IP Geolocation Accuracy Evaluation Based on Crowdsourcing

Guangyu Zhu¹, Guoming Ren², Xiang Li¹, Xiaoye Li¹,* and Yongpeng Ti²

¹National Computer System Engineering Research Institute of China, Beijing 100083, China
²Zhengzhou Aiwen Computer Technology Co., Ltd. Zhengzhou, Henan Province 450016, China

*Corresponding author email: xiaoyel2@illinois.edu

Abstract. IP geolocation is a crucial technology in the study of many fields, such as network measurement, fraud detection, and location-based advertising, etc. For over a decade, numerous researches and applications in this area have gained significant advances. In recent years, since more and more physical devices are connected to cyberspace, evaluating the accuracy of IP geolocation has gained much more attentions. In this paper, we first illustrate the concept and applications of the IP geolocation. Then, we design and propose an evaluation method on IP geolocation accuracy based on crowdsourcing. We evaluate the accuracy of a popular high-precision IP geolocation system and find out it reaches 98.88% correct on city-level geolocation, and the median error distance of street-level geolocation is 0.13km. Finally, the future development direction of IP geolocation accuracy evaluation is discussed.

Keywords: IP geolocation; Accuracy; Evaluation; Crowdsourcing.

1. Introduction

1.1. The Concept of IP Geolocation
IP geolocation is used to determine the exact geographic address of a target device by its IP address[1]. IP geolocation is widely used by target advertising, online fraud detection, cyber-attacks attribution and so on[2]. In the network layer, physical devices can be off-limits from communication limitations of physical space and then achieve point-to-point interconnection in cyberspace. An IP address is the identifier of a host to access the Internet in cyberspace. Geographical location information is an important entity spatial information. Through IP geolocation technology, the target IP can be connected with the geographical location of the equipment and its various value factors based on its location.

1.2. The Applications of the IP Geolocation
Location-related value includes commercial value, scientific value, security value and other aspects. IP geolocation is mainly used in targeted content push, network performance optimization, social network, Internet of things, network security and other fields.

In the field of content push, ISPs push the nearby content related to the region to customers according to their IP addresses. For example, the browser pushes local news to its customer, WeChat pushes the nearby store to its customer, and Bing provides the special search according to the logged-in user [3].

In the field of social networking, nearby content push and nearby person push are provided according to the customer’s common login IP location, i.e., WeChat will recommend nearby friends to its user.
In the field of network performance optimization, location information provided by IP-based geolocation offers better network communication performance for network nodes. For example, adjacent peers can be provided to nodes in the P2P network, thus reducing the long-distance communication cost [4].

Network security issues exist in the above fields so that IP location can be applied to customer identification, IP camouflage detection as well as physical system attack and protection. Through the corresponding location of the user’s IP, ISP verifies the user identities and protects the user privacy data. The IP geolocation can associate the IP address of the visitor with its location, and verify whether the IP address is a fraudulent IP.

1.3. IP Geolocation Accuracy

In the field of IP geolocation, error distance between IP geolocation results and its real location is used as the criterion of IP geolocation accuracy. There are several typical IP geolocation algorithms, such as Octant[5], CBG[6], TBG[7], SLG[8] and GBRF[9]. The comparison results are shown in table 1.

| Algorithm | Experimental Data Sets | Geolocation Error (KM) |
|-----------|------------------------|------------------------|
| Octant    | 51 PlanetLab nodes, as vantage point and target | median error: 35 maximum error: 277 |
|           | Two types of nodes: 42 hosts in the western European RIPE network, 95 hosts in the American AMP network | |
| CBG       | American AMP network. When testing the CBG geolocation performance, selecting one of the same kind of nodes as the target, and the rest as the vantage point | average error: 78, 182 median error: 22, 95 |
|           | Vantage point: 68 PlaneLab nodes in North America | |
| TBG       | Target set: (1)Non-academic, 22 Sprint network nodes; (2)Academic, 128 hosts from universities in the United States | average error: 194, 253 |
|           | Vantage point: 163 publicly available ping servers, 136 traceroute servers | |
| SLG       | Target set: 88 PlaneLab nodes; 72 <IP, postal address > peer by residents | average error: 0.69, 2.25 maximum error: 5.24, 8.1 |
|           | by residents | |
| GBRF      | Vantage point: 2 probe sources autonomously deployed in Beijing and Chengdu | City-level accuracy>92% |
|           | Target set: a total of 1,024 target IP located in Henan province | |

From the table 1, the number of the target set used by the typical geolocation methods, i.e., Octant, CBG, TBG, SLG and GBRF, respectively is 51, 137 (42+95), 150 (22+128), 160 (88+72) and 1,024 in the experimental test. So these typical localization algorithms use a small number of target sets, and the target with the largest number is also a thousand level set.

Crowdsourcing refers to the practice that a company or organization outsources the work tasks performed by employees in the past to non-specific (and typically large) public volunteers in a free and voluntary way. Crowdsourcing is usually done by individuals, however, if related to a collaborative task, it could also appear in relying on open source individual production. Jeff Howe, a reporter of Wired magazine, first introduced the concept of crowdsourcing in June 2006.

Based on this, aiming at collecting a large number of IP with location information as a test target set of IP geolocation, we propose an evaluation method based on crowdsourcing, build a crowdsourcing
platform to collect IP with location information, and evaluate the accuracy of a currently popular high precision IP geolocation system based on this platform.

2. Design of Crowdsourcing Platform

2.1. Architecture Diagram
The Architecture diagram of the crowdsourcing platform is shown in figure 1. Combining with the map component, end development framework (mainly composed of wxss, JS, wxml, JSON, etc.), general development component, GPS authorization and login authorization of online social platform, the front-end of crowdsourcing platform realizes visual map display. The map can be adaptive according to the geolocation accuracy, providing high-precision IP geolocation result information of the user's IP; at the same time, the crowdsourcing platform collects the user's GPS position information with their authorization, and provides GPS calibration function, which is convenient for users to judge the accuracy of IP geolocation.

Through token authentication with online social platform server, the back-end implements user-based access control restrictions. For users who can access, we offer a high precision IP geolocation service by querying the database, recording the geolocation log, and storing the geolocation log in a database. Through the configuration items provided by the online social platform official website, the back-end security call service host is set to realize the business technology connection between the front end and the back end.

2.2. Functions
Figure 2 is the geolocation result of the crowdsourcing platform, which mainly implements the following three functions.
(1) Get the public network IP address of the current Internet device, i.e., the Internet IP is 115.60.134.193 in figure 2.
(2) By querying the IP geolocation database with ultra-high precision, we get the detailed geographic information and IP geolocation extension information of the current IP, and visualize them on the map. Its location center is shown in figure 2. The detailed geographic location information includes continent, country, province, zip code, and time zone, with values of Asia, China, Henan province, 450004, and UTC + 8 (representing time zone 8) in figure 2.
IP geolocation extension information includes geolocation precision, operator, application scenario, accuracy and consistency. Among IP geolocation extension information, the geolocation precision of the return value is street, operator is China Unicom, application scenario is residential user, accuracy is five stars and consistency is four stars, as shown in figure 2.

![Geolocation result of crowdsourcing platform.](image)

Figure 2. Geolocation result of crowdsourcing platform.

(3) In the case of access to the user authorization, we obtain the user's GPS location and take the GPS location as the real set for testing the accuracy of IP geolocation, which is represented by a small circle with an arrow in Figure 2. The crowdsourcing platform displays the user location and geolocation results to the user, assisting the user to determine the accuracy of the geolocation results.

3. Experiment

3.1. Experimental Data

First, we screened the IP that appeared from October 1 to December 31 in 2019 and obtained 301,119 IP containing GSP location information.

Secondly, because the use scope of mobile network IP is usually concentrated in a province-level region and their geolocation precision is also at province-level, it is not necessary to consider when conducting the experiment of city consistency and error distribution of street-level IP. Therefore, we obtain 296,626 non-mobile networks IP by filtering non-mobile network IP. In figure 3, a) and b) are respectively the usage range and geolocation results of the same China mobile 223.104.106.233. In Figure 3, the use scope of 223.104.106.233 is concentrated in Henan province, with a small number in Tianjin city, Chengdu city, Sichuan province and Wuhan city, Hubei province. Meanwhile, the
geolocation precision of 223.104.106.233 is at the province-level, and the geolocation result is Henan province, China.

Finally, due to the network is not updated timely, users may report the location of the last WIFI connection to the crowdsourcing platform, resulting in incorrect GPS location collection. Thus, we obtain 293,006 IP that appears at least twice and whose GPS location is close to each other.

3.1.1. Application Distribution. Application scenarios are divided by the type of IP user. 1) the enterprise special IP, whose user is small companies, also covers the region where the company is located in; 2) residential IP, whose user is ordinary residential users, usually covers one or more adjacent residential cells;3) the special export IP usually used by major industries is employed by their multiple branches;4) the data center IP, whose users are data center that provides network hosting services or cloud operators, such as the IP used by Ali cloud server; 5) the organization IP, whose users has its own AS by non-operator organizations. Such IP is often used within organizations, such as Microsoft and HP; 6) the university IP, whose users are the University, such as the IP used by Zhengzhou University.

Table 2 shows the distribution of 293,006 IP application scenarios. As can be seen from Table 2, the number of enterprise special IP is 142,782(48.73%), ranking the first. The number of residential IPs is 1,614(48.34%), ranking the second. The number of special export IP, university IP and organization IP is 5,362, 2,565 and 306 respectively, ranking third, fourth and fifth respectively. The number of data center IP is 306(0.10%) with the fewest number.

Table 2. The distribution of application scenario.

| Application scenario | IP number | Percentage |
|----------------------|-----------|------------|
| Enterprise Special   | 142,782   | 48.73%     |
| Residential          | 141,639   | 48.34%     |
| Special Export       | 5,362     | 1.83%      |
| University           | 2,565     | 0.88%      |
| Organization         | 352       | 0.12%      |
| Data center          | 306       | 0.10%      |
| In total             | 293,006   | 100.00%    |
3.1.2. Precision. Table 3 shows the distribution of 293,006 IP geolocation precision. As can be seen from Table 3, the number of street-level IPs is 286,951(97.93%), ranking the first. The number of district-level and city-level IPs is 3,686(1.26%) and 1,124(0.38%) respectively, ranking second and third respectively. The number of province-level IPs is 1,245(0.42%), ranking the last.

Table 3. The distribution of geolocation precision.

| Precision     | IP number | Percentage |
|---------------|-----------|------------|
| Street-level  | 286,951   | 97.93%     |
| District-level| 3,686     | 1.26%      |
| City-level    | 1,124     | 0.38%      |
| Province-level| 1,245     | 0.42%      |
| In total      | 293,006   | 100.00%    |

3.2. Experiment Results and Analysis

There are two groups of experiments for 293,006 IP. In the first experiment, we analysed city consistency of 291,761 IP with city level and above geolocation precision, covering 286,951 street-level IP, 3,686 district-level IP and 1,124 city-level IP.

In the second experiment, we analysed the error distance distribution for 286,951 street-level IPs.

3.2.1. Experiment 1: City-level Consistency. First, we use the API of the Baidu map or Google map to obtain the country, province and city information corresponding to these latitudes and longitudes; second, we compare if the city of between IP actual location and geolocation results is consistent, and judge city-level consistency of the IP geolocation results.

The IP number of the same city between the actual location and geolocation results is 289,724, accounting for 98.88%. And the IP number of the different cities between the actual location and geolocation results is 3,282, accounting for 1.12%.

There are two reasons why cause city-level inconsistency. Specifically: (1) Residential IPs in remote or sparsely populated areas can cover a large area and are distributed in different neighboring cities; (2) For the data center IP, the geolocation result is the real location of the machine room, while the crowdsourcing platform reports the use IP geolocation.

3.2.2. Experiment 2: The Error Distance Distribution of Street-level IP. For 293,006 Street-level IP, we calculate the distance between their real location and the center point of the geolocation result, as shown in Figure 3.
As you can see in Figure 3, the median error distance of the popular IP geolocation system is 0.13 km; the maximum error distance is 66.65 km. The error distances of 70%, 80%, 90% and 95% are 1.39 km, 2.24 km, 3.70 km and 3.70 km respectively.
From the largest number of enterprise dedicated lines’ IP and residential users’ IP, this paper analyses the median error distance of geolocation results in detail. Figure 4 and figure 5 respectively show their error distance distribution diagram.
In view of the largest number of enterprise special and residential IP, we analyse the median error distribution between the real location and geolocation results of them in detail.
Figure 4 shows that the median error distance of street-level enterprise special IP is 0.07 km. The maximum error distance is 13.92 km. The error distances of 70%, 80%, 90% and 95% are 0.09 km, 0.11 km, 0.12 km and 0.14 km respectively.

![Figure 4](image)

**Figure 4.** The error distance distribution of street-level enterprise special IP.

We can see from Figure 5 that the median error distance of street-level residential IP is 1.74 km. The maximum error distance is 66.65 km. The error distances of 70%, 80%, 90% and 95% are respectively 2.74 km, 3.72 km, 7.49 km and 11.76 km.

![Figure 5](image)

**Figure 5.** The error distance distribution of street-level residential IP.
4. Related Work
Chen et al. [10] develop a new IP geolocation algorithm that uses the nearest common router. This algorithm gets the geolocation of the nearest common router by estimating the delay and distance between the router and landmark, and then assign the router’s location to target IP. This algorithm reduces the accumulated relative delay error. Zhao et al. [11] introduce a street-level geolocation algorithm based on the router multilevel partitioning. The algorithm makes a hypothesis that each router has a relatively stable service object for a period of time. Based on this hypothesis, the geographic service of the router can be inferred by analysing the connection between routers and landmarks. Finally, the algorithm can estimate the geolocation of the target IP according to routers’ service ranges. This algorithm reduces 31.36% of error comparing to Wang’s SLG algorithm. Most of the previous researches focused on how to increase the precision and accuracy of their geolocation results, and very few scholars focused on the accuracy evaluation system. We focused on how to increase the target test set in attempt to build the evaluation platform.

5. Conclusion and Future Work
We design and propose the crowdsourcing platform to collect the target IP with location information, and increase the number of target sets in the IP geolocation experiment. Based on the collected IPs, we evaluate the accuracy of a popular IP geolocation system. The results show that: a) the city consistency of the IP geolocation system reaches 98.88%; b) in the case of street-level IP, the median error of the IP geolocation system is 0.13km, achieving high-precision IP geolocation.

How to collect a large number of IPs with location information and how to use the limited IP with location efficiently will be one of the future directions of high-precision IP location system.

References
[1] J. A. Muir and P. C. V. Oorschot, “Internet geolocation: evasion and counter evasion,” in Journal of ACM Computing Surveys (CSUR’09), 42(1):4, 2009.
[2] Z. Wang, H. Li, Q. Li, W. Li, L. Sun, “Towards ip geolocation with intermediate routers based on topology discovery,” in Journal of Cybersecurity, 2(1):1-14, 2019.
[3] “Bing,” Microsoft, https://cn.bing.com/.
[4] O. Abboud, A. Kovacevic, K. Graffi, K. Pussep, R. Steinmetz, “Underlay awareness in P2P systems: Techniques and challenges,” in Proceeding of the 23rd IEEE Int’l Symp. on Parallel and Distributed Processing (IPDPS), 2009.
[5] B. WONG, I. STOYANOV, E. G. SIRER, “Octant: a comprehensive framework for the geolocalization of Internet hosts,” in Proceeding of The 4th Usenix Symposium on Networked Systems Design & Implementation, 2007.
[6] B. Gueye, A. Ziviani, M. Crovella, S. Fdida, “Constraint-based geolocation of internet hosts,” in Journal of ACM IEEE/ACM Transactions on Networking, 14(6): 1219-1232, 2006.
[7] E. KATZ-BASSETT E, J. P. JOHN, and A. KRISHNAMURTHY, “Towards IP geolocation using delay and topology measurements,” in Proceeding of The 6th ACM Sigcomm Conference on Internet Measurement, 2006.
[8] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, C. Huang, “Towards street-level client-independent IP geolocation,” in Proceeding of The 8th Conference on Networked Systems Design and Implementation (NSDI), 2011.
[9] F. ZHAO, Y. SONG, F. LIU, K. Ke, J. Chen, X. Luo, “City-Level Geolocation Based on Routing Feature,” in Proceeding of The 29th International Conference on Advanced Information Networking and Applications, 2015.
[10] J. Chen, F. Liu, Y. Shi, and X. Luo, “Towards IP location estimation using the nearest common router,” in Journal of Internet Technology, 19(6): 2097–2110, 2018
[11] F. Zhao, R. Xu, R. Li, M. Zhu, X. Luo, “Street-Level Geolocation Based on Router Multilevel Partitioning,” in Journal of IEEE Access, 7: 59237–59248, 2019.