidate landmark obtaining and street-level landmark evaluation are the key links of the proposed algorithm. This section elaborates on the feasibility and the performance of the proposed algorithm.

4.1 Algorithm Feasibility Analysis

In the data preprocessing stage, domain names and their corresponding institutions can be obtained; the IP addresses corresponding to the domain names can be obtained through DNS parsing. In the candidate landmark obtaining stage, all the domain names corresponding to the IPs can be obtained through IP reverse query, and the domain names are corresponding to the geographical locations of the institutions. Hence, the correspondence between available IPs and the domain names can be used to describe the candidate landmarks. Table 2 shows the correspondence between some IP addresses (attributions), the domain names corresponding to the IPs, and the cities where the institutions corresponding to the domain names are located.

Table 2. Correspondence between IP addresses, their corresponding domain names and the cities where the institutions are located

IP address (attribution)	Corresponding domain name	Institution location
60.247.18.3 (Beijing)	www.bnu.edu.cn english.bnu.edu.cn	Beijing
219.142.121.58 (Beijing)	email.bnu.edu.cn	Beijing
59.38.32.56 (Zhuhai)	www.bnuz.edu.cn	Zhuhai
61.172.193.59 (Shanghai)	**www.bnuz.edu.cn** sww.changning.scn cnq.sh.gov.cn www.shskkq.com www.changning.cn **www.zzsmfc.com**	**Zhuhai** Shanghai Shanghai Shanghai Shanghai **Zhengzhou**
183.57.48.35 (Shenzhen)	mx.elevatorbuy.com	**Shanghai**
220.181.14.139 (Beijing)	www.163.com and 21 domain names	Beijing

It can be seen from Table 2 that there is not one-to-one correspondence between the server IP and the domain name. Besides, there is a phenomenon that the IP attribution is not consistent with the city where the corresponding institution is located. As shown in Table 2, the attribution of IP address 61.172.193.59 is Shanghai; the corresponding Web domain names are respectively www.bnuz.edu.cn, sww.changning.scn, cnq.sh.gov.cn, www.shskkq.com, www.changning.cn and www.zzsmfc.com; the institutions are respectively located in Zhuhai, Shanghai, Zhengzhou, etc. According to Step 3 in the candidate landmark obtaining stage of the proposed algorithm, the institutions corresponding to Web domain names www.bnuz.edu.cn and www.zzsmfc.com are respectively located in Zhuhai and Zhengzhou, which are not consistent with that the attribution of IP address 61.172.193.59 is Shanghai. Therefore, candidate landmarks (61.172. 193.59, 22.39019(N), 113.50923(E)) 、(61.172.193.59, 34.76791(N), 113.71742(E)) should be excluded from possible candidate landmarks; similarly, the attribution of IP address 183.57.48.35 is Shenzhen, and the institution corresponding to Email domain name mx.elevatorbuy.com is located in Shanghai. Candidate landmarks (183.57.48.35), 31.23925 (N), 121.47459 (E)) should be excluded.

It can also be seen from Table 2 that there is a phenomenon that an IP corresponds to multiple domain names. For example, the Web domain names corresponding to IP address 60.247.18.3 include www.bnu.edu.cn and english.bnu.edu.cn. Furthermore, there is a phenomenon that a domain name corresponds to multiple IPs. For example, the domain names corresponding to IP address 220.181.14.139 include 22 domain names such as www.163.com. At this time, the IP attribution is consistent with the city where the institution corresponding to the domain name is located. It is difficult to directly judge the real geographical location of the IP. This algorithm uses it as a candidate landmark for further evaluation.

According to the candidate landmark obtaining stage of the proposed algorithm, the candidate landmarks finally obtained from Table 2 are shown in Table 3.

As can be seen from Table 2 and Table 3, compared with the Web-Based landmark obtaining algorithm, which uses each IP and its corresponding domain name as candidate landmarks, the proposed algorithm effectively reduces the number of candidate landmarks.

Table 3. Candidate landmarks obtained from Table 2

Corresponding Institution location	Landmark corresponding IP
Beijing	60.247.18.3
Beijing	219.142.121.58
Zhuhai	59.38.32.56
Shanghai	61.172.193.59
Beijing	220.181.14.139

In the street-level landmark evaluation stage of the proposed algorithm, according to the feature that the landmark and the target directly connected to the nearest common router are geographically close [18], the SLG geolocation algorithm is used to take the existing accurate landmark having the minimum relative delay with the candidate landmark as the location estimation of the target. The reliability of the landmark is judged based on the geolocation error. As shown in Figure 4, let T be the candidate landmark to be evaluated and L be an existing accurate landmark; the existing accurate landmark L and the candidate landmark T have a nearest common router. Among all the existing accurate landmarks having a nearest common router with T, L is the landmark having the minimum relative delay; R is the evaluation threshold; and r is the actual physical distance between the candidate landmark T and the existing accurate landmark L. In Figure 4, if $r < R$, the candidate landmark T is evaluated as a reliable street-level landmark.

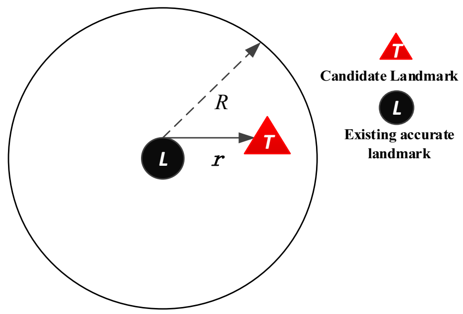


Figure 4. Setting thresholds to evaluate street-level landmarks

In the street-level landmark evaluation stage of the proposed algorithm, compared with the Web-Based landmark obtaining algorithm, which only stores the landmarks whose IPs and domain names are one-to-one corresponding, the proposed algorithm can effectively evaluate the candidate landmarks and effectively correct some landmarks mistakenly deleted or evaluated by the Web-Based landmark obtaining algorithm.

Table 4 shows some candidate landmarks of the proposed algorithm.

Table 4. Candidate landmarks corresponding to IP addresses 210.22.116.78 and 198.4.23.145

IP address	Corresponding domain name	Institution name	Latitude (N)	Longitude (E)
210.22.116.78	shiba.hpe.cn	Shanghai No. 8 Middle School	31.21478	121.49171
	www.hpe.sh.cn	Huangpu District Education Center	31.20816	121.48622
	gm.hpe.sh.cn	Shanghai Guangming Middle School	31.22831	121.47583
198.49.23.145	afanyc.com	Afanyc art exhibition hall	40.72213	-74.00151

As can be seen from Table 4, the domain names corresponding to IP address 210.22.116.78 are shiba.hpe.cn, www.hpe.sh.cn, and gm.hpe.sh.cn, and the corresponding institutions are Shanghai No. 8 Middle School, Huangpu District Education Center, and Shanghai Guangming Middle School respectively; the domain name corresponding to IP address 198.4.23.145 is afanyc.com, and the corresponding institution is the New York afanyc art exhibition hall. Based on the screening rule of IP and domain one-to-one correspondence, the Web-Based landmark obtaining algorithm excludes the candidate landmark corresponding to IP address 210.22.116.78, and evaluates the candidate landmark corresponding to IP address 198.4.23.145 as a reliable landmark.

The evaluation results of the proposed algorithm on the candidate landmarks in Table 4 are shown in Table 5.

Table 5. Evaluation results on the candidate landmarks in Table 4

Candidate landmark corresponding to IP address	Institution name	Geolocation error
210.22.116.78	Shanghai No. 8 Middle School	19.873km
	Huangpu District Education Center	**3.893km**
	Shanghai Guangming Middle School	26.531km
198.49.23.145	**Afanyc art exhibition hall**	**49.381km**

As can be seen from Table 5, the proposed algorithm sets the evaluation threshold R to 5 km, and evaluates the candidate landmarks by using the SLG geolocation algorithm. For the candidate landmark corresponding to IP address 210.22.116.78, the geographical location of the institution with the smallest geolocation error is the Huangpu District Education Center. The geolocation error satisfies the set threshold, and the proposed algorithm evaluates it as a reliable street-level landmark. That is, the Web-Based landmark obtaining algorithm deletes this reliable street-level landmark by mistake. For the candidate landmark corresponding to IP address 198.49.23.145, the geolocation error is 49.381 km, which exceeds the set threshold, and the proposed algorithm excludes it. That is, the Web-Based landmark obtaining algorithm rates it as a reliable landmark by mistake.

The Web-Based landmark obtaining algorithm takes the geographic location of an institution whose IP is corresponding to its domain name as a candidate landmark. Among them, there may be invalid landmarks whose IP attributions are inconsistent with the cities where the institutions are located. However, the proposed algorithm can exclude some invalid landmarks by using the corresponding landmark screening rule. Taking New York City's 5,000 Web server IPs and their corresponding domain names as examples, Table 6 shows candidate landmark obtaining by the two algorithms and evaluation results via the SLG algorithm.

Table 6. Candidate landmark obtaining on New York City by the two algorithms and evaluation results via the SLG algorithm

Quantity of IPs	Quantity of domain names	Algorithm	Quantity of candidate landmarks	Quantity of reliable landmarks after evaluation
5000	124,537	Web-Based Algorithm [6]	208,122	552
		Proposed Algorithm	21,224	1,287

It can be seen from Table 6 that the Web-Based landmark obtaining algorithm obtains 208,122 candidate landmarks, while the proposed algorithm obtains 21,224 candidate landmarks and excludes 186,898 candidate landmarks whose IP attributions are inconsistent with the geographical locations of the institution, effectively reducing the size of the candidate landmarks. When the candidate landmarks obtained by the two algorithms are evaluated by the SLG algorithm, the evaluation costs for the Web-Based obtaining landmark algorithm is higher due to a larger number of candidate landmarks; whereas the number of reliable landmarks obtained by the proposed algorithm is significantly greater than that obtained by the Web-Based landmark obtaining algorithm, despite the number of candidate landmarks obtained by the proposed algorithm is relatively small.

4.2 Algorithm Performance Analysis

The Web-Based landmark obtaining algorithm mainly includes two stages: institution information query and candidate landmark screening. The first stage is similar to the data preprocessing stage of the proposed algorithm. Both are to obtain related data such as IPs and domain names. The difference lies in that in the second stage, the proposed algorithm screens candidate landmarks according to the screening rule, which effectively reduces the size of candidate landmarks and reduces the computational overhead of landmark evaluation. It should be noted that the Web-Based landmark obtaining algorithm obtains a large number of invalid landmarks. As shown in Table 6 above, if evaluation is carried out by directly using the SLG algorithm, 208,122 candidate landmarks need to be evaluated, of which computational overhead is huge compared with the 21,224 candidate landmarks of the proposed algorithm. Taken together, by comparing the candidate landmark screening and geolocation evaluation of the proposed algorithm with the search and exclusion process of the Web-Based landmark obtaining algorithm, the proposed algorithm has higher complexitiesy. Nonetheless, at present, the obtaining of landmarks attaches more importance to the reliability of landmarks. The proposed algorithm can effectively correct some landmarks mis-deleted and mis-evaluated by the Web-Based landmark obtaining algorithm. In this sense, the proposed algorithm has certain advantages.

5 Experimental Results

The experiments in this paper include four parts: candidate landmark obtaining, street-level landmark evaluation, comparison with the Web-Based landmark obtaining algorithm, and landmark reliability verification.

The experimental setups include five aspects: yellow page selection, tool usage (IP geolocation database, online map), probe source deployment, detection policy setting, and existing accurate landmark selection, which are specifically shown in Table 7.

Table 7. Experimental setups

Yellow page selection	China: 34 major cities	Chinese Enterprise Yellow Pages[1], NetEase Email Yellow Pages[2], China Telecom Yellow Pages[3], The Chinese Yellow Pages[4], The Global Yellow Pages [5], Huangye 88[6], 114 Yellow Pages[7], HK Yellow Pages[8], Chinese Enterprises Online[9], China University Information Inquiry
	US: New York, Orlando, Atlanta	Superpages[10], YP.com, Yellowpages[11]
Tool usage	IP database: Baidu database[12], Maxmind database Online map: Google Map[13]	
Probe source deployment	China: China Telecom, 1 in Beijing, Shanghai and Zhengzhou respectively US: PlanetLab[14] node, 2 in Los Angeles and San Francisco respectively, 1 in Seattl	
Policy	ICMP、UDP、TCP、ICMP-PARIS、UDP-PARIS	
Accurate landmarks	China: 1000 in Beijing, 800 in Shanghai, 500 in Hong Kong, Xi'an and Zhengzhou respectively US: 1000 in New York City, 800 in Orlando and Atlanta respectively	

Note. [1]Enterprise Yellow Pages, http://www.88152.com; [2]163mail. http://y.mail.163.com/country; [3]China Telecom Yellow Pages. http://www.yellowpage.com.cn; [4]Cnlist.org. http://www.cnlist.org/; [5]Qqhy. http://www.qqhyw.com; [6]Yellowpage88. http://www.huangye88.com; [7]114chn. http://www.114chn.com/; [8]Yellowpage of HK. http://www.yp.com.hk/; [9]Chinese Enterprises Online. http://www.71ab.com/; [10]Superpages. https://superpage.com/; [11]Yellowpages. https://www.yellowpages.com/; [12]Baidu Api. http://lbsyun.baidu.com/index.php?title=webapi/ip-api; [13]Google Maps. http://www.maps.google.com/; [14]Planetlab.http://www.planet-lab.org/

Table 7 shows that this paper selects some Chinese and American yellow pages with relatively fixed data formats to conduct experiments. In terms of tool usage, for China, the relatively reliable Baidu database is selected, and for the United States, the relatively reliable Maxmind database is selected. Regarding network detection, one way is to use multiple source deployment to perform Traceroute detection, and the other is to adopt multi-protocol detection way to improve the integrity of the obtained topology information. With regard to the selection of existing accurate landmarks, the IPs with known geographical locations are rated as accurate landmarks via Wifi hotspot collecting, multi-database comparison and public network platform PlanetLab and RIPE NCC query [19-20].

The public information of the institutions in Beijing, Shanghai, Hong Kong, Xi'an, Zhengzhou of China, and New York City, Orlando, and Atlanta of the United States is extracted from a total of 13 yellow pages in the experiment setups. The obtained Web domain names and email domain names are analyzed. Results are shown in Table 8 (see Appendix for details of the experiments on other Chinese cities).

Table 8. Domain names and IP obtaining results

The total quantity of domain names	Quantity of Web domain names	Quantity of Email domain names	The total quantity of IPs	Quantity of Web server IPs	Quantity of Email server IPs
187,947	155,996	31,951	398,567	346,753	51,814

5.1 Candidate Landmark Obtaining Experiment

As shown in Table 8, the experiment extracts more than 180,000 domain names and obtained nearly 400,000 server IPs. Among them, regarding the distribution of the quantity of domain names, the quantity of Web domain names is obviously greater than that of Email domain names; and regarding the distribution of the quantity of IPs, the quantity of Web server IPs is obviously greater than that of Email server IPs.

All the domain names corresponding to IPs are obtained through IP reverse query, and the correspondence between the domain names and the IPs is established. The results are shown in Table 9.

Table 9. IP Reverse query results

Query the total number of IPs	Quantity of Web server IPs	Quantity of Email server IPs	Corresponding to the total quantity of domain names	Quantity of Web domain names	Quantity of Email domain names
398,567	346,753	51,814	2911,393	2891,790	19,603

As shown in Table 9, the experiment obtains more domain names through IP reverse query. Regarding the distribution of the quantity of domain names, the quantity of the domain names obtained by the Web server IP significantly increases. But the Email server IP usually only corresponds to the Email server name, and the quantity of the obtained domain names decreases.

According to the screening rule of the proposed algorithm, the domain names whose IP attributions are not consistent with the cities where their corresponding institutions are located are excluded. The results are shown in the Table 10.

Table 10. Inconsistent domain name exclusion results

City	The total quantity of IPs	The total quantity of domain names	The quantity of the excluded domain names	The quantity of the retained domain names	Quantity of Web domain names	Quantity of Email domain names
Beijing	19,404	345,762	253,206	92,556	89,896	2,660
Shanghai	17,374	328,732	246,401	82,331	79,865	2,466
Hong Kong	12,966	298,754	224,674	74,080	72,487	1,593
Xi’an	10,319	189,747	136,464	53,283	52,224	1,059
Zhengzhou	8,960	183,421	149,596	33,825	32,908	917
New York	45,075	598,138	328,964	269,174	264,967	4,207
Orlando	29,919	489,771	303,317	186,454	183,720	2,734
Atlanta	29,509	477,068	318,925	158,143	154,176	3,967

It can be seen from Table 10 that there are a certain number of IPs that are not consistent with the cities where the institutions corresponding to the domain names are located. According to the screening rule, such domain names will be excluded, effectively reducing the quantity of the IPs that are corresponding to multiple domain names.

By establishing the correspondence between the domain names and the IPs, the inconsistent domain names are screened. The obtaining results on the IPs corresponding to candidate landmarks in eight cities are shown in Table 11.

Table 11. Candidate landmarks obtaining results

City	Quantity of candidate landmarks	The total quantity of IPs	Quantity of Web server IPs	Quantity of Email server IPs	Corresponding to the total quantity of domain names
Beijing	97,891	19,404	15,583	3,821	92,556
Shanghai	89,877	17,374	14,552	2,822	82,331
Hong Kong	91,228	12,966	10,725	2,241	74,080
Xi’an	58,329	10,319	8,630	1,689	53,283
Zhengzhou	40,331	8,960	7,920	1,040	33,825
New York	299,633	45,075	39,248	5,827	269,174
Orlando	228,621	29,919	27,234	2,685	186,454
Atlanta	198,743	29,509	26,588	2,921	158,143

It can be seen from Table 11 that among the results, there are numerous phenomena that an IP corresponds to multiple domain names, and a domain name corresponds to multiple IPs. And each IP that corresponds to its related domain name (the geographical location of an institution) would be rated as a candidate landmark. In this way, the experiment obtains a certain number of candidate landmarks.

5.2 Street-level Landmark Evaluation Experiment

In the street-level landmark evaluation experiment, 2.5 km, 5 km and 10 km are used as street-level landmark evaluation thresholds respectively. The existing accurate landmarks in the cities in the experimental setups are used as the reference points of the SLG algorithm to evaluate the candidate landmarks in the above 8 cities. The experimental results are shown in Table 12.

Table 12. Street-level landmark evaluation results

City	Quantity of existing accurate landmarks	Quantity of evaluation landmarks	Quantity of evaluation landmarks		
			Threshold = 10km	Threshold = 5km	Threshold = 2km
Beijing	1,000	97,891	**7,507**	2,322	721
Shanghai	800	89,877	**5,025**	1,998	691
Hong Kong	500	91,228	**5,154**	2,452	781
Xi’an	500	58,329	**4,368**	1,438	423
Zhengzhou	500	40,331	**3,429**	1,087	367
New York	1,000	297,633	**12,563**	6,721	2,033
Orlando	800	228,621	**8,661**	4,087	1,524
Atlanta	800	198,743	**9,253**	4,612	1,391

It can be seen from Table 12 that the quantity of reliable landmarks obtained through experiments varies under different accuracy conditions. Taken together, as the geolocation accuracy increases, the quantity of the obtained reliable landmarks decreases.

5.3 Comparison with the Web-Based Landmark Obtaining Algorithm

5.3.1 Comparison on the Quantity of the Obtained Landmarks

In the above landmark obtaining experiment, some landmarks are consistent with the landmarks obtained by the Web-Based algorithm. In order to compare the two algorithms, the Web-Based landmark obtaining algorithm is adopted to obtain candidate landmarks for the 17 common institutions such as universities, middle schools, primary schools, hospitals, hotels and governments in the target cities, and evaluate them. The comparison results on the quantities of candidate landmarks obtained by the two algorithms and the street-level landmarks after evaluation are shown in Table 13 and Table 14, respectively.

Table 13. Comparison on the quantities of candidate landmarks obtained by the two algorithms

City	Proposed Algorithm	Web-Based Algorithm [6]
Beijing	97,891	487,329
Shanghai	89,877	438,521
Hong Kong	91,228	553,289
Xi'an	58,329	297,634
Zhengzhou	40,331	273,225
New York	299,633	836,723
Orlando	228,621	693,657
Atlanta	198,743	599,932
Total	1102,653	4180,310

Table 14. Comparison on the quantities of street-level landmarks obtained by the two algorithms

City	Proposed Algorithm	Web-Based Algorithm [6]
Beijing	7,507	1,834
Shanghai	5,025	1,711
Hong Kong	5,154	2,933
Xi'an	4,368	957
Zhengzhou	3,429	793
New York	12,563	3,672
Orlando	8,661	2,889
Atlanta	9,253	3,067
Total	55,960	17,856

It can be seen from Table 13 that the quantity of candidate landmarks obtained by the proposed algorithm is significantly less than that of the Web-Based landmark obtaining algorithm, decreasing a total of 3,077,657 candidate landmarks. It can be known from the proposed algorithm rule that the proposed algorithm excludes a large number of invalid landmarks whose IP attributions are inconsistent with the cities where their corresponding institutions are located, which reduces the size of the candidate landmarks, and can effectively lower the algorithm overhead for landmark evaluation in the next stage. It is consistent with the analysis of the candidate landmark obtaining stage in Section 4.1.

As can be seen from Table 14, the proposed algorithm obtains significantly more street-level landmarks than the Web-Based landmark obtaining algorithm, augmenting 38,104 street-level landmarks. It can be known from the proposed algorithm rule that the landmarks obtained by the proposed algorithm are evaluated by the SLG geolocation algorithm, and have high reliability. However, some landmarks obtained by the Web-Based landmark obtaining algorithm may be mis-evaluated.

5.3.2 Correction on the Mis-deleted and Mis-evaluated Landmarks

In order to show the proposed algorithm's correction results on the mis-deleted and mis-evaluated landmarks by the Web-Based landmark obtaining algorithm, the experiment takes the above-mentioned unscreened 346,753 Web server IPs obtained through yellow pages as objects to compare the evaluation results of the two algorithms. Results are shown in the Table 15.

As can be seen from Table 15, in the evaluation results on the landmarks corresponding to the Web server IPs in the 8 cities, the quantity of reliable landmarks evaluated by the proposed algorithm is significantly more than that of the Web-Based landmark obtaining algorithm, augmenting a total of 40,753 reliable landmarks. Among them, apart from a small number of candidate landmarks that cannot be effectively evaluated due to unreachable detection and inaccurate delay measurement, the proposed algorithm effectively corrects a total of 48,361 landmarks mis-deleted and mis-evaluated by the Web-Based landmark obtaining algorithm. It is consistent with the analysis of the street-level landmark evaluation stage in Section 4.1.

Table 15. Evaluation results on the landmarks corresponding to Web server IPs

City	The quantity of reliable landmarks evaluated by Web-Based Algorithm [6]	The quantity of reliable landmarks evaluated by the proposed Algorithm	The total quantity of corrected landmarks	Corrected mis-deleted landmarks	Corrected mis-evaluated landmarks
Beijing	631	6,908	7,282	6,277	1,055
Shanghai	647	4,326	4,500	3,679	851
Hong Kong	429	4,513	5,311	4,084	1,227
Xi'an	371	2,755	2,686	2,384	302
Zhengzhou	338	2,129	2,070	1,791	279
New York	1,089	9,977	10,721	8,888	1,833
Orlando	987	7,289	7,339	6,302	1,037
Atlanta	953	8,301	8,452	7,348	1,104
Total	5,445	46,198	48,361	40,753	7,878

5.4 Verification on the Reliability of Landmarks

A sufficient number of evenly-distributed reliable landmarks can provide more network topology information around target IPs and meet the higher-precision geolocation requirements [6, 14-15]. In order to verify the reliability of the landmarks obtained by the proposed algorithm, the randomly selected 80% of the existing accurate landmarks (recorded as the original landmark set), the sum of 80% of the existing accurate landmarks and the reliable street-level landmarks obtained by the proposed algorithm (recorded as the extend landmark set) are used as the reference points of the SLG algorithm respectively. Geolocation is performed on the remaining 20% of the existing accurate landmarks. The cumulative probability of geolocation errors is shown in Figure 5, which shows the accuracy of the two landmark sets under the same geolocation error. The abscissa in the figure is the geolocation error, and the ordinate is the cumulative probability.

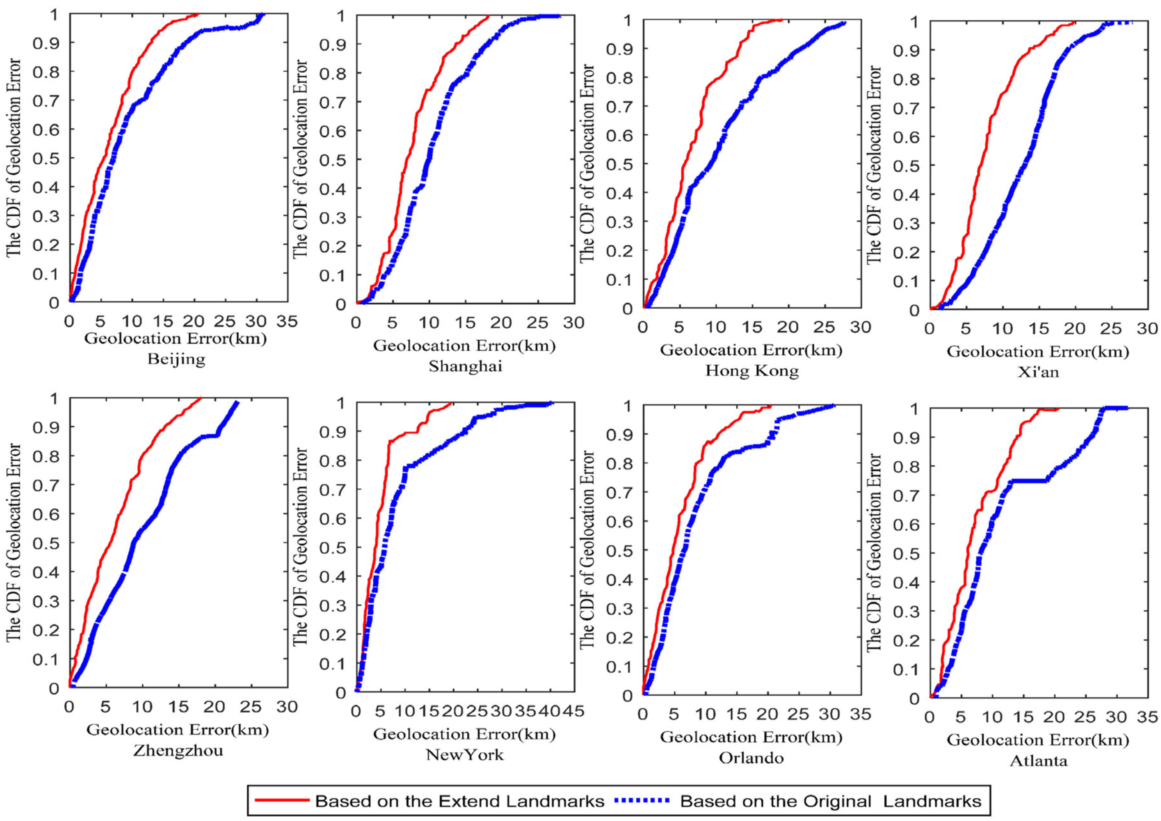


Figure 5. Comparison on the cumulative probability of geolocation errors

As can be seen from Figure. 5, it is obvious that from the original existing accurate landmarks to the adding of the reliable landmarks obtained by the proposed algorithm, the total number of landmarks that are used as the reference points of the SLG algorithm has increased. What's more, the distribution thereof is more intensive, showing the geolocation accuracy on the target IPs is significantly improved. It verifies that the street-level landmarks obtained by the proposed algorithm are of higher reliability.

Further analysis is carried out on the geolocation results based on the two landmark datasets. In the 8 target cities, the quantities of the target IPs successfully geolocated by the original landmark dataset are 181, 139, 82, 69, 71, 159, 137 and 139, respectively. The quantities of the target IPs successfully geolocated by the extend landmark dataset are 189, 147, 88, 79, 77, 179, 143 and 152, respectively. That is, the geolocation success rate of the extend landmark dataset is significantly better than that of the original landmark dataset. The geolocation errors of the target IPs successfully geolocated by the both landmark datasets are shown in the Table 16.

Table 16. Comparison on the geolocation errors of the two landmark datasets

City	Quantity of successful geolocating Target	Quantity of original landmark	Average / maximum error	Quantity of extend landmark	Average / maximum error
Beijing	181	800	9.845km/31.452km	8,307	**5.327km/20.021km**
Shanghai	139	640	10.406km/24.991km	5,665	**5.652km/17.889km**
Hong Kong	82	400	13.831km/27.601km	5,554	**5.593km/19.563km**
Xi'an	69	400	14.343km/24.908km	4,768	**6.078km/19.858km**
Zhengzhou	71	400	11.807km/23.645km	3,829	**5.389km/17.507km**
New York	159	800	8.851km/30.013km	13,363	**4.028km/19.897km**
Orlando	137	640	8.993km/30.047km	9,301	**4.518km/20.211km**
Atlanta	139	640	15.691km/26.121km	9,893	**7.531km/20.012km**

As can be seen from Table 16, in the geolocation results of the same target IPs, the average geolocation error and maximum error of the 8 cities' extend landmark sets are significantly smaller than those of the geolocation results of the original landmark sets. That is, the geolocation accuracy is effectively improved.

6 Conclusion

Street-level landmarks are the important basis for achieving high-accuracy IP geolocation. Based on the assumption that institutions often set basic network service facilities inside their own institutions, the proposed algorithm extracts the domain names corresponding to the institutions in yellow pages, parses the corresponding IPs, screens landmarks according to whether an IP attribution is consistent with the cities where all possible corresponding institutions are located, and evaluates candidate landmarks by using the SLG algorithm. The experimental results show that the proposed algorithm can significantly augment the quantity of reliable street-level landmarks, and can effectively correct some landmarks mis-deleted and mis-evaluated by the Web-Based landmark obtaining algorithm, thereby providing better landmark support for the geolocation of network target IPs.

Next, we will consider the updating timeliness of the institution information in yellow pages, and extend the types of the obtained landmarks to FTP servers, and DNS servers, etc. to enhance the quantity and reliability of the obtained street-level landmarks further. Meanwhile, we will also focus on how to quantify the impacts of the quantity and distribution of landmarks on geolocation accuracy.

Acknowledgements

The work presented in this paper is supported by the National Key R&D Program of China (No. 2016YFB0801303, 2016QY01W0105), the National Natural Science Foundation of China (No. U1636219, U1736214, 61379151, 61602508 and 61772549), Plan for Scientific Innovation Talent of Henan Province (No. 2018JR0018) and the Key Technologies R&D Program of Henan Province (No. 162102210032).

References

[1] J. A. Muir, P. C. V. Oorschot, Internet Geolocation: Evasion and Counterevasion, *ACM Computing Surveys*, Vol. 42, No. 1, pp.1-22, December, 2009.

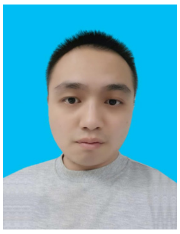
[2] V. N. Padmanabhan, L. Subramanian, An Investigation of Geographic Mapping Techniques for Internet Hosts, *ACM SIGCOMM Computer Communication Review*, Vol. 31, No. 4, pp. 173-185, October, 2001.

[3] R. Koch, M. Golling, L. Stiemert, G. D. Rodosek, Using Geolocation for the Strategic Preincident Preparation of an IT Forensics Analysis, *IEEE Systems Journal*, Vol. 10, No. 4, pp. 1338-1349, February, 2016.

[4] M. Gharaibeh, H. Zhang, C. Papadopoulos, J. Heidemann, Assessing Co-locality of IP Blocks, *IEEE Computer Communications Workshops (INFOCOM WKSHPS)*, San Francisco, California, USA, 2016, pp. 503-508.

[5] B. Wong, I. Stoyanov, E. G. Sirer, Octant: A Comprehensive Framework for the Geolocation of Internet Hosts, *the 4th USENIX Conference on Networked Systems Design and Implementation(NSDI)*, Cambridge, MA, USA, 2007, pp. 23-36.
[6] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, C. Huang, Towards Street-level Client-independent IP Geolocation, *8th USENIX Conference on Networked Systems Design and Implementation(NSDI)*, Boston, MA, USA, 2011, pp. 27-36.
[7] B. Gueye, A. Ziviani, M. Crovella, S. Fdida, Constraint-based Geolocation of Internet Hosts, *IEEE/ACM Transaction on Networking*, Vol. 14, No. 6, pp. 1219-1232, December, 2006.
[8] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, Towards IP Geolocation Using Delay and Topology Measurements, *6th ACM SIGCOMM Conference on Internet Measurement*, Rio de Janeiro, Brazi, 2006, pp. 71-84.
[9] L. Doherty, S. J. Kristofer, L. El Ghaoui, Convex Position Estimation in Wireless Sensor Networks, *the 30th IEEE Conference on Computer Communications (INFOCOM), Anchorage*, AK, USA, 2001, pp. 1655-1663.
[10] P. Bojan, P. Aleksandar, C. Zlatko, Uses of w3c's Geolocation API, *IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, Hungary, 2010, pp. 319-322.
[11] J. Wang, X. Xu, M. Yang, T. Zang, IP Geolocation Technology Research Based on Network Measurement, *IEEE Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, Warsaw, Poland, 2016, pp. 892-897.
[12] J. Chen, F. Liu, F. Zhao, G. Zhu, S. C. Ding, A SC-Vivaldi Network Coordinate System Based IP Geolocation Method, *Journal of Internet Technology*, Vol. 17, No. 1, pp. 119-127, January, 2016.
[13] F. Zhao, W. Shi, Y. Gan, Z. Peng, X. Luo, A Localization and Tracking Scheme for Target Gangs Based on Big Data of Wi-Fi Locations, *ClusterComputing*, pp. 1-12, January, 2018. doi:10.10007/s10586-018-1737-7.
[14] H. Jiang, Y. Liu, J. N. Matthews, IP Geolocation Estimation Using Neural Networks with Stable Landmarks, *Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 170-175, 2016. doi: 10.1109/INFCOMW.2016.7562066.
[15] G. Zhu, X. Luo, F. Liu, An Algorithm of City-level Landmark Mining Based on Internet Forum, *the 18th International Conference on Network-Based Information Systems*, Taipei, Taiwan, 2015, pp. 294-301.
[16] C. Guo, Y. Liu, W. Shen, Mining the Web and the Internet for Accurate IP Address Geolocations, *International Conference on Computer Communications (INFOCOM)*, Rio de Janeiro, Brazil, 2009, pp. 2841-2845.
[17] H. Liu, Y. Zhang, Y. Zhou, Mining Checkins from Location-Sharing services for Client-independent IP Geolocation, *IEEE Conference on Computer Communications (INFOCOM)*, Toronto, Canada, 2014, pp. 619-627.
[18] J. Chen, F. Liu, X. Luo, F. Zhao, G. Zhu, A Landmark Calibration Based IP Geolocation Approach, *Eurasip Journal on Information Security*, Vol. 1, No. 29, pp. 1-4, December, 2016.
[19] N. Spring, L. Peterson, A. Bavier, Using PlanetLab for Network Research, *ACM Sigops Operating Systems Review*, Vol. 40, No. 1, pp.17-24, January, 2006.
[20] H. Li, P. Zhang, Z. Wang, Changing IP Geolocation from Arbitrary Database Query Towards Multi-databases Fusion, *IEEE Computers and Communications*, Heraklion, Greece, 2017, pp. 1150-1157.

Biographies



Te Ma received the B.S. degree from the Zhengzhou Science and Technology Institute China, in 2016. Currently, he is pursuing a M.S. degree in the State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou Science and Technology Institute, China. His research interests include network entity geolocation and network entity landmark acquisition and evaluation technology.

Fenlin Liu is currently a professor of State Key Laboratory of Mathematical Engineering and Advanced Computing, and Zhengzhou Information Science and Technology Institute. His research interests include information hiding and security theory. He is the author or co-author of more than 100 refereed international journal and conference papers.

Xiangyang Luo received the M.S. degree and the Ph.D. degree from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2004 and 2010, respectively. He is currently an associate professor of State Key Laboratory of Mathematical Engineering and Advanced Computing, and Zhengzhou Information Science and Technology Institute.

Meijuan Yin was born in Anhui Province, China at November, 1977. She was conferred a M.Sc. in computer science by State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou, China, in 2003. She is working on the Ph.D. in computer software and academic of the same Laboratory.

Ruixiang Li received the B.S. degree from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2016. He is studying for the M.S. degree in State Key Laboratory of Mathematical Engineering and Advanced Computing, and Zhengzhou Information Science and Technology Institute.

Appendix

Candidate Landmarks Obtaining Results

This section shows the results of the acquisition of candidate landmarks in the remaining major cities in China. The results are shown in the following table.

Table 1. Domain names and IP obtaining results

The total quantity of domain names	Quantity of Web domain names	Quantity of Email domain names	The total quantity of IPs	Quantity of Web server IPs	Quantity of Email server IPs
1692301	1584391	107910	2352897	1723582	629315

Table 2. IP reverse query results

Query the total number of IPs	Quantity of Web server IPs	Quantity of Email server IPs	Corresponding to the total quantity of domain names	Quantity of Web domain names	Quantity of Email domain names
2352897	1723582	629315	51863921	51792525	71396

Table 3. Candidate landmarks obtaining results

City	Quantity of candidate landmarks	The total quantity of IPs	Quantity of Web server IPs	Quantity of Email server IPs	Corresponding to the total quantity of domain names
Shijiazhuang	39643	9083	8504	579	32831
Shenyang	62821	12589	10833	1726	57283
Harbin	50234	11453	10493	960	48861
Hangzhou	76389	17021	15177	1844	73751
Fuzhou	35678	8864	7852	1012	31629
Jinan	44356	13663	11763	1900	38307
Guangzhou	10753	19578	16828	2750	82877
Wuhan	62337	16543	14221	2322	58998
Chengdu	42898	14067	12684	1203	39667
Kunming	33890	8099	7424	675	28631
Lanzhou	41679	9271	8083	1188	38599
Taipei	106743	18921	15339	3522	94088
Nanning	35321	11872	10563	1309	28273
Yinchuan	27556	5831	5288	543	24387
Taiyuan	43564	8854	7934	920	30391
Changchun	39601	9207	8521	686	36709
Nanjing	98777	18582	15987	2595	91966
Hefei	49752	11672	10003	1669	40820
Nanchang	31054	9402	8771	631	28994
Changsha	48653	12031	10727	1304	40438
Haikou	34755	7489	6921	568	26303
Guiyang	38651	8261	7607	654	32836
Xining	31563	6299	5792	507	24528
Hohhot	28997	6663	6232	431	27953
Lhasa	17832	3832	3698	134	15284
Urumqi	24322	7653	6924	729	19823
Macao	69807	9571	8887	684	40237
Chongqing	53479	13381	11928	1453	48326
Tianjin	79925	16301	14533	1768	69049