Hindawi
Security and Communication Networks
Volume 2020, Article ID 8843188, 15 pages
https://doi.org/10.1155/2020/8843188

*Research Article*

# Evaluator: A Multilevel Decision Approach for Web-Based Landmark Evaluation

**Meijuan Yin, Wen Yang, Xiaonan Liu, and Xiangyang Luo**

*China State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450002, China*

Correspondence should be addressed to Meijuan Yin; raindot_ymj@163.com

Received 29 March 2020; Accepted 15 June 2020; Published 15 July 2020

Academic Editor: Clemente Galdi



Street-level landmarks are an important basis for street-level IP geolocation, and the web-based landmark is one of the main sources of street-level landmarks. Considering the existing street-level landmark evaluation methods having low accuracy and strict constraints, this paper analyses the causes and evaluation idea of invalid web-based candidate landmarks and proposes Evaluator, a web-based landmark evaluation approach. Evaluator adopts the idea of the decision tree to filter invalid landmarks layer by layer and comprehensively estimates the quantitative reliability of candidate landmarks with public data and services to obtain reliable landmarks. This paper proposes the domain name system (DNS) distributed query algorithm to effectively resolve all IP addresses of a domain name, which provides data support for Evaluator to filter candidate landmarks. Meanwhile, this paper also proposes a reverse verification algorithm to obtain all domain names of an IP address, which provides an important reference to calculate the reliability of a reliable landmark. In addition, gradient descent is used to assess the parameters of the reliability estimating model, which effectively improves the robustness of Evaluator. Experiments show that reliable landmarks from Evaluator reduce the geolocation error of 100 targets in Hong Kong from 7.30 km to 3.91 km, compared with the landmark verifying method (LVM), one of the latest web-based landmark evaluation methods. Moreover, Evaluator significantly improves the evaluation coverage based on the same geolocation accuracy with street-level landmark evaluation (SLE), one of the latest landmark evaluation methods.

## 1. Introduction

IP geolocation [1] is a technique to locate an Internet host using its IP address. A growing number of applications, e.g., targeted advertising, location-based content customizing, and network performance optimizing, are widely used with the help of IP geolocation. A series of landmark-based approaches [2–9] are widely used due to their high accuracy and high reliability. Hence, a great number of high-precision street-level landmarks have become the key foundation for IP geolocation. Hundreds of millions of web servers have been distributed all over the world with a sufficient number of scales. According to a survey [10] on web servers by NetCraft in May 2019, the number of websites responding to the survey reached 235,011,143. These web servers are ideal landmarks for IP geolocation for their stable performance and the relative fixed relation between their geography locations and IP addresses. Consequently, we call this kind of landmark as web-based landmarks. Web-based landmarks are obtained from online map mining and web page mining. The locations obtained by both methods are the geographic location of the website owners or their management organization, but what actually needed is the geographical location of the website server. In the early days of the Internet, the two are often in the same position. However, with the rapid development of the Internet, the development of virtual hosting, content delivery networks (CDNs), and cloud service networks has led to complex relationships of their position, which reduces the location accuracy of web-based landmarks. Hence, an effective method is needed to evaluate the reliability of web-based landmarks.

Some researchers have studied the evaluation of web-based landmarks. Guo et al. [11] proposed the Structon

method to mine and evaluate landmarks. They referred to the /24 IP segment and AS and BGP routing table information of the landmark and adopted the majority voting algorithm to improve the coverage and accuracy rate of the landmarks. However, they can only obtain city-level landmarks and cannot meet the needs of street-level positioning. Wang et al. [6] proposed a web-based landmark verifying method, or landmark verifying method (LVM) for short, which evaluated a landmark by comparing the results of http requests constructed by its IP address and domain name. They effectively filtered invalid landmarks and greatly improved the accuracy of web-based landmarks. However, they not only incorrectly deleted a large number of web-based landmarks whose website does not support direct access by IP addresses, but also cannot identify invalid landmarks of a CDN network and virtual hosting configured with default domain names, which reduce the accuracy of evaluation results.

The landmark evaluation ideas proposed by some researchers are also applicable to the evaluation of web-based landmarks. Shavitt and Zilberman [12] proposed a landmark evaluation method based on the point of presence (PoP) level network analysis, which uses the majority voting algorithm to determine the city-level location of landmarks, and greatly improved their location accuracy. Wang et al. [13] established an evaluation machine learning model based on the routing strategy and evaluated the city-level location of landmarks. They significantly improved the accuracy of landmarks based on the reduction of evaluation costs. Zhu et al. [14] proposed the E-GeoTrack algorithm to vote on the city of a landmark based on the implicit information in the nearest router. However, these methods can only determine the city-level location of the landmarks. Li et al. [15] proposed a street-level landmark evaluation (SLE) method based on the nearest common router. They grouped the candidate landmarks according to their nearest router and calculated their reliability from the distribution of the constraint relationship between their distance and their delay. They realized the evaluation of street-level landmarks and greatly improved the accuracy of the landmark location based on the LVM. However, they can only evaluate the landmarks sharing the most common routes which is a strict constraint for landmarks. Furthermore, they had to repeatedly measure the delay at different time periods.

Considering that the existing methods have a low accuracy or a strict constraint, this paper proposes Evaluator for web-based landmark evaluation, which adopts the multilevel decision method for layer-by-layer filtering and evaluation. First, all the IP addresses of a domain name are obtained by the DNS distributed query algorithm, and then according to the domain name and IP mapping relationship of the landmarks, the invalid candidate landmarks are filtered. Second, the reliability of a landmark is calculated with the mapped domain number, Whois information, and IP location database information of its IP address. Finally, candidate landmarks with reliability exceeding the given threshold are selected to be reliable landmarks. After testing the accuracy of the evaluation results of five cities in China and the US, we found that Evaluator significantly improved the accuracy of the evaluation effect based on LVM and maintained a much better evaluation coverage than SLE which had almost the same geolocation accuracy.

The rest of this paper is organized as follows. In Section 2, we survey the causes and evaluation strategies of invalid candidate landmarks. In Section 3, we describe Evaluator and its methodology to filter and evaluate the candidate landmarks. Then, in Section 4, we present our experimental results. Finally, we conclude the whole paper in Section 5.

## 2. Analysis of the Invalid Candidate

The web server serves as a bridge to connect the geographic location to the IP address of the candidate landmarks. The invalid candidate landmarks usually refer to landmarks whose associated locations are not the same as the actual location of the corresponding device of the IP address. The causes can be divided into two categories: one is that the mapping from the domain name to the location is abnormal, and the other is that the mapping from the domain name to the IP address is abnormal.

*2.1. Abnormal Mapping Relationship between Domain and Location.* Some candidate landmarks share the same domain name but declare different geographical locations. There are 2 reasons: (1) measurement error: inconsistent measurement standards of different data sources, and measurement errors make the geographical locations of the candidate landmarks with the same domain name vary, but their locations are very close (usually less than 1 km); (2) multibranch: branch offices of enterprises, chain stores, business outlets, university branches, etc., with worldwide geographical locations (usually more than 10 km away), however, always share the same domain name.

Evaluation idea: we can evaluate the above two cases according to the size of the distribution range of the locations claimed by the candidate landmarks sharing the same domain name.

*2.2. Abnormal Mapping Relationship between Domain and IP Address.* Abnormal mapping relationship between the domain and the IP address means that the location extracted from the website of the domain name is inconsistent with the actual location of the corresponding IP. According to the quantity relationship between the two, it is divided into four classes: one domain name mapping to multiple IP addresses, multiple domain names mapping to one IP address, multiple domain names mapping to multiple IP addresses, and one domain name mapping to one IP address. We observed Chinese candidate landmarks mined from Google Maps and found 1,286,877 domain names mapped by 83,096 IP addresses. The results of the mapping relationships between these domain names and IP addresses are shown in Figure 1.

We can find that most websites in China use the IP addresses belonging to the "one domain name mapping to one IP address" class. These websites use 53301 IP addresses to account for 60.8% of all IP addresses. Meanwhile, 12924 IP addresses (14.7%) belong to the "one domain name

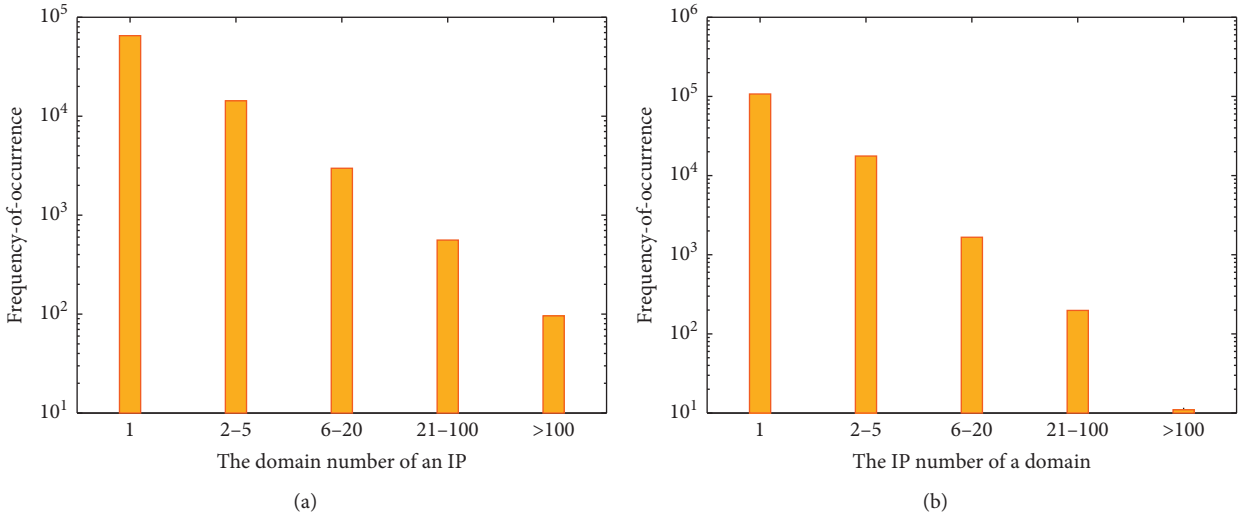



Figure 1: The mapping relationship between domain names and IP addresses in China. (a) Number of domain names for an IP address. (b) Number of IP addresses for a domain name.

mapping to multiple IP addresses" class, 15118 IP addresses (17.3%) belong to the "multiple domain names mapping to one IP address" class, and 6293 IP addresses (7.2%) belong to the "multiple domain names mapping to multiple IP addresses" class. In this section, we will analyse the abnormal mapping relationship and the causes as well as the corresponding evaluation ideas.

*2.2.1. One Domain Name Mapping to Multiple IP Addresses.* This type of landmark includes CDN networks and server load balancing. The server nodes of a CDN network often spread all over the country and even around the world. However, a landmark mined from the web page or the online map just claim the location of the user organization, which cannot be used as a reliable landmark. For example, the website of Lenovo's customer service (support.lenovo.com.cn) uses the CDN network of Unicom, which is mapped to 66 IP addresses distributed over 10 network segments, as shown in Figure 2. Nevertheless, the IP addresses of server load balancing corresponds to a set of back-end servers located inside the organization, which can be used as reliable landmarks.

Evaluation idea: from the obvious difference of the distribution of IP addresses in the above two cases, we can see that if the IP addresses mapped by a domain name are distributed in the same network segment, this kind of mapping is highly likely to be the case of the server load balancing and unlikely to be the case of the CDN network. The subnet distribution of a group of IP addresses is usually used to determine whether these IP addresses are in the same network segment. Consequently, we can distinguish the above two cases according to the subnet distribution of the IP addresses mapped by a domain name and exclude the invalid landmarks possibly belonging to CDN networks. This paper strictly excludes the invalid landmarks of the suspected CDN networks by checking whether the IP addresses of its domain name are distributed over multiple /24 network segments.

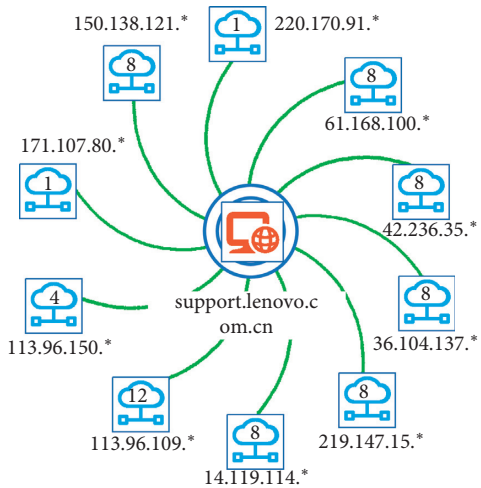



Figure 2: IP addresses of support.lenovo.com.cn.

*2.2.2. Multiple Domain Names Mapping to One IP Address.* This type of landmark is generally divided into three cases: virtual hosting [16], subdomain, and alternate domain name. A candidate landmark which belongs to virtual hosting provides the location of the user company or headquarters, but its server with an IP address provided by the ISP is deployed in the room of an Internet data center (IDC). The two have significant location differences and cannot be used for IP geolocation. For example, Figure 3 shows the IP address 110.173.196.3 of Alibaba, Hangzhou, Zhejiang Province, China, which maps to 260 domain names. The organizations corresponding to these domains are distributed in 20 provinces in China. Meanwhile, the subdomains and alternate domain names, other domain names except

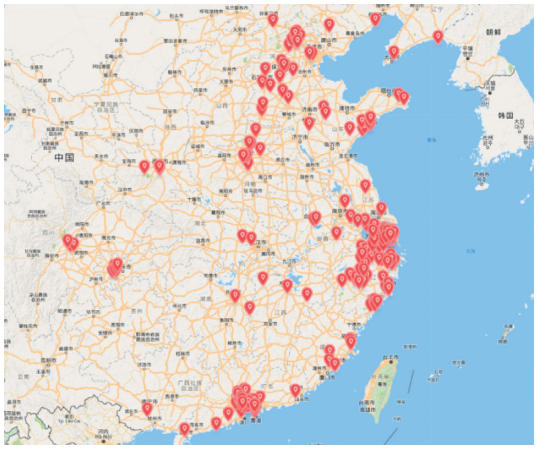


Figure 3: The claimed locations of 110.173.196.3's domain name.

the main domain name of a website, usually share the IP address and geographical location with the main domain name, and landmarks of this type can be used as reliable landmarks.

Evaluation idea: to begin with, we pick out the candidate landmarks which belong to virtual hosting by two ways. One is judging the distribution of the locations claimed by the candidate landmarks with different domain names sharing one IP address, as different domain names are mapped to one IP address by virtual hosting and the locations of their corresponding organizations are usually widely distributed. An alternative way is comparing the web page responses to an IP address request and the domain name request of one landmark, respectively, as the web page response to the domain name request is usually different from that to the IP address request for landmarks belonging to virtual hosting. And then, since one website generally has only several different subdomain or alternate domain names, we can obtain the number of domain names carried by the IP address of a candidate landmark and infer the probability that these domain names belong to the subdomain or alternate domain names and estimate the reliability of the landmark based on the domain name number.

*2.2.3. Multiple Domain Names Mapping to Multiple IP Addresses.* This kind of landmark is usually used by websites hosted on the cloud service. The accesses to this kind of websites are redirected to different physical hosts with unknown locations, which cannot be used as landmarks for IP geolocation. Figure 4 shows the complex mapping relationship between the 8 domain names and the 8 IP addresses of China Unicom cloud services.

Evaluation idea: same as in Section 2.2.1.

*2.2.4. One Domain Name Mapping to One IP Address.* This kind of landmark can bind a geographic location to the IP address, but this location is not necessarily the true location of the web server, and there are 2 kinds of invalid landmarks. A candidate landmark which belongs to server hosting hosts its server to the Internet data center. However, the geographic location claimed is the location of its user organization. Cloud server with an exclusive IP address has its IP address corresponding to one domain name uniquely, but the actual location is far apart.

Evaluation ideas: we compare the claimed information of a candidate landmark with the registration organization and its location information of the landmark's IP address and infer the reliability according to the conflict degree of two information. The organization and its location to which an IP address is assigned can be obtained by query Whois servers [17]. The consistency of the acquired information with the claimed information of a candidate landmark is a strong evidence of the accuracy of the landmark. When the acquired location information is not exactly the same as the claimed information of a candidate landmark, and if the name of the administrative district, such as the province or city, in the obtained location information is consistent with that in the claimed location information, the accuracy of the landmark can also be proved to a certain extent. Moreover, IP geolocation databases provide city-level granular IP location information. Studies have shown that this information is highly credible at the national and provincial levels and lower at the city level [18, 19]. Although the credibility of the city information of an IP address provided by IP geolocation databases is not high, the information can still be used to infer the reliability of a candidate landmark. Assuming that the server hosting or the cloud service provider serves $n$ cities, the probability that the landmark is incorrectly located to city A is $1/n$, and the probability that an IP geolocation database incorrectly declares the landmark's IP address in city A depends on the number of cities, say $m$, in the province in which it is located. Namely, when the administrative district information of a candidate landmark agrees with that provided by IP geolocation databases, the probability that both information is incorrect is no more than $1/mn$, which is far less than $1/n$. Consequently, the consistency between the location information claimed by the landmark and that provided by IP geolocation databases indicates the landmark's reliability. Therefore, by comparing the Whois registration information of a landmark's IP address with the claimed information of the landmark, this paper adjusts the landmark's reliability according to the similarity of the organization name, administrative district name in the location information, and domain name in two information and then compares the city name in the landmark's claimed location information with that provided by IP geolocation databases to correct the reliability of the landmark.

In summary, candidate landmark evaluation can be carried out as follows. Firstly, candidate landmarks which belong to the multibranch are excluded according to the distribution range of the location claimed by the candidate landmarks sharing a domain name. Secondly, candidate landmarks which belong to CDN networks and cloud services are excluded according to the subnet distribution of the IP address mapped by their domain name. Thirdly, candidate landmarks which belong to virtual hosting are excluded according to the distribution range of the locations of candidate landmarks sharing an IP address. Finally, the

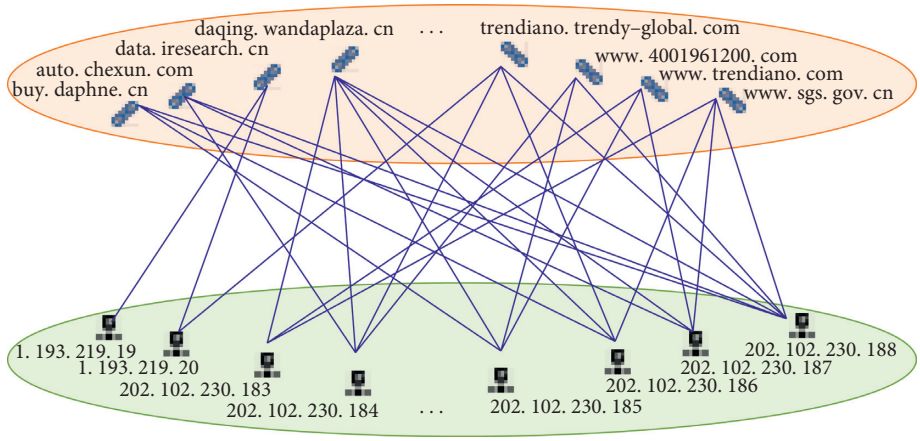



Figure 4: Mapping relationship of IP addresses and domain names in the cloud service of China Unicom.

reliability of a candidate landmark is calculated by the following information: the comparison result of the web page responses to its IP address request and its domain request, the number of domain names mapped by its IP address, the matching result with registration organization, and its location information in the administrative district level of its IP address.

## 3. The Evaluator Framework

In order to filter out reliable landmarks with quantified reliability from millions of candidate web-based landmarks, we propose the approach Evaluator. Evaluator classifies and filters the candidate landmarks belonging to CDN networks, virtual hosting, and cloud server according to the characteristics of invalid landmarks and then estimates the reliability of landmarks with layer-by-layer correction. An overview of Evaluator is shown in Figure 5.

The core input of Evaluator is a collection of candidate landmarks, which will be processed by the invalid landmark filtering module and the landmark reliability estimating module, respectively, before being reliably outputted.

The invalid landmark filtering module excludes obvious invalid landmarks according to the characteristics of invalid landmarks by four steps, domain location filtering, identical IP filtering, identical domain name filtering, and web page request filtering. And, the step of identical domain name filtering is based on the information of all the IP addresses mapped by a domain name, which are obtained by the method of DNS distributed query proposed in this paper.

The landmark reliability estimating module uses a variety of public information and services to evaluate the credibility of the surviving landmarks. This module includes two stages: reliability initialization and reliability correction. The initial value of the reliability for the landmarks reserved after invalid landmark filtering is set in the first stage, reliability initialization. The initialized reliability of a landmark is then corrected step by step through three steps, IP reverse lookup, Whois matching, and IP geolocation database verification, in the second stage, reliability correction. And, a reliability estimating model is completely constructed finally in this module. There are some important parameters in the reliability estimating model, and only after these parameters have been set reasonable values can this model be used to accurately to estimate reliabilities of landmarks. Therefore, an additional module of the Evaluator framework is presented, which is named model parameter assessing.

The model parameter assessing module, the support of the landmark reliability estimating module, evaluates the parameters of reliability estimating model by the gradient descent algorithm based on the initial parameter values and ends up with the final reasonable values of these parameters.

*3.1. Invalid Landmark Filtering.* The invalid landmark filtering contains four steps without strict sequence. In order to improve filtering efficiency, each processing step is performed in the ascending order of their time complexity sequence: (1) Domain location filtering: the multibranch is filtered by verifying the location scatter of candidate landmarks with the same domain name; (2) identical IP filtering: the virtual hosting and cloud server cases are filtered by verifying the location scatter of candidate landmarks with the same IP address; (3) identical domain name filtering: the CDN network is filtered by verifying IP segments in which the IP address of the landmark's domain name stayed; (4) web page request filtering: the virtual hosting and the cloud server survived in (2) are filtered by verifying whether the web page requests with the IP address and the domain name of the candidate landmark returns a nonempty and disparate result. The filtering rules of each step are given as follows, in which the meaning of the notations is as shown in Table 1.

*(1) Domain Location Filtering.* The candidate landmarks are grouped according to their domain name, and then the radius of each group's geographic location scatter is obtained. Finally, the groups with distribution radius exceeding $R_D$ are filtered. The filtering rule is shown as follows:

$$\forall_d \forall_l [dom(l) = d \wedge radi(loc(l)) > R_D] \longrightarrow del(l). \quad (1)$$

Considering the preciseness of the landmark location is one of the main factors determining the geolocation accuracy, hence the threshold $R_D$ of domain location filtering

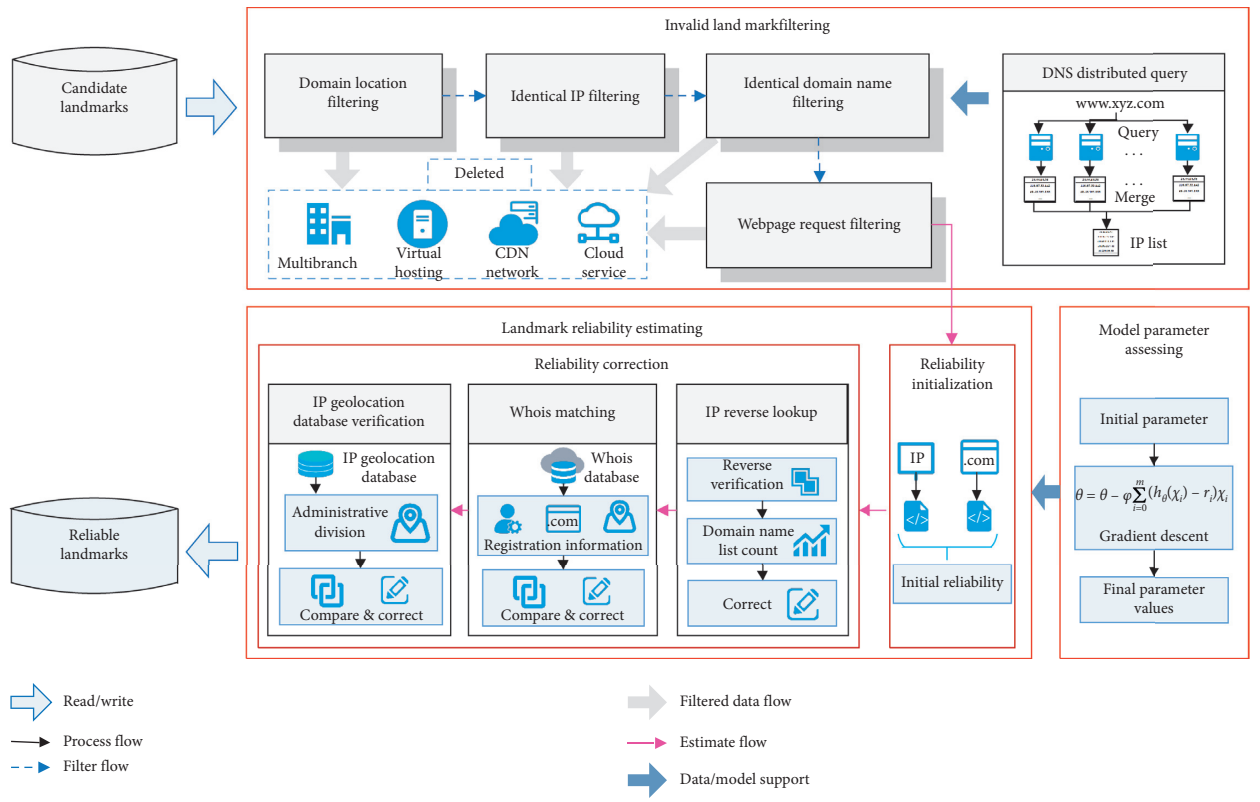



Figure 5: The overview of Evaluator.

Table 1: The expression in the filtering rule.

| Notation | Description |
|---|---|
| $l$ | A landmark |
| $dom(l)$ | The domain of landmark $l$ |
| $radi(S_{loc})$ | Scatter radius of location set $S_{loc}$ |
| $del(l)$ | Delete landmark $l$ |
| $segNum(S_{ip})$ | The number of IP segments in which IP addresses of $S_{ip}$ stay |
| $null$ | The result is empty |
| $d$ | A domain |
| $loc(l)$ | The geography location of landmark $l$ |
| $ip(l)$ | The IP address of landmark $l$ |
| $ipSet(d)$ | The IP address set mapping to $d$ |
| $res(i)$ | The result of web page request |

should be set according to the mean error that the geolocation method accepts. The geolocation accuracy of the street-level geolocation (SLG) algorithm on the residential network and the online map dataset is 2.25 km and 2.11 km [6], respectively, while the latest geolocation method [8] claims that its average positioning accuracy in Beijing reaches 3.8 km, consequently we set it to 2 km (less than the average error of the existing positioning accuracy). Similarly, when filtering the IP address with different declaration positions of the domain name, the scatter radius threshold of the declared location is also set to 2 km. The identical IP filtering also sets the same threshold.

*(2) Identical IP Filtering.* The candidate landmarks are grouped according to their IP addresses, and then the groups with geographic location scatter distribution radius exceeding $R_D$ will be deleted. The filtering rule is shown as follows:

$$\forall_{IP}\forall_l [ip(l) = IP \Lambda radi(loc(l)) > R_D] \longrightarrow del(l). \quad (2)$$

*(3) Identical Domain Name Filtering.* The /24 IP segments in which all IP addresses of a domain name stay are obtained, and then the candidate landmarks whose domain name's IP addresses occupy two or more IP segments will be excluded. The filtering rule is shown as follows:

$$\forall_l \{segNum[ipSetdom(l)]\} > 1 \longrightarrow del(l). \quad (3)$$

We want to obtain all the IP addresses of a domain name by the DNS query for identical domain name filtering, but the round-robin DNS [20] causes each DNS query to only

obtain the IP addresses in one record instead of all records, and the DNS cache [21] technology making repeated queries to the same server often returns the same result. Therefore, this paper proposes the DNS distributed query, an algorithm which uses servers all over the world to perform DNS query from different paths, and merges the responses to obtain a complete result. The steps are as follows:

Step 1: assign multiple DNS servers distributed around the world

Step 2: send a record query request of a domain name to the servers, respectively, and extract an IP address list from the response

Step 3: merge the IP address lists gained by each server and then obtain the query result after data deduplication

If these servers refer to the same superior server, it is possible to obtain the DNS cache data of the server and affect the query effect. Accordingly, these servers should be topologically dispersed as much as possible. If a query request is sent to enough DNS servers, all the IP addresses registered by a domain name would be obtained.

*(4) Web Page Request Filtering*. The home page of a landmark's organization is accessed through its domain name and IP address, respectively. The landmark will be deleted if the responses to two web page requests are different nonempty web pages. The filtering rule is shown as follows:

$$\forall_l \{res[ip(l)] \neq null \wedge res[dom(l)] \neq null \wedge res[ip(l)] \neq res[dom(l)]\} \longrightarrow del(l). \quad (4)$$

*3.2. Landmark Reliability Estimating.* The estimating landmark reliability first sets the initial value of reliabilities for the landmarks reserved after filtering in the reliability initialization stage and then corrects the reliabilities step by step in the reliability correction stage. In the reliability correction stage, some public information, e.g., IP mapping data, IP registration information, and IP Geolocation data, is collected, in which key attributes are extracted to be compared with the information claimed by the landmarks to correct their reliabilities step by step. IP mapping data are obtained by IP reverse lookup, in which the number of domain names mapped by an IP address is a key factor affecting the accuracy of landmarks. IP registration information is acquired by querying to Whois servers, in which the information, such as the organization, its location, and domain name of a landmark IP, is a strong evidence of the accuracy of the landmark information. IP geolocation data are obtained from some IP geolocation databases, in which the city information of a landmark IP is a reference for landmark accuracy. Consequently, based on the above three types of key attribute information, the reliability of a landmark is corrected, respectively, through the corresponding three steps: IP reverse lookup, Whois matching, and IP geolocation database verification, according to the importance of three information. Finally, a landmark with the reliability exceeding the threshold $\tau$ is considered to be a credible one.

*3.2.1. Reliability Initialization.* The initial reliability $r_0$ of a landmark is determined on the results of the web page request with its IP address and domain name, respectively:

$$r_0 = \begin{cases} \alpha, & res_{IP} = res_{domin} \neq null, \\ \beta, & res_{IP} = null, res_{domin} \neq null, \\ \gamma, & res_{domin} = null, \end{cases} \quad (5)$$

where $res_{IP}$ and $res_{domin}$ are the results of the web page request with its IP address and domain name, respectively, and *null* means a result without any content.

A landmark is considered to be reliable, if its $res_{IP}$ and $res_{domin}$ are nonempty and consistent, and its initial reliability is set to be $\alpha$. When $res_{IP}$ is null and $res_{domin}$ not, the landmark may be unreliable in virtual hosting and the cloud server case, or it may be a reliable one that does not support IP accessing, and its initial reliability is set to be $\beta$. A landmark is considered to be invalid or its service is temporarily unresponsive, if both its $res_{IP}$ and $res_{domin}$ are null, and its initial reliability is set to be $\gamma$, where we have $1 \geq \alpha > \beta > \gamma > 0$.

*3.2.2. Reliability Correction*

*(1) IP Reverse Lookup*. The distinct domain name list mapped by a landmark's IP address is obtained from the reverse lookup websites [22–26] and in which the landmark's domain name is added. Finally, the total number $n$ is gained, and the reliability of the landmark is corrected to be $r_1$ according to $n$:

$$r_1 = (1 - p_d) r_0 + p_d * f(n), \quad (6)$$

where $p_d$ is the reliability correction weight of IP reverse lookup and $f(n) \in (0, 1)$ denotes the reliability correction.

IP reverse lookup aspires an exact $n$, which means that the list of domain names obtained must be complete, i.e., all the domain names carried by the IP address should be obtained. However, we cannot obtain all domain names of many IP addresses or contain incorrect domain names because most IP reverse lookup websites have low coverage and untimely updates. To this end, this paper proposes reverse verification whose steps are as follows:

Step 1: the domain lists of the target IP address are gained from IP reverse lookup websites

Step 2: the complete domain name list is obtained after merging the obtained domain name lists and removing the duplicate domains

Step 3: DNS distributed query is performed for each domain name in the complete domain name list, and the domain name that cannot resolve the target IP is deleted, and the domain name list carried by the target IP address is obtained

*(2) Whois Matching*. The registration organization, domain name, and location information in the administrative district level corresponding to the landmark IP address are obtained by the Whois query and are compared with the

landmark information. Finally, the reliability is corrected according to the comparing results:

$$r_2 = (1 - p_w)r_1 + p_w * w_r, \tag{7}$$

$$w_r = k_c w_{c+} k_o w_{o+} (1 - k_c - k_o) w_d, \tag{8}$$

$$w_c = k_{co} w_{co+} k_{pr} w_{pr+} \left(1 - k_{co} - k_{pr}\right) w_{ci}, \tag{9}$$

where $w_r$ and $p_w$ indicate the reliability and the weight of Whois matching, respectively; $k_c$ and $k_o$ indicate the weight of location information in the administrative district level and the organization name registered in the Whois database, respectively; $w_c$, $w_o$, and $w_d$ with the value of 0-1, are the matching degree, respectively, of the registration location information in the administrative district level, organization, and domain name generated by the longest common subsequence (LCS) approach [27]; $k_{co}$ and $k_{pr}$ are the weights of the national and provincial administrative districts, respectively; $w_{co}$, $w_{pr}$, and $w_{ci}$ are the matching granularity of administrative district information between Whois registration information and landmark information, and their values are set as shown in Table 2.

*(3) IP Geolocation Database Verification.* Shavitt and Zilberman [12] provide insight into 7 IP geolocation databases by PoP analysis and find that most of the IP geolocation databases have quite high accuracy, and ip2location [28] has better coverage and accuracy. Accordingly, the administrative district information of a landmark is compared with ip2location DB90, and the correction coefficient $l_r$ is calculated according to the matching degree to obtain the landmark reliability $r$:

$$r = (1 - p_l)r_2 + p_l l_r, \tag{10}$$

$$l_r = k_{co} w_{Lco+} k_{pr} w_{Lpr+} \left(1 - k_{co} - k_{pr}\right) w_{Lci}, \tag{11}$$

where $p_l$ is the weight of the IP geolocation database and $w_{Lco}$, $w_{Lpr}$, and $w_{Lci}$, with the value shown in Table 2, are the matching granularity of administrative district information between information in the IP geolocation database and landmark information.

Combining equations (5)–(11) gives us Evaluator's final model of estimating landmark reliability:

$$\begin{aligned} r = {} & (1 - p_l)(1 - p_w)(1 - p_d)r_0 + (1 - p_l)(1 - p_w)p_d e^{1-n} \\ & + (1 - p_l)p_w k_c k_{co} w_{co} \\ & + (1 - p_l)p_w k_c k_{pr} w_{pr} + (1 - p_l)p_w k_c \left(1 - k_{co} - k_{pr}\right) w_{ci} \\ & + (1 - p_l)p_w k_o w_o \\ & + (1 - p_l)(1 - k_c - k_o)p_w w_d + p_l k_{co} w_{Lco} + p_l k_{pr} w_{Lpr} \\ & + p_l \left(1 - k_{co} - k_{pr}\right) w_{Lci}, \end{aligned} \tag{12}$$

where the values of $n$, $w_{co}$, $w_{pr}$, $w_{ci}$, $w_o$, $w_d$, $w_{Lco}$, $w_{Lpr}$, and $w_{Lci}$ are obtained in the steps of reliability initialization, IP reverse lookup, Whois matching, and IP geolocation database verification, respectively, and $\alpha$, $\beta$, $\gamma$, $p_l$, $p_w$, $p_d$, $k_c$, $k_{co}$, $k_{pr}$, and $k_o$ are the parameters of the estimating model.

Table 2: The matching granularity of administrative district information.

| Country | Province | City | $w_{co}/w_{Lco}$ | $w_{pr}/w_{Lpr}$ | $w_{ci}/w_{Lci}$ |
|---|---|---|---|---|---|
| Not match | Match or not | Match or not | 0 | 0 | 0 |
| Match | Not match | Not match | 1 | 0 | 0 |
| Match | Match | Not match | 1 | 1 | 0 |
| Match | Match | Match | 1 | 1 | 1 |

*3.3. Model Parameter Assessing.* In order to assess reasonable parameters of the reliability estimating model, we first manually label $m$ candidate landmarks within the area to be evaluated, calculate the error between the geographic location declared therein and the actual location, and then generate the landmark credibility $r$ according to the rules in Table 3. Meanwhile, we obtain the values of attributes, i.e., $n$, $w_{co}$, $w_{pr}$, $w_{ci}$, $w_o$, $w_d$, $w_{Lco}$, $w_{Lpr}$, and $w_{Lci}$ of the landmarks, and then based on these attributes, we assess the estimating model parameters $\alpha$, $\beta$, $\gamma$, $p_l$, $p_w$, $p_d$, $k_c$, $k_{co}$, $k_{pr}$, and $k_o$ by the gradient descent algorithm.

For convenience of description, this paper simplifies formula (12) to $r = h_\theta(x) = \theta^T X$, where $\theta$ represents the parameter to be evaluated and $X$ represents the variables, i.e., $n$, $w_{co}$, $w_{pr}$, $w_{ci}$, $w_o$, $w_d$, $w_{Lco}$, $w_{Lpr}$, and $w_{Lci}$. The steps of parameter estimation are as follows:

Step 1: set the initial value $\theta_0$ of $\theta$.

Step 2: adjust the value of $\theta$ according to the $m$ group data so that the objective function $J(\theta) = (\partial/\partial\theta)(1/2)\sum_{i=0}^{m}(h_\theta(x) - r)^2$ decreases in the direction of the gradient. The adjustment formula is as follows:

$$\theta = \theta - \delta \frac{\partial}{\partial\theta} J(\theta) = \theta - \delta \sum_{i=0}^{m} (h_\theta(x_i) - r_i)x_i, \tag{13}$$

where $\delta$ denotes the step size and is set to 0.01 empirically.

Step 3: the change of $J(\theta)$ before and after the adjustment of $\theta$ is tested. When $|J(\theta_j) - J(\theta_{j-1})| > \tau_{Th}$, Step 2 will be iterated; otherwise, the adjustment is ended and $\theta_j$ is set to be the final estimating model parameters, where $\tau_{Th}$ is a threshold value and is set to 0.13 empirically.

To achieve the goal of rapid convergence, we set the initial values of parameters in the estimating model empirically as shown in Table 4.

## 4. Experiments and Results

In order to verify the validity of Evaluator, we evaluated candidate landmarks collected from several cities and tested the performance of gained reliable landmarks. All experiments were run on a 256 G RAM, 1.92 GHz CPU server using the Windows Server 2012 R2 Stander. All Evaluator algorithms are implemented in C# and run in the VS2015 environment.

Table 3: Landmark deviation and confidence comparison.

| Geography error | Reliability | Geography error | Reliability |
|---|---|---|---|
| <2 km | 1 | >20 km (same city) | 0.2 |
| 2–5 km | 0.9 | Different city but same province | 0.1 |
| 5–20 km | 0.5 | Different province | 0 |

Table 4: The initial values of parameters in the estimating model.

| Parameter name | $\alpha$ | $\beta$ | $\gamma$ | $p_l$ | $p_w$ | $p_d$ | $k_c$ | $k_{co}$ | $k_{pr}$ | $k_o$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Initial value | 0.95 | 0.5 | 0.1 | 0.5 | 0.2 | 0.5 | 0.2 | 0.01 | 0.1 | 0.4 |

*4.1. Data Description.* The experimental data are divided into two parts: candidate landmark dataset consisting of candidate landmarks from 5 provinces/cities in China and the United States, used for evaluation and parameter estimation of Evaluator; verification dataset consisting of landmarks with known location in partial cities, used for verifying the performance of Evaluator.

*4.1.1. Candidate Landmark Dataset.* We collected the candidate landmark dataset from Google Maps in February 2018. It contains organization information of Beijing, Zhengzhou, Hong Kong, New York, and Los Angles and IP addresses resolved from their domain name, as shown in Table 5, where candidate landmarks refer to all the landmarks obtained, domain names mean the distinct domain names included in the candidate landmarks, and IP addresses obtained refers to the distinct IP addresses of the candidate landmarks.

*4.1.2. Verification Dataset*

*Terminal Dataset.* 100 IP addresses used by the terminal with known detailed locations from the above five cities are, respectively, selected and used as verification datasets to test the geolocation effect of the reliable landmarks from Evaluator.

*WIFI Dataset.* Since the WIFI signal transmission distance is limited (not more than 100 m), the distance between the mobile terminal and the WIFI hotspot is relatively close. Accordingly, the WIFI-based landmark obtained by the mobile terminal has high accuracy. We manually collected 146 WIFI-based landmarks in 6 districts of Zhengzhou City, Henan Province, and 119 in Beijing to test the geolocation effect of Evaluator landmarks.

*4.2. The Determination of Parameters*

*4.2.1. DNS Distributed Query Parameters.* We evenly selected 10,000 domain names in 629 cities distributed in 63 provinces of China, Japan, and the US and then adopt the DNS distributed query approach by inquiring 120 DNS servers scattered in six continents (excluding Antarctica) and 15 countries. We find that the number of IPs obtained from the 23 servers is no longer increased. Namely, we can get all the IP addresses of a domain with 23 DNS servers repeatedly resolving it. Consequently, 23 DNS servers with fast processing speed and stable performance are selected for the domain name resolution. The country distribution is shown in Table 6.

Table 5: Candidate landmark dataset.

| City | Candidate landmarks | Domain names | IP addresses obtained |
|---|---|---|---|
| Beijing | 57452 | 14909 | 16360 |
| Zhengzhou | 11707 | 3025 | 2900 |
| Hong Kong | 19363 | 10952 | 12432 |
| New York | 340713 | 193344 | 96978 |
| Los Angeles | 86420 | 45206 | 41811 |

Table 6: The country distribution of DNS servers.

| DNS server | Country |
|---|---|
| 114.114.114.114 | China |
| 119.29.29.29 | China |
| 223.5.5.5 | China |
| 205.252.144.228 | China |
| 180.76.76.76 | China |
| 202.12.27.33 | Japan |
| 168.126.63.1 | Korea |
| 202.138.103.100 | India |
| 8.26.56.26 | United States |
| 208.67.222.222 | United States |
| 8.8.8.8 | United States |
| 4.2.2.1 | United States |
| 209.166.160.36 | Canada |
| 24.153.22.15 | Canada |
| 77.88.8.1 | Russia |
| 84.200.69.80 | Germany |
| 194.167.105.130 | France |
| 193.0.14.129 | UK |
| 85.255.112.161 | Ukraine |
| 122.201.69.71 | Australia |
| 203.166.97.9 | Australia |
| 200.221.11.100 | Brazil |
| 41.138.66.60 | South Africa |

*4.2.2. Reverse Verification Parameters.* For the purpose of testing the results of reverse verification, we randomly selected 1000 IP addresses, which host websites in mainland China, and applied reverse verification by chinaz.com, ip138.com, aizhan.com, hackertarget.com, and viewdns.info. Finally, we verified the accessibility of the domain names, as shown in Table 7.

Table 7 shows that reverse verification has greatly improved both the IP coverage rate and the domain name accessibility, in spite of the total number of domain names obtained, which are even fewer than the ones some websites returned. This is because there are a large number of expired invalid domain names in the results of these websites, and we have filtered them in reverse verification, which has resulted

Table 7: The result of IP reverse lookup.

| Websites | Resultant IP number | Domain name number | Accessibility (%) |
|---|---|---|---|
| chinaz.com | 336 | 2563 | 79 |
| ip138.com | 981 | 21870 | 83 |
| aizhan.com | 676 | 56886 | 81 |
| mxtoolbox.com | 881 | 98562 | 85 |
| domaintools.com | 915 | 74195 | 89 |
| Reverse verification | 982 | 85845 | 96 |

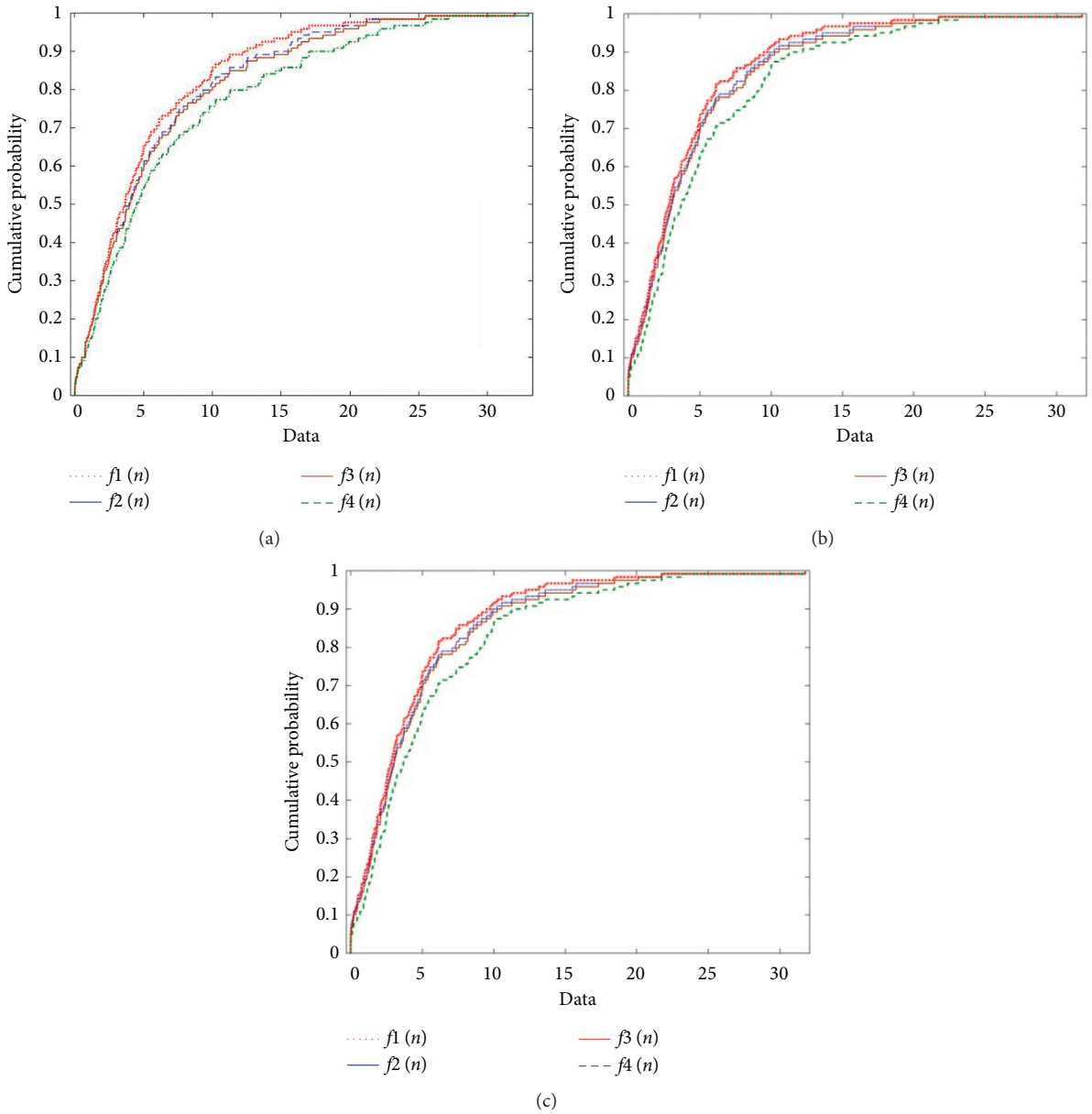



Figure 6: The influence of $f(n)$ on the geolocation effect. (a) $\tau = 0.3$. (b) $\tau = 0.5$. (c) $\tau = 0.8$.

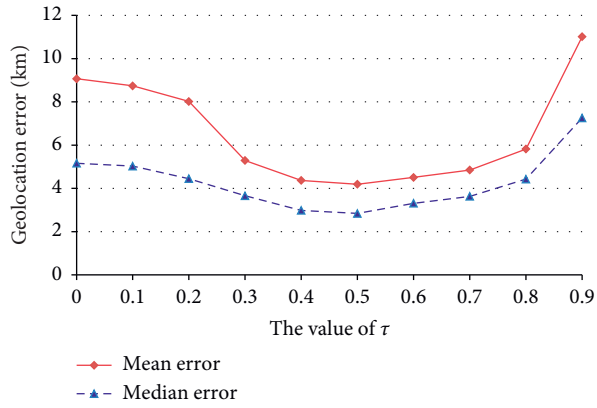



Figure 7: The influence of $\tau$ on the geolocation effect.

in a decrease in the magnitude of domains obtained, but acquired better timeliness and reliability.

#### 4.2.3. Reliability Estimating Model Parameters

*(1) Estimating Function f(n).* The reliability of a landmark decreases as the number of domain names mapped by its IP address grows. Accordingly, $f(n)$ is a monotonically decreasing function, which conforms to $f(1) = 1$ and has $f(n) \longrightarrow 0$ when $n \longrightarrow \infty$. In order to test the influence of $f(n)$ on the evaluation effect, we set $f(n)$ as an inverse proportional function $f1(n) = n^{-1}$, exponential function $f2(n) = e^{1-n}$, logarithmic function $f3(n) = (1/(1 + \log n))$, and arctangent function $f4(n) = ((\pi/2) - \arctan n)/((\pi/2) - \arctan 1)$, respectively, to evaluate the candidate landmarks in Beijing. Finally, the reliable landmarks ($\tau = 0.3, 0.5, 0.8$) are used to locate the 119 IPs in the WIFI dataset. The geolocation results are shown in Figure 6.

It can be seen from Figure 6 that *f(n)* has a significant influence on the geolocation effect, and the best result is obtained when *f(n)* takes the exponential function, and the median error is 5.2 km, 4.19 km, 4.32 km at $\tau = 0.3, 0.5$, and 0.8, respectively. When *f(n)* takes the inverse proportional function, the median error is 5.81 km, 4.45 km, and 4.44 km at $\tau = 0.3, 0.5$, and 0.8, respectively. When *f(n)* takes the arctangent function, the median error is 6.00 km, 4.51 km, and 4.61 km at $\tau = 0.3, 0.5$, and 0.8, respectively. The geolocation result of the 2 are basically the same and slightly worse than the exponential function. When *f(n)* takes the logarithmic function, the median error is 7.08 km, 5.48 km, and 5.05 km at $\tau = 0.3, 0.5$, and 0.8, respectively, which is obviously worse than the first three. In addition, it can be found that the influence of *f(n)* on the evaluation result gradually decreases with the increase in the reliability. This is because the IP addresses of the candidate landmarks with high reliability are often mapping to only one domain name, and the value of *f(n)* does not affect the evaluation results.

*(2) Credibility Threshold $\tau$.* In order to determine the value of the reliability threshold $\tau$, this paper uses Evaluator to evaluate the candidate landmarks in Beijing and then locates the IP address in the WIFI dataset with the reliable landmarks under different values of $\tau$. The geolocation result is shown in Figure 7.

As shown in Figure 7, when $\tau$ is between 0.4 and 0.7, the landmark geolocation error stays in an ideal interval, and the best choice of the reliability threshold is 0.5.

*4.3. The Performance of Evaluator.* In order to verify the geolocation effect of the reliable landmarks from Evaluator, this paper uses Evaluator and LVM to evaluate the candidate landmark dataset, respectively. Then, we set the IP addresses in the terminal dataset to be geolocation targets and locate them with the candidate landmarks (after city-level screening), the landmarks trusted by LVM, and the landmarks trusted by Evaluator, respectively. The geolocation performance is shown in Table 8 and Figure 8. Table 8 lists the mean geolocation error of three methods, respectively, in five cities, and Figure 8 shows the cumulative probability of their geolocation errors.

Table 8 and Figure 8 suggest the following:

(1) When we geolocate an IP address in the terminal dataset with the landmarks evaluated by Evaluator or LVM, the mean geolocation errors, respectively, in Beijing, Zhengzhou, Hong Kong, New York, and Los Angeles are significantly less than the results of the method with the candidate landmarks. This shows that the methods Evaluator and LVM have greatly improved the geolocation accuracy, and Evaluator has obvious advantages over LVM in the meantime.

(2) It can be seen that the geolocation result in the USA outperformed than that in China. This is because the stagnation of map service development in China has been slow to update after the Google company withdrew from the Chinese market. We can also find this from the quantitative distribution of the candidate landmarks mined.

Since SLE requires candidate landmarks to be proved multiple times from 22:30 to 06:30 (next day), the time cost is large. Accordingly, we only evaluate Zhengzhou and Beijing with SLE. In order to compare the geolocation result of the two methods, this paper uses the WIFI dataset as the geolocation targets and locate it with landmarks trusted by

Table 8: The mean geolocation error (ME) of three methods in different cities.

| City name | ME of Evaluator (km) | ME of LVM (km) | ME of candidate landmarks |
|---|---|---|---|
| Beijing | 4.98 | 8.88 | 16.23 |
| Zhengzhou | 5.11 | 9.24 | 17.02 |
| Hong Kong | 3.91 | 7.30 | 9.32 |
| New York | 4.01 | 7.65 | 10.54 |
| Los Angeles | 4.70 | 8.04 | 10.73 |

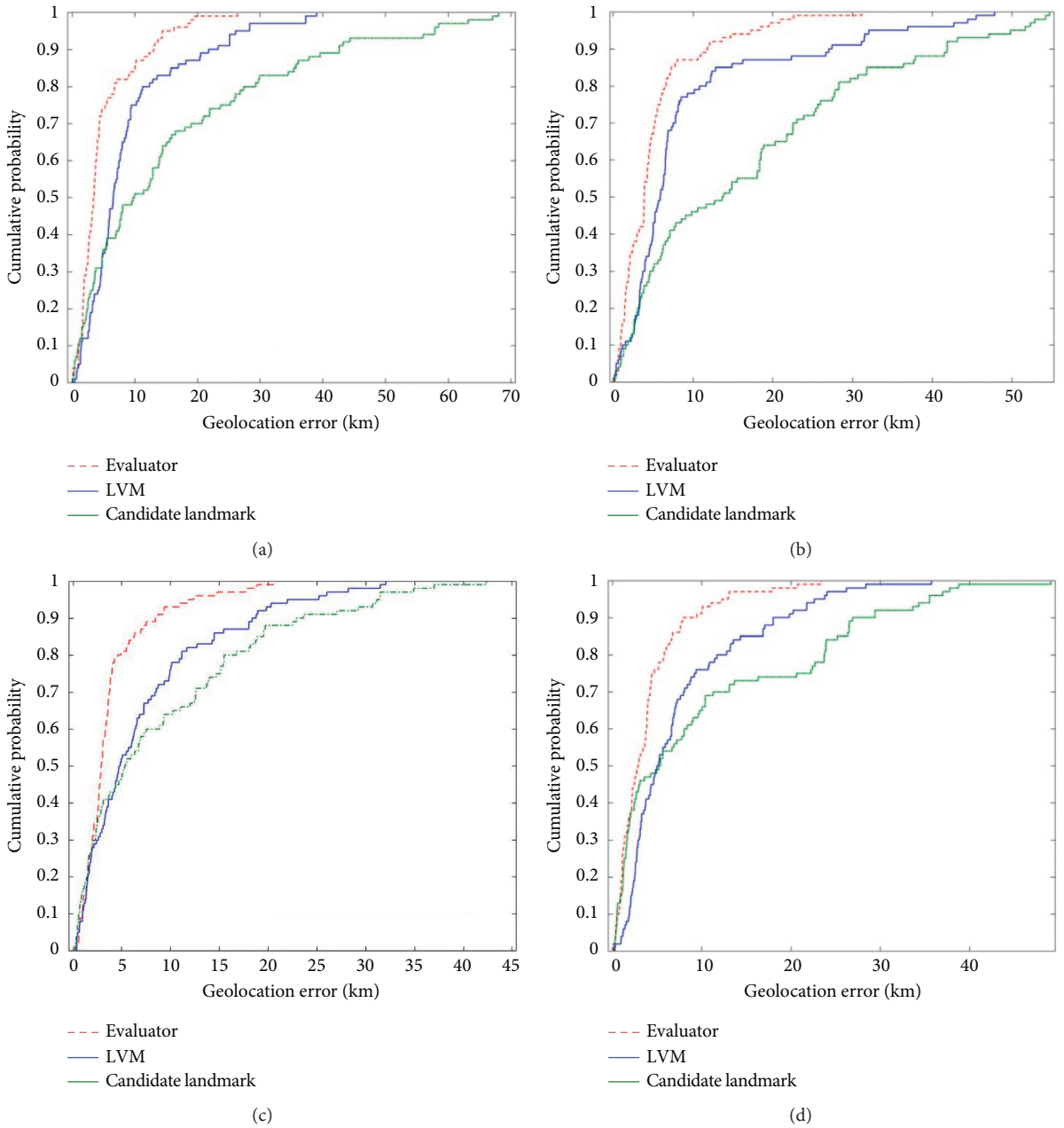
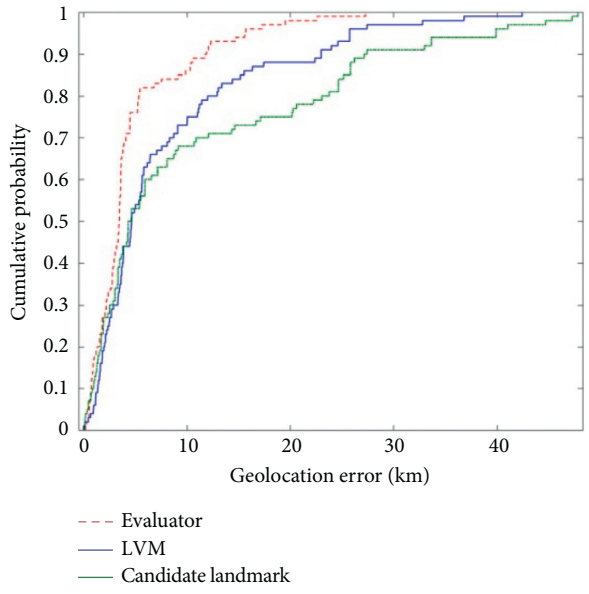



Figure 8: Continued.

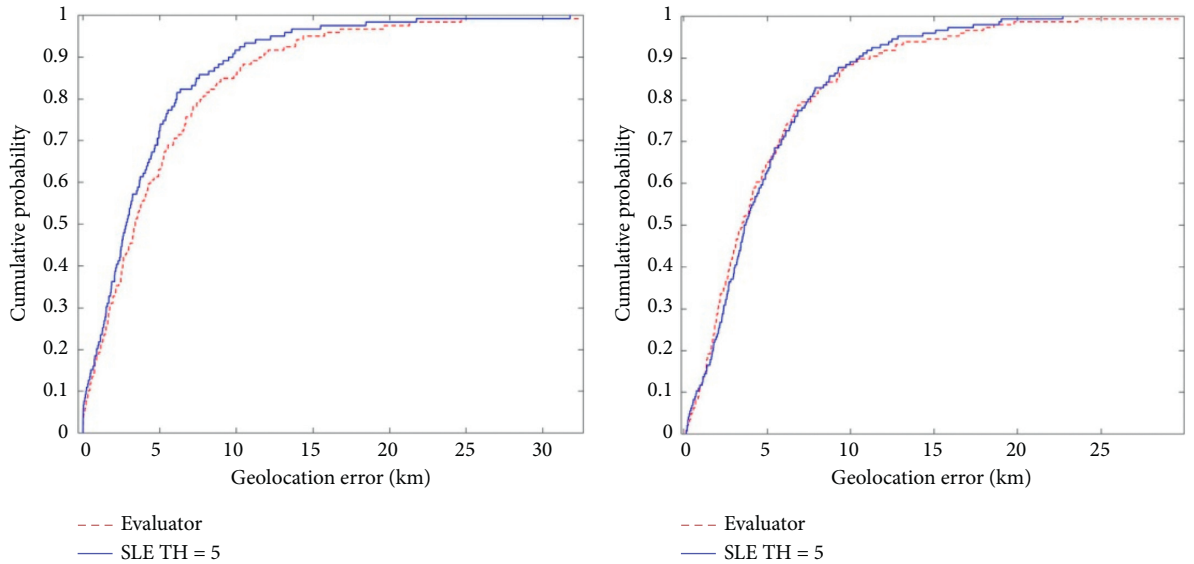



(e)

Figure 8: The geolocation result of 3 landmark sets. (a) Beijing. (b) Zhengzhou. (c) Hong Kong. (d) New York. (e) Los Angeles.

Table 9: The mean geolocation error (ME) of Evaluator and SLE.

| City name | ME of Evaluator (km) | ME of SLE (km) |
|---|---|---|
| Beijing | 4.98 | 4.71 |
| Zhengzhou | 5.02 | 4.95 |




(a) (b)

Figure 9: The geolocation result of Evaluator and SLE. (a) Beijing. (b) Zhengzhou.

Evaluator and landmarks trusted by SLE (TH = 5), respectively. Table 9 shows the mean geolocation error of Evaluator and SLE in two cities. Figure 9 depicts the geolocation results of Evaluator and SLE. At the same time, we calculate the rate of the landmarks that can be evaluated, respectively, by SLE in the candidate landmarks of two cities, and the evaluable landmark coverage rate is shown in Table 10.

Table 10: The evaluable landmark coverage rate of SLE.

| City name | Candidate landmark number | Evaluable landmark number | Evaluable coverage rate (%) |
|---|---|---|---|
| Beijing | 11099 | 2036 | 10.34 |
| Zhengzhou | 4921 | 776 | 15.77 |

From Table 9 and Figure 9, we can see that the geolocation accuracy of Evaluator is at the same level with SLE. However, the evaluable coverage rates of candidate landmarks of SLE in two cities are both less than 16% as shown in Table 10, while Evaluator can cover all candidate landmarks. Evaluator improves the average evaluable coverage rate of the two cities by 82.45%, and the improvement will be more significant in areas with low candidate landmark density.

## 5. Conclusions

Based on the analysis of invalid web-based landmarks, we introduced Evaluator, a web-based landmark evaluation approach, which adopts a multilevel decision-making method to hierarchically filter invalid landmarks by their common characteristics and uses the public data, e.g., public service and third-party free database, to assess the reliability of landmarks. In this paper, we propose a DNS distributed query which breaks through the limitations to obtain a much better result and introduce reverse verification to significantly improve the coverage and accuracy of the IP reverse lookup result. Meanwhile, we use the gradient descent method to determine the estimating model parameters of relevant regions with a small number of manually labelled data and improve the robustness and reliability of Evaluator. Experiment indicates that Evaluator significantly improves the current method by reducing the geolocation error and increasing the evaluation coverage.

However, there are still a few invalid landmarks missed by Evaluator, which limit the further improvement of the geolocation accuracy of Evaluator. Our future work will focus on combining topology detection and delay measurement to further improve the accuracy of web-based landmark evaluation.

## Data Availability

The datasets analysed or generated during the study cannot be made publicly available for some reasons. For readers interested in this article, please contact us and we may provide partial data appropriately as the case may be.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

The work presented in this paper was supported in part by the National Key R&D Program of China under Grants 2016QY01W0105 and 2016YFB0801303, the National Natural Science Foundation of China under Grants U1636219, 61602508, 61772549, U1736214, and 61572052, and the Plan for Scientific Innovation Talent of Henan Province under Grant 2018JR0018.

## References

[1] J. A. Muir and P. C. V. Oorschot, "Internet geolocation," *ACM Computing Surveys*, vol. 42, no. 1, pp. 1–23, 2009.

[2] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4, pp. 173–185, 2001.

[3] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of Internet hosts," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.

[4] E. Katz-Bassett, J. P. John, A. Krishnamurthy et al., "Towards IP geolocation using delay and topology measurements," in *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, pp. 71–84, Rio de Janeriro, Brazil, October 2006.

[5] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: A comprehensive framework for the geolocalization of internet hosts," in *Proceedings of the 4th USENIX Conference on Networked Systems Design & Implementation*, p. 23, Cambridge, MA, USA, April 2007.

[6] Y. Wang, D. Burgener, M. Flores et al., "Towards street-level client-independent IP geolocation," in *Proceedings of the USENIX Conference on Networked Systems Design and Implementation (NSDI)*, vol. 11, p. 27, Boston, MA, USA, March 2011.

[7] H. Jiang, Y. Liu, and J. N. Matthews, "IP geolocation estimation using neural networks with stable landmarks," in *Proceedings of the IEEE Computer Communications Workshops*, pp. 170–175, San Francisco, CA, USA, April 2016.

[8] Z. Wang, Y. Chen, H. Wen, L. Zhao, and L. Sun, "Discovering routers as secondary landmarksfor accurate IP geolocation," in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 1–5, Toronto, Canada, September 2017.

[9] J. Chen, F. Liu, X. Luo et al., "A landmark calibration-based IP geolocation approach," *EURASIP Journal on Information Security*, vol. 2016, pp. 1–11, 2016.

[10] https://news.netcraft.com/archives/category/web-server-survey/.

[11] C. Guo, Y. Liu, W. Shen et al., "Mining the web and the internet for accurate ip address geolocations," in *Proceedings of the IEEE INFOCOM-The 28th Conference on Computer Communications*, pp. 2841–2845, Rio de Janeiro, Brazil, April 2009.

[12] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

[13] T. Wang, K. Xu, J. Song et al., "An optimization method for the geolocation databases of Internet hosts based on machine learning," *Mathematical Problems in Engineering*, vol. 2015, Article ID 972642, 17 pages, 2015.

[14] G. Zhu, X. Luo, F. Liu, and J. Chen, "An algorithm of city-level landmark mining based on internet forum," in *Proceedings of the International Conference on Network-Based Information Systems*, pp. 294–301, Taipei, Taiwan, September 2015.
[15] R. Li, Y. Sun, J. Hu, T. Ma, and X. Luo, "Street-level landmark evaluation based on nearest routers," *Security and Communication Networks*, vol. 2018, Article ID 2507293, 12 pages, 2018.
[16] https://en.wikipedia.org/wiki/Virtual_hosting.
[17] L. Daigle, "WHOIS protocol specification," IETF, Fremont, CA, USA, RFC 3912, 2004.
[18] D. Komosný, M. Vozňák, and S. U. Rehman, "Location accuracy of commercial IP address geolocation databases," *Information Technology & Control*, vol. 46, no. 3, pp. 333–344, 2017.
[19] H. Li, Y. He, R. Xi et al., "A complete evaluation of the chinese IP geolocation databases," in *Proceedings of the International Conference on Intelligent Computation Technology and Automation*, pp. 13–17, IEEE, Nanchang, China, June 2015.
[20] https://en.wikipedia.org/wiki/Round-robin_DNS.
[21] https://en.wikipedia.org/wiki/Name_server#Caching_name_server.
[22] http://s.tool.chinaz.com/same.
[23] http://site.ip138.com/.
[24] https://dns.aizhan.com/.
[25] https://hackertarget.com/reverse-dns-lookup/.
[26] https://viewdns.info/reverseip/.
[27] K. Hakata and H. Imai, "Algorithms for the longest common subsequence problem," *Genome Informatics*, vol. 24, pp. 664–675, 1977.
[28] https://www.ip2location.com/databases/.