Towards IP Location Estimation Using the Nearest Common Router

Jing-ning Chen[1], Fen-lin Liu[1], Ya-feng Shi[2], Xiangyang Luo[1]

[1] State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou Science and Technology Institute, China
[2] School of Mathematics and Statistics, Kashgar University, China
jingning_chen@sina.com, fenlinliu@vip.sina.com, shyf-shyf@163.com, luoxy_ieu@sina.com

Abstract

As for the large error of geolocation algorithm based on shortest relative delay in the real Internet, a new IP geolocation algorithm is proposed in this paper, which could get the geographic location estimation of Internet host using the nearest common router between the target host and landmarks. The proposed algorithm finds the nearest common router between the target host and landmarks through topology discovery and takes the network entity with known location as passive landmark; estimates the delay between nearest common router and landmark, determines the geographic location of this router, then mapping the target host to this location. The algorithm analysis and experimental results show that, compared with the geolocation algorithm based on the shortest relative delay, proposed algorithm could reduce the accumulated error caused by relative delay, as well as the average errors of geolocation results effectively, and then it is expected to improve the accuracy of IP geolocation.

Keywords: IP geolocation, Nearest common router, Landmark, Law of cosines

1 Introduction

Internet geolocation technology is an important network application technology, which aims to determine the geolocation of a target node with IP identifier in a certain level of granularity [1]. Since every host directly connected to the Internet can be identified with a unique IP address, and Internet geolocation technology usually use the IP address to get geographical coordinate of the host, so Internet geolocation is also called IP geolocation [2]. AS the basis of location-based service, IP geolocation Technology has strong theory significance and application value for many applications, such as targeted advertising, the continuity and supervision of cloud service, screening of sensitive network entity, network forensics of illegal acts like Internet fraud and attack, develop the deployment strategy of the network infrastructure and discover fault nodes [3]. Therefore, it is very important to carry out researches on the IP geolocation technology.

There are some public tools available on the Internet can provide geolocation services, such as IP2Location, Maxmind, Quova, Geobytes and so on. These tools usually consist in building a database to keep the mapping between IP blocks and a geographic location based on cooperative or some unique technical [4-6]. Both of [7-8] analyze the accuracy of those database, get the conclusion that geolocation databases can claim country-level accuracy, but certainly not city-level. The existing IP geolocation technology can be divided into two categories, one is collaboration, and the other is non-cooperation. The former obtains the location of target IP from the existing registration information, such as Whois databases [9], DNS database [10-11] and data provided by ISP. Because of the incomplete and outdate registration information, the errors of those methods are usually much large. The latter kind of methods can get the location of the target IP through the delay measurement and topology discover, without the information from Internet service provider. Traditional non-cooperation IP geolocation technologies are mainly based on the network measurement technology. The representative methods includes CBG [12] (Constraint-Based Geolocation), TBG [13] (Topology-Based Geolocation), SLG (Street-Level Geolocation) [14], GeoGet [15] and so on. CBG is first to introduce the network distance constraints and multilateration to IP geolocation. Its core idea is to treat the measured values of the delay between the probe points and the target as a set of constraints, thus the position of the target can be estimated by using the distances between the target and a sufficient number of probe points, and this idea are still useful [16-17]. Inspired by positioning ideas in sensor networks, Katz-Bassett et al. [13] proposed the TBG algorithm, which using the delays between probe points and the target or

*Corresponding Author: Xiangyang Luo; E-mail: luoxy_ieu@sina.com
DOI: 10.3966/160792642018121907013

the intermediate router and the delay among each hop, then calculates the distance constrains between the probe point and target IP and intermediate routers, as well as the constrains between the adjacent intermediate nodes, thus the positioning problem is transformed into the problem of solving a semidefinite programming problem. SLG uses the idea of approaching tier by tier to get the location of the target. After getting a coarse-grained estimation region of the target from first and second tier using the CBG algorithm, a large number of landmarks in the region can be obtained. Finally, the target IP is located to the position of the landmark which has the shortest relative delay to target IP. GeoGet considers the strength of the correlation between delays and the distance is closely related to the level of network connectivity. When the methods of calculating the distance constraints based on CBG and TBG fail to get the target position, the principle that the shortest delay comes from the nearest probe still holds. Therefore closest-shortest still can give geolocation results.

However, despite a decade of development, IP geolocation has made a great progress, there are still some flaws. For those above methods: when the delay-distance correlation is not moderate or strong, and there is no probe around the target IP, the delay from probe to target IP will much large, and then it will be difficult for CBG and SLG (tier 1) to get effective geolocation result due to large distance constraints. Combined with hop latency and hints of the intermediate routers, TBG relies on global optimization to estimate the intermediate routers and target at the same time. While reduce the geolocation error of intermediate routers, it may introduce more error for target IP. Geoget geolocalize the target IP by taking the Web servers as passive landmarks, and it can avoid the situation under which delay may not be gotten by active measurement, and the consequence is that only those hosts which connect to the master server actively can be geolocalized. In addition, the IP geolocation methods based on machine learning and statistics are proposed in [18-21], but the realizations of this kind of methods are usually complex, and the accuracy much depends on the training data. While there are also some lightweight geolocation methods to reduce the burden of delay measurement [22-24], they may lack applicability. The existing IP geolocation studies show that, although there are some methods can geolocalize internet hosts, due to the inaccurate delay measurement, incomplete path information and the number of landmarks, the geolocation errors of those methods are generally large and cannot meet the needs of high-precision geolocation.

In this paper, we investigate the delay-distance relationship in the weakly connected Internet region (Zhengzhou City, capital of Henan Province, China), combined with topology information, a geolocation algorithm using the nearest common router is proposed, and this algorithm takes geolocation result of the router which is the most similar to the target in the aspect of network topology as location estimation for the target IP. Different from the common IP geolocation methods based on the delay-distance relationship between the probe and target or the relative delay-distance between the landmarks and target, this paper investigates the delay-distance relationship between the routers close to target IP and landmarks, and then proposed algorithm can improve the geolocation accuracy of the target IP whose nearest common router can be determined by the landmarks related to this router and the delay between them. The algorithm analysis and experimental results show that, the proposed algorithm could work with a single probe, and the average error and the maximum error is better than the algorithm based on the shortest relative delay in [14].

The rest of this paper is organized as follows: Section 2 analyzes the shortness of the existing geolocation method based on the shortest relative delay in the weakly connected Internet region, and proposes the problem to be solved in this paper. Section 3 details the basic idea, principles framework and main steps of the proposed algorithm. Section 4 discusses the error analysis of the proposed algorithm. Section 5 shows the experimental results and Section 6 concludes.

2 Problem Formulation

IP Geolocation scenario is usually can be shown as Figure 1. At first, the typical IP Geolocation methods calculate the delay-distance conversion coefficient for the probe, or use the 2/3C (C is the speed of light) as a conversion coefficient directly. Then according to delay (from probe to target IP) and conversion coefficient of the probe, calculate the distance constraints from probes to the target IP. Finally, get the geolocation results using multilateration. SLG which is one of the IP geolocation methods with smallest geolocation error, introduces the concept of relative delay, and utilizes the idea of approaching tier by tier to geolocalize the target. IP. In tier 1, SLG uses 4/9C (as conversion coefficient) and CBG, and geolocalizes target IP into a coarse-grained region. In tier 2, SLG estimates the relative delay between landmarks (in the above coarse-grained region) and target, takes those landmarks as probes, and then geolocalizes target IP into a fine-grained region like tier 1. In tier 3, SLG estimates the relative delay between landmarks (in the above fine-grained region) and target, and takes the location of landmark which has shortest relative delay to the target IP as the geolocation result. The above procedure shows that, while put SLG into the real IP geolocation applications, requires two preconditions: (1) delay and distance are strongly correlated; (2) relative delay and distance are strongly correlated and the shortest relative delay corresponding to the shortest distance. SLG could achieve very good results in

strongly connected Internet region, while the Internet connection is weak, whether the above two preconditions is also satisfied or not? The following will answer this question, while take the PlanetLab nodes and Zhengzhou Internet hosts as example.

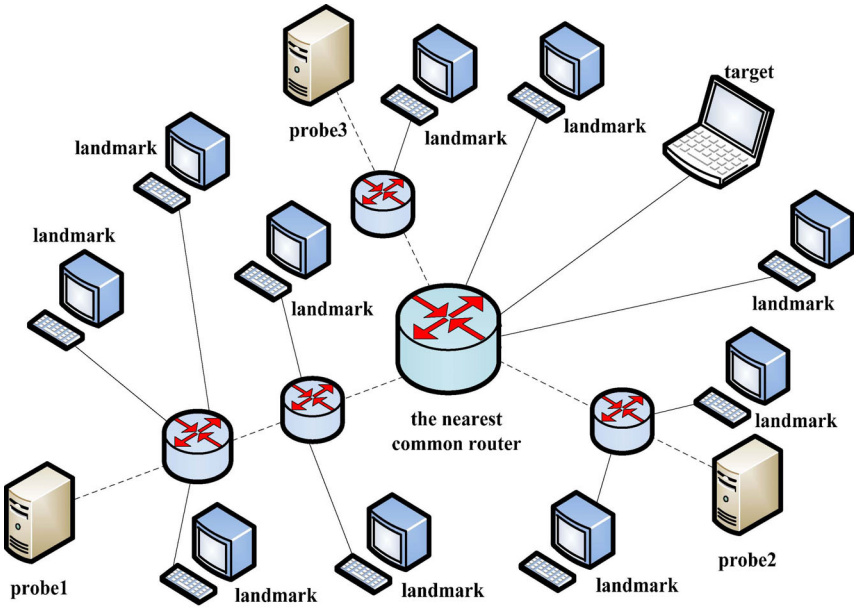

Figure 1. IP Geolocation scenario

While analyzing the delay-distance correlation: For PlanetLab nodes, using planetlab1.uta.edu to as probe and the other 208 nodes as landmarks (the distribution is shown in Figure 2), those landmarks are selected because the delay from planetlab1.uta.edu to the 208 nodes can be measured within ten days at 2014.01.08 ~ 2014.01.17 and the delay value could be measured (between 0ms to 100ms); For Zhengzhou Internet hosts, using the host located in (34.816129N, 113.535455E) (N is north latitude, E is east longitude) as probe and 417 Internet hosts as landmarks (the distribution is shown in Figure 5, subject to Google Map free API, here only show 300 landmarks, the rest 117 landmarks are all nearby, the detailed distribution as shown in Figure 6). While analyzing the relative delay-distance correlation: For Zhengzhou Internet hosts, the probe is same as above and the landmarks are 176 hosts (distribution as shown in Figure 8) of above 417 landmarks, delay and path of each of 176 hosts could be measured.

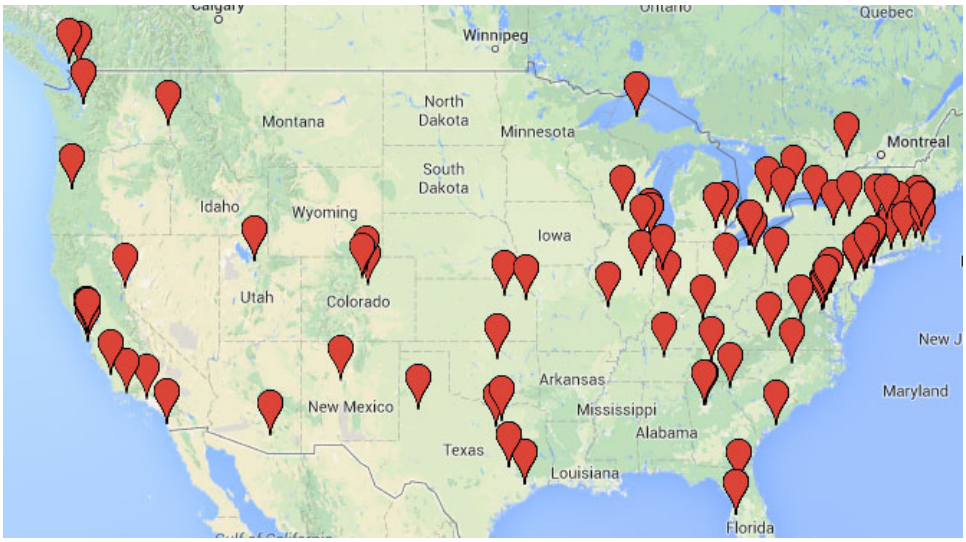

Figure 2. The distribution of 208 Planetlab node

2.1 Delay-Distance Correlation

For PlanetLab nodes, we measure the delay from planetlab1.uta.edu to other 208 nodes of Planetlab, and calculate the corresponding geographical distance. All the (delay, distance) pairs are shown in Figure 3. For 417 Zhengzhou landmarks, we measure the delay from probe to landmarks and calculate the corresponding geographic distance. All the (delay, distance) pairs are

shown in Figure 7. The correlation coefficient of delay and distance $corr(de, di)$ could be calculated using formula (1) in [15], and the $corr(de, di)$ also could be used to measure Internet connectivity. The formula (1) is shown as follows:

$$corr(de, di) = \frac{cov(de, di)}{sqrt(D(de)) \times sqrt(D(di))} \quad \textbf{(1)}$$

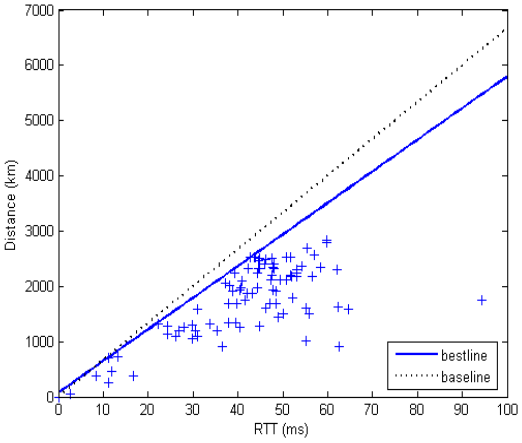

Figure 3. Bestline and baseline of CBG

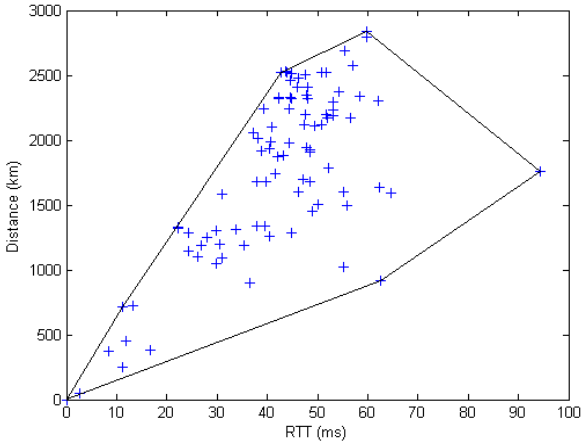

Figure 4. Negative and positive constraint of Octant

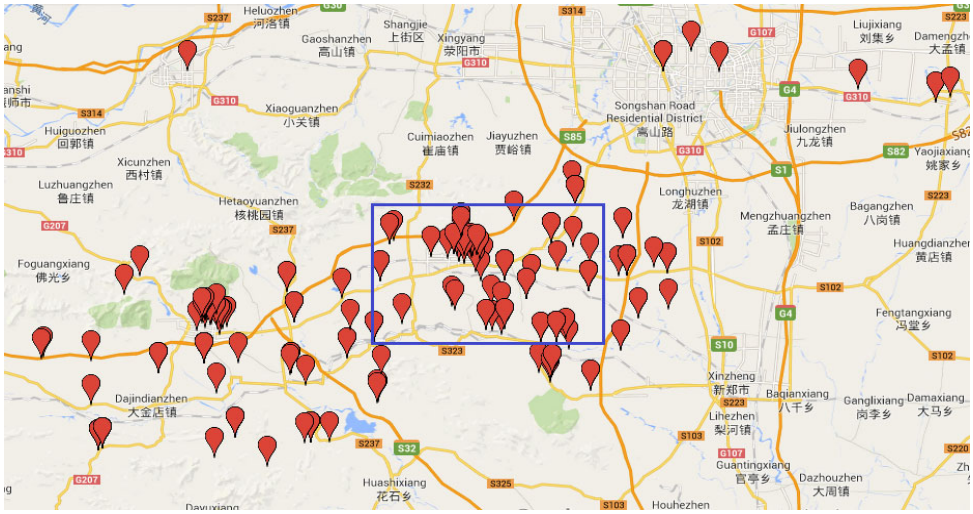

Figure 5. Distribution of 300 Zhengzhou landmarks

Figure 6. Detailed distribution of landmarks in the blue box of Figure 5

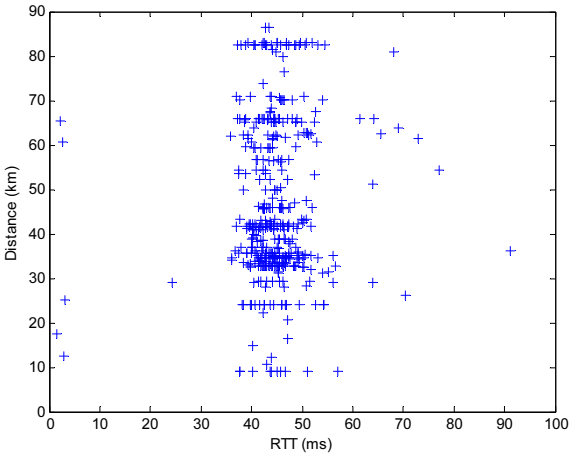

Figure 7. (delay, distance) pairs of 417zhengzhou landmarks

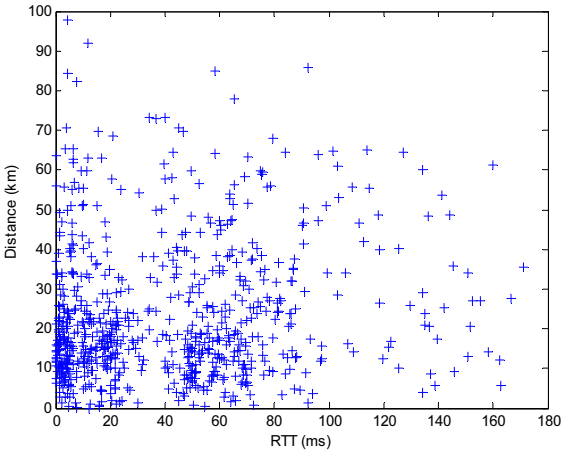

Figure 8. (relative delay, distance) pairs of 176 zhengzhou landmarks

In formula (1), the delay between Internet nodes is de, the distance between Internet nodes is di, the deviation of delay is $D(de)$, the deviation of distance is $D(di)$, and the covariance between delay and distance is $cov(de, di)$.

Then calculate the correlation between delay and distance of Planetlab nodes and Zhengzhou landmarks through the formula (1). For PlanetLab nodes, the correlation coefficient is 0.7439 and the bestline $(y = 57.2055x + 71.8018)$ of CBG is shown in Figure 3, and the baseline is corresponding to the delay-distance relationship while the conversion coefficient is 4/9C. Straightforward negative and positive constraint of Octant [25] is shown as Figure 4. For Zhengzhou landmarks, the correlation between delay and distance is very weak, so it is difficult for CBG, TBG, SLG and Octant to get an effective constraint from probe to target IP.

2.2 Relative Delay-Distance Correlation

Wang et al. [14] introduced the concept of relative delay. For two landmark nodes A and B, the nearest common router between A and B is R, while the round-trip delay from probe P to A, B, and R is $RTT(P, A)$, $RTT(P, B)$ and $RTT(P, R)$, the relative delay between A and B as shown in formula (2):

$$RltRTT(A, B) = (RTT(P, A) - RTT(P, R)) + (RTT(P, B) - RTT(P, R)) \quad (2)$$

After knowing the relative delay between landmarks and the targets, Wang et al. [14] uses 4/9C as conversion coefficient and computes distance constraints from landmarks to target. For 176 landmarks of Zhengzhou (distribution as shown in Figure 9), all the (relative delay, distance) pairs are shown in Figure 8, and then it is obvious that the relative delay and distance correlation is very weak, and the 4/9C is too loose to geolocalize.

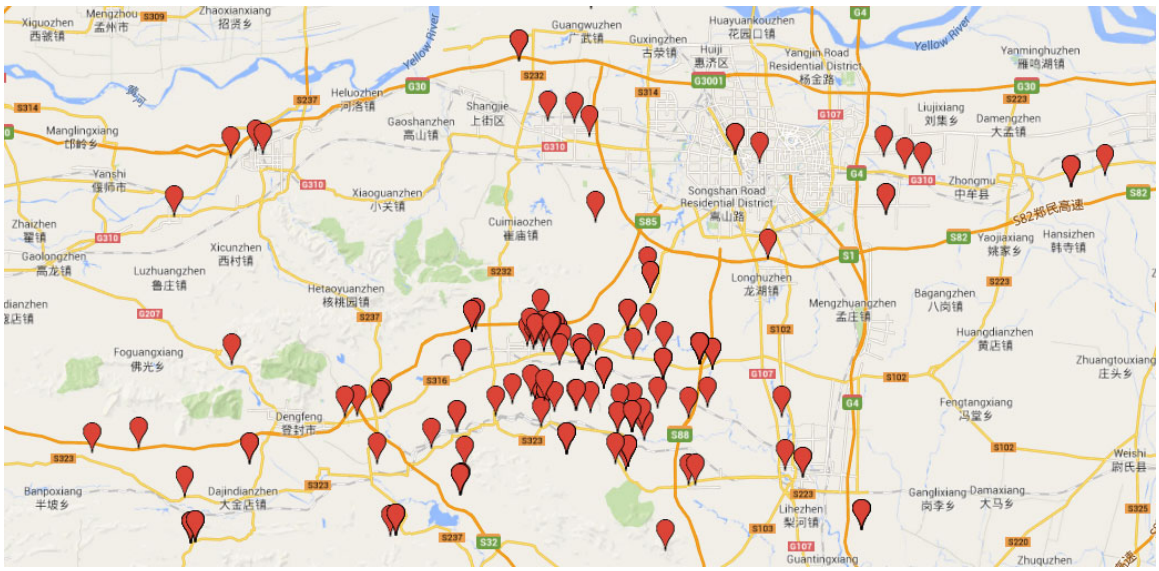

Figure 9. Ddistribution of 176 Zhengzhou landmarks

When geolocalizing the target into a fine-grain region, [14] selects the landmark with shortest relative delay to the target IP, and argues that the target's location is same as this landmark's. In order to analyze the correspondence between the relative delay and shortest distance, this paper uses the method of calculating Disk-rank-of-shortest-delay in [15] to calculate the Disk-rank-of-shortest-relative-delay. For a landmark entity A, find out all (relative delay, distance) pairs related to A and sort those pairs according to distance from smallest to largest, then choose the ranking position (starting from 0) of the pair with the shortest delay, say r. Finally, r divided by the total number of (relative delay, distance) pairs related to A, say Dist-rank-of-shortest-delay. The corresponding CDF of Dist-rank-of- shortest-delay is shown in Figure 10. When the Dist-rank-of-shortest-delay. Of the landmark A is 0, it means that the landmark which have the shortest relative delay with A is just the one which is closest to A. As can be seen from Figure 10, for the geolocation algorithm [14] based on shortest relative delay, there is only less than 30% of the target could be geolocalized to its nearest landmark.

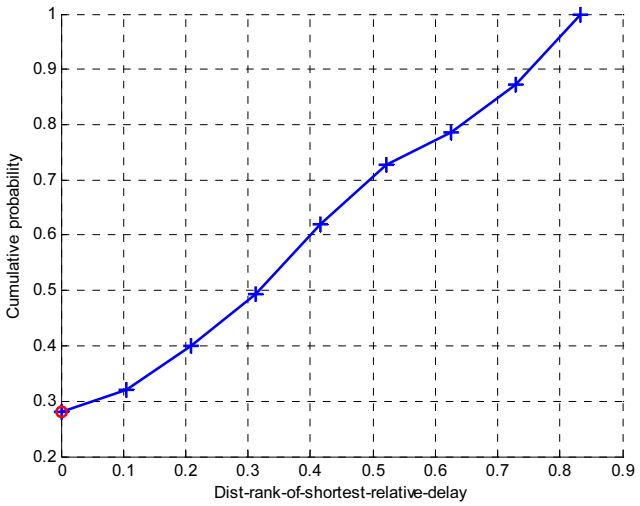

Figure.10. CDF of Dist-rank-of-shortest-delay for Zhengzhou landmarks

As known from the above experiments, the correlation between delay and distance of PlanetLab nodes is rather strong, and use 2/3C, 4/9C or bestline (such as CBG), the geolocation algorithms based on delay measurement can obtain effective constraints for target and get the geolocation result. In the Internet region we study: The correlation between delay and distance is very weak, as well as the relative delay and distance. So it is difficult to construct distance constraints for target IP from the probes and landmarks; As the formula (2) shown, geolocation algorithms based on the relative delay will introduce two delay error: *RTT(P, A) - RTT(P, R)* and *(RTT(P, B) - RTT(P, R)* ; In many cases, geolocation algorithms based on the shortest relative delay cannot get the nearest landmark for target IP.

In order to overcome the above problems and improve the accuracy of IP geolocation in the weakly connected real Internet region, an IP location estimation algorithm using the nearest common router is proposed in this paper and assigns geographic location of the nearest common router to the target IP. Calculate a conversion coefficient for each nearest common router, the proposed algorithms will expect to overcome the larger geolocation error caused by the fixed conversion coefficient to some extent; Calculate the distance constraint from the delay between the nearest common router and landmarks, the proposed algorithms will eliminate the accumulated error caused by relative delay; Geolocalize the nearest common router by utilizing the location of landmarks directly connected to this router, the proposed algorithms could avoid the error caused by the existing geolocation algorithms based on the shortest relative delay, and then improve the geolocation accuracy of target IP.

3 Location Estimation Algorithm Using the Nearest Common Router

In the real Internet environment, the locations of target entities are usually close to the last-hop router[14, 23]. Thus, the last-hop router on the path from probe to target IP is usually the center of distribution of all locations where the target IP maybe located in. Therefore, it is a reasonable way mapping the target IP to this location of nearest common router between the target host and landmarks, while the expected error is minimized. Based on this idea, this paper presents an IP location estimation algorithm using the nearest common router.

3.1 Schematic Diagram

The schematic diagram of the proposed algorithm is shown in Figure 11, and our work is the blue dashed box part of the diagram. The basic idea of IP location estimation algorithm based on the nearest common router is that: when landmarks connected to the nearest common router not less than 3, combined with the delay between the nearest common router and landmark, the geographical location of that router could be determined, and this location can be used as the location estimation of target IP. The key steps are as follows: using delay and path probe packets, get the delay and path from probe to target IP and landmark; according to path from probe to target IP and landmarks, find the landmarks which have the nearest common router with target IP and this router; according to delay from probe to this router and delay from probe to those landmarks, calculate the delay between the router and landmarks; according to the delay and the location of the landmarks, calculate the conversion coefficient between this router and landmarks using the law of cosines; according to the

conversion coefficient and delay, calculate the distance between this router and landmarks; get the geolocation result of the router using multilateration and take the location of this router as the location estimation of the target IP.

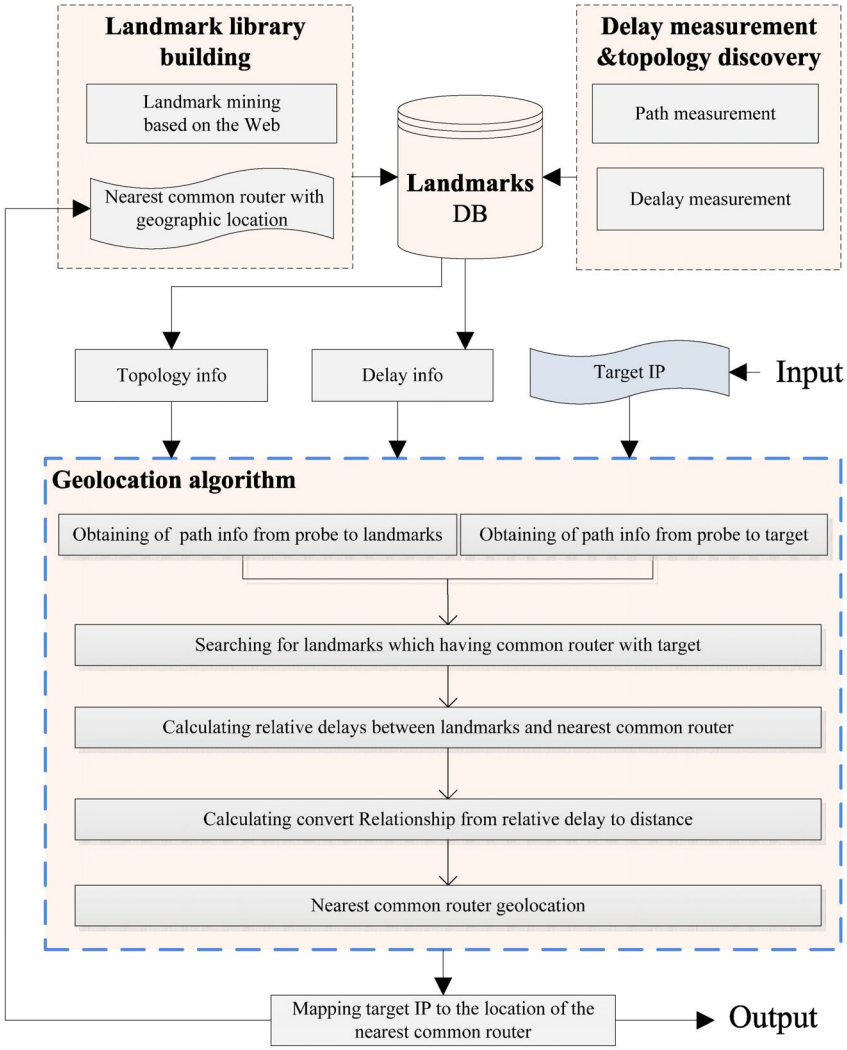

Figure 11. The schematic diagram of proposed algorithm

Algorithm: NCRGeolocation()

Input: landmark DB_L and target IP

Output: (R_{lat}, R_{lng}), the location of target IP

1. $Traceroute(IP, DB_L) \rightarrow Path_T, Path_L$
2. $Relevance(Path_L, Path_T) \rightarrow R, L_1, ..., L_n$
3. $Compute_1(R, L_1, ..., L_n) \rightarrow t_{R1}, ..., t_{Rn}$
4. $Choose(L_1, ..., L_n) \rightarrow L_i, L_j, L_k$
5. $Compute_2(L_i, L_j, L_k, t_{Ri}, t_{Rk}, t_{Rn}) \rightarrow \delta$
6. $Multilaterate(L_i, L_j, L_k, t_{Ri}, t_{Rk}, t_{Rn}, \delta) \rightarrow (R_{lat}, R_{lng})$

Among the above: $Traceroute(IP, DB_L)$ is used to collection the path $Path_T$ from probe to the target IP and the paths $Path_L$ from probe to the landmarks in DB_L; $Relevance(Path_L, Path_T)$ is used to find the nearest common router R between landmarks and target IP, record R and the corresponding landmarks $L_1, ..., L_n$; $Compute_1(R, L_1, ..., L_n)$ is used to calculate the delay $t_{R1}, ..., t_{Rn}$ between R and $L_1, ..., L_n$; $Choose(L_1, ..., L_n)$ is used to select three available landmarks L_i, L_j, L_k; $Compute_2(L_i, L_j, L_k, t_{Ri}, t_{Rk}, t_{Rn})$ is used to calculate the conversion coefficient δ of delay and distance using the law of cosines, according to L_i, L_j, L_k and t_{Ri}, t_{Rk}, t_{Rn}; $Multilaterate(L_i, L_j, L_k, t_{Ri}, t_{Rk}, t_{Rn}, \delta)$ is used to get the geolocation result of R using multilateration and assign this location to the location estimation of the target IP.

From the above schematic diagram and key processes, we can know that the key steps of the proposed algorithm includes: the path detection,

looking for the nearest common router, delay measurement, calculating the conversion coefficient of delay and distance and mapping the nearest common router, etc. Among them, the path detection means that probes send traceroute packages to landmarks and target IP, obtain the intermediate routers interfaces from probes to destination hosts and then identify and merge multiple interfaces of the same router; looking for the nearest common router means that finding the common routers which are both on the path from the probe to landmarks and target IP, and the closest one to target from a topological point of view; different from the traceroute packages of the path detection, delay measurement sends ping packages to obtain the RTTs between probes and destination hosts, and because of the network congestion and circuitous paths, accurate delay usually cannot be gotten through once measurement, and this usually needs multiple measurements and take the minimum RTT as the final delay; The candidate landmark selection refers to, when there are multiple landmarks directly connected to the nearest common router, choose 3 landmarks (calculate the relative delay from those landmarks to nearest common router, and then take out the three landmarks with smallest delay) to geolocalize the nearest common router.

The two most important steps of the algorithm is that: (1) calculating the conversion coefficient of delay and distance; (2) mapping the nearest common router. The details are as follows.

3.2 Calculating Conversion Coefficient of Delay and Distance

Due to the locations of landmarks are known, after the nearest common router between landmarks and target IP and the delay between this router and landmarks are obtained, the conversion coefficient of delay and distance while the packet forwarded by this router could be calculated using the law of cosine.

For the distribution scenarios of Internet entities shown in Figure 12, there are 5 landmarks which have nearest common router with the target IP. 3 landmarks are selected from the 5 landmarks to calculate the conversion coefficient of delay and distance. Target IP entity referred to as T. The nearest common router referred to as R. The selected three landmark labeled as A, B and C. The distance between A and B referred to as d_1, the distance between A and C referred to as d_2, the distance between A and C referred to as d_3. The delay between R and A, B, C referred to as t_1, t_2, t_3, respectively. The conversion coefficient of delay and distance referred to as δ, and then the distance between R and A, B, C is δt_1, δt_2, δt_3, respectively. The abstracted geolocation model corresponds to Figure 12 is shown in Figure 13.

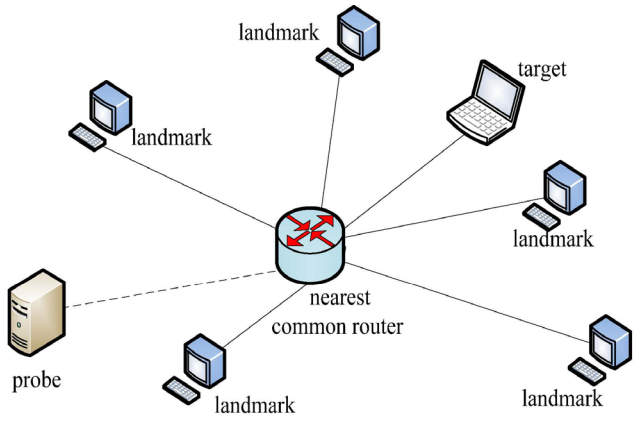

Figure 12. The distribution scenarios of Internet entities

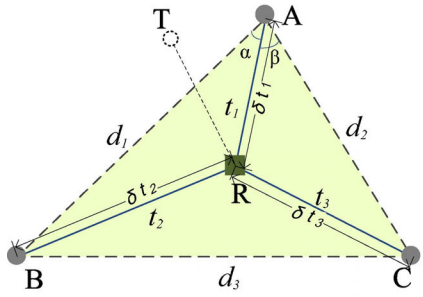

Figure 13. Abstracted geolocation model

In Figure 13, $\angle BAR=\alpha$, $\angle RAC=\beta$. Since the locations of A, B and C are known, the triangle $\triangle ABC$ is uniquely determined and $\angle BAC$ is a constant, known from the law of cosines:

$$\alpha=arccos(((d_1^2+(\delta t_1)^2-(\delta t_2)^2)/(2\delta d_1 t_1)) \quad (3)$$

$$\beta=arccos(((\delta t1)^2+d_2^2-(\delta t_3)^2)/(2\delta d_2 t_1)) \quad (4)$$

$$\angle BAC=arccos((d_1^2+d_2^2-d_3^2)/(2d_1 d_2)) \quad (5)$$

Take formula (3-5) into the expression: $\angle BAC = \angle BAR+\angle RAC = \alpha+\beta$, equation (6) could be obtained:

$$\begin{aligned} &arccos((d_1^2+d_2^2-d_3^2)/(2d_1d_2))= \\ &\quad arccos((d_1^2+(\delta t_1))^2-(\delta t_2))^2)/(2\delta d_1 t_1))) \\ &\quad +arccos(((\delta t_1)^2+d_2^2-(\delta t_3))^2)/(2\delta d_2 t_1))) \end{aligned} \quad (6)$$

In the equation (6), δ is the only one unknown variables, thus solving this equation, the solution of the equation is the conversion coefficient of delay and distance between the nearest common router and landmarks.

When there are more than three landmarks directly connected to the nearest common router, different landmarks selection strategies will result in different conversion coefficients. Based on the assumption that the error of the small delay is relatively small, three

landmarks while the delay from which to the nearest common router is smallest, will be selected.

3.3 Mapping the Nearest Common Router

After get the conversion coefficient of delay and distance, three equations (7-9) can be created:

$$distance\left(lat_1,\ lng_1,\ x,\ y\right)=r_1 \quad (7)$$

$$distance\left(lat_2,\ lng_2,\ x,\ y\right)=r_2 \quad (8)$$

$$distance\left(lat_3,\ lng_3,\ x,\ y\right)=r_3 \quad (9)$$

The (latitude, longitude) of landmark A is $\left(lat_1, lng_1\right)$, the (latitude, longitude) of landmark B is $\left(lat_2, lng_2\right)$, the (latitude, longitude) of landmark C is $\left(lat_3, lng_3\right)$. The distance from R to A is r_1 ($r_1=\delta t_1$), the distance from R to B is r_2 ($r_2=\delta t_2$), the distance from R to C is r_3 ($r_3=\delta t_3$). $distance\left(lat_i, lng_i, lat_j, lng_j\right)$ is the function that calculate the geographical distances between two points (lat_i, lng_i) and (lat_j, lng_j) using Vincenty's formula [26]. The solution of equations (7-9) is the latitude and longitude of the nearest common router.

4 Error Analysis of Proposed Algorithm

4.1 Analysis of Accumulated Error

Wang et al. [14] argued that, when using a sufficient numbers of traceroute servers, the path between landmark and target IP connected through nearest common routers can represent the direct path between them, and the relative delay can be taken as the delay between the landmark and target IP (equivalent to use the landmark as the probe, directly measure the delay from the landmark to target IP). Thus based on shortest ping [13] (measuring delay from many probes to the destination IP, and select the probe with shortest delay as the geolocation result), target IP can be mapped to the location of the landmark with the shortest relative delay to target IP.

From the above analysis, as shown in Figure 13, the delay t_i' ($i=1,2,3$) adopted in the geolocation algorithm based on relative delay, and $t_i'=t_i+t$, while t_i ($i=1,2,3$) is the delay between landmarks and the nearest common router and t is the delay between the target IP and the nearest common router. In the real Internet, because of the impact of network congestion, circuitous paths and other factors, there is always an error in the estimation of t_i and t. Thus the geolocation algorithm based on relative delay has more accumulated error due to the sum of t_i and t, and in the result of the large geolocation error. The IP location estimation algorithm proposed in this paper based on the nearest common router between target IP and landmarks, uses the delay t_i between the nearest common router and landmarks (directly connected to target IP through the nearest common router), only introduces once error of delay, and then this algorithm is expected to improve the geolocation accuracy.

4.2 Maximum and Average Error Analysis

In the real network environment, target IP and landmarks with the nearest common router to the target IP are usually distributed around the nearest common router, and the probability of every Internet host locates in each location could be seemed as the same. It means that, there exists a circle that uses the location of the nearest common router as the center, the distance between nearest common router and the landmark which is farthest to the nearest common router as the radius, and target IP and landmarks are uniformly distributed within this circle. While O is center of the circle, r is the radius and ε is the distance between the nearest common router and target IP, then generally $\varepsilon \le r$.

Comparison of the maximum errors: The geolocation algorithm based on shortest relative delay does not consider the angle determined by landmark, target IP and the near common routers. For the distribution scenarios is as shown in Figure.14. When the relative delay between T and L is smaller than the relative delay between T and the other four landmarks, the geolocation algorithm based on shortest relative delay will map T to the location of L. However, among the five landmarks, L is the farthest one to T in fact, and in this case, the geolocation result corresponding to the maximum error $\varepsilon+r$. At the same time, the IP location estimation algorithm using the nearest common router proposed in this paper selects three landmarks in the Figure 13, and gets the geographic location of R according to the delay between the three landmarks and R. Finally, take this location as the location estimation of T and the geolocation error is ε. It is obvious $\varepsilon < \varepsilon + r$, and then the maximum error of IP location estimation algorithm proposed in this paper is smaller than geolocation algorithm based on the shortest relative delay.

Comparison of the average errors: As the case shown in Figure 15, the geolocation algorithm based on shortest relative delay will map the T to one of landmarks within the circle. Suppose there're n landmarks within the circle, denoted by $L_i, ..., L_n$, the distance between L_i and T is $de\left(L_i, T\right)$. Because the distribution of landmarks within the circle can be regarded as uniform distribution, then $de\left(Li, T\right) \in \left[0, \varepsilon+r\right]$, and $de\left(Li, T\right)$ could be a set of uniform

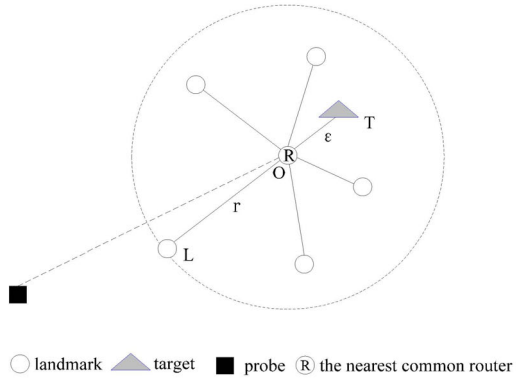

Figure 14. Maximum error comparison of two algorithms

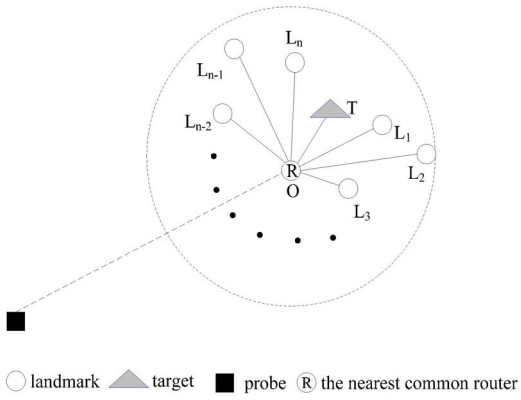

Figure 15. Average error comparison between two algorithms

values within $[0,\varepsilon+r]$, while each probability is $1/n$. So the average error d_{mean} of the geolocation algorithm based on shortest relative delay is shown as formula (10).

$$\begin{aligned} d_{mean} &= E\left(de(L_i,T)\right) = \frac{1}{n}\sum_{i=1}^{n} de(L_i,T) = \frac{1}{n}\sum_{i=1}^{n}\frac{i(\varepsilon+r)}{n} \\ &= \frac{1}{n}\left(\frac{\varepsilon+r}{n}+\frac{2(\varepsilon+r)}{n}+\ldots+(\varepsilon+r)\right) \\ &= \frac{1}{n}\left(\frac{\varepsilon+r}{n}(1+2+\ldots+n)\right) \\ &= \frac{(\varepsilon+r)(n+1)}{2n} \end{aligned} \tag{10}$$

Different from the geolocation algorithm based on shortest relative delay, the IP location estimation algorithm proposed in this paper calculates the conversion coefficient of delay and distance for R. When the number of landmarks connected to R is more than 3, the geographic location of R can be gotten using multilateration, and then maps T to this location. So the geolocation error is ε.

Usually $\varepsilon \le r$, thus:

$$\varepsilon \le \frac{\varepsilon+r}{2} < \frac{(\varepsilon+r)(n+1)}{2n} = d_{mean} \tag{11}$$

So, the average error of IP location estimation algorithm proposed in this paper is also smaller than geolocation algorithm based on the shortest relative delay.

5 Experimental Results

In terms of experimental implementation, the database which composes with IP addresses located in Zhengzhou and Shanghai is constructed, while the locations of those IP are known. In order to verify the effectiveness of the proposed algorithm, this paper carries out experiments in both cases: the probe and target IP locate in the same city and the probe and target IP locate in different cities. Generally, the paths from different probes to the same target IP are similar at the last two or three hops, so in most cases, one probe is enough for geolocation algorithm based on the shortest relative delay in [14] and IP location estimation algorithm proposed in this paper. The single probe used in experiment locates at (34.816129N, 113.535455E) (N is northern latitude, E is east longitude).

5.1 Probe and Target IP are Located in the Same City

In this experiment, the data set is composed of 176 landmarks (distribution as shown in Figure 8), and all these landmarks are located in Zhengzhou city and its affiliated county (city).

Before calculating the conversion coefficient of delay and distance between nearest common routers and landmarks, it is need to find out the nearest common router between target IP and landmarks. For instance, for target IP 120.194.19.227 and 120.194.19.229, their paths are shown in Table 1. Combined with the paths from probe to those landmarks, 120.194.30.42 is selected as the nearest common router. There are 12 landmarks connected to 120.194.30.42, as shown in Table 2.

Among 4 solutions of equation (3), only positive solutions are chosen as δ, because δ refers to a conversion coefficient between delay and distance. Geolocation error of proposed algorithm of this paper shows that the smaller the value of δ, the smaller the gelocation error. Therefore, when the equation (3) has multiple solutions, the smallest positive solution is selected as the value of δ. From Table 2, 6 landmarks (2 groups) connected to 120.194.30.42 that have smaller delay are selected to map the nearest common router.

Table 1. The paths from probe to 120.194.19.227and 120.194.19.229

Probe_IP	Router_IP	Target_IP	RouterHop	Probintime
10.104.171.78	218.29.102.1	120.194.19.227	3	2014/4/18 8:20
	61.168.251.69		4	2014/4/18 8:20
	61.168.32.125		5	2014/4/18 8:20
	219.158.16.89		6	2014/4/18 8:20
	219.158.11.114		7	2014/4/18 8:20
	219.158.38.214		8	2014/4/18 8:20
	221.176.15.85		9	2014/4/18 8:20
	221.183.8.109		10	2014/4/18 8:20
	221.183.12.18		11	2014/4/18 8:20
	221.176.98.6		12	2014/4/18 8:20
	120.194.30.42		13	2014/4/18 8:20
	120.194.19.227		14	2014/4/18 8:20
	218.29.102.1	120.194.19.229	3	2014/4/18 8:20
	61.168.18.217		4	2014/4/18 8:20
	61.168.195.49		5	2014/4/18 8:20
	219.158.21.117		6	2014/4/18 8:20
	219.158.11.54		7	2014/4/18 8:20
	219.158.38.214		8	2014/4/18 8:20
	221.176.16.33		9	2014/4/18 8:20
	221.183.8.109		10	2014/4/18 8:20
	221.183.12.22		11	2014/4/18 8:20
	221.176.99.6		12	2014/4/18 8:20
	120.194.30.42		13	2014/4/18 8:20
	120.194.19.229		14	2014/4/18 8:20

Table 2. The nearest common router and related landmarks of 120.194.19.227 and 120.194.19.229

Target_IP	Nnearest common router_IP	Landmark_IP	Latitude	Longitude
120.194.19.227 120.194.19.229	120.194.30.42	120.194.19.251	34.31920442	112.90632
		120.194.21.99	34.46173592	113.12939
		120.194.21.109	34.41836217	113.41749
		120.194.21.110	34.46383206	113.50942
		120.194.24.113	34.44353613	113.26704
		120.194.24.137	34.33220076	113.82434
		120.194.24.142	34.33220076	113.82434
		120.194.24.153	34.37341641	113.27229
		120.194.24.160	34.41911437	112.76502
		120.194.24.168	34.31920442	112.90632
		120.194.24.174	34.40685687	113.15662
		120.194.24.182	34.37146776	113.27132

According to the above 6 landmarks, solve equation (3) to get two group solutions for the conversion coefficient of delay and distance δ. The two group solutions are [30.2704, -30.2704, 21.2643, -21.2643] and [7.4086, -7.4086, 2.5137, -2.5137], respectively. Since the conversion coefficient cannot be a negative, choose the positive solution. Then take four positive solutions [30.2704, 21.2643, 7.4086, 2.5137] into equations (4-6), the corresponding geographic locations of router 120.194.30.42 are (34.4394, 112.8683), (34.3076, 112.6946), (34.255, 113.274) and (34.3885, 113.2277), and map the two target IP 120.194.19.227 and 120.194.19.229 to the above locations. The error comparison between geolocation algorithm based on the shortest relative delay in [14] and proposed algorithm are as shown in Table 3 (the unit of geolocation error is kilometer).

Table 3. The error of 120.194.19.227 and 120.194.19.229 for two algorithms

Target IP	shortest relative delay	δ=30.270	δ=21.2643	δ=7.4086	δ=2.5137
120.194.19.227	35.92	54.73	36.61	21.01	7.11
120.194.19.229	28.58	46.38	27.07	25.69	10.58

For the nearest common router 120.194.30.46 and 171.8.240.146, there are 34 landmarks which take those two routers as the last hop (remove the 6 landmarks that chosen to map the nearest common router from 40). Figure 16 shows the cumulative probability comparison between geolocation algorithm based on the shortest relative delay and IP location estimation algorithm based on the nearest common router, while take the 34 as target IP. The average error and maximum error of IP location estimation algorithm using the nearest common router are 34.8482km and 68.3527km, respectively, while the geolocation algorithm based on the shortest relative delay in [14] are 51.0131km and 85.7670km. It can be seen that both of the average and maximum error of the proposed algorithm are smaller than the algorithm based on shortest relative delay, which is consistent with analysis in the Section 4.2. In addition, the red dotted line in Figure 16 shows that the median errors of the two algorithms are 34.8482km and 51.0131km. So the median error of the proposed algorithm is smaller than the algorithm in [14] too.

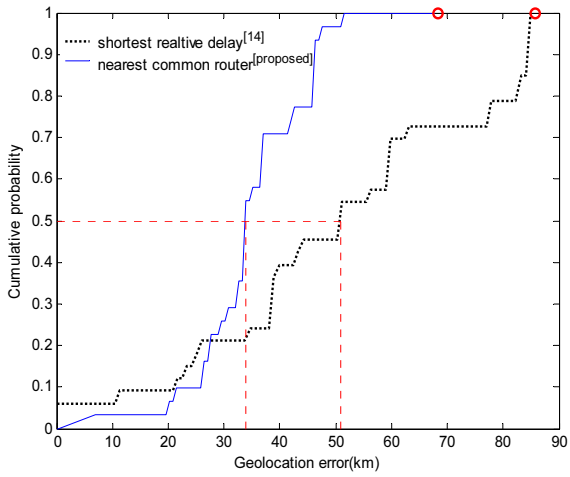

Figure 16. Error distance comparison of the two algorithms.

For Figure 16, while the "shortest relative delay" is the geolocation algorithm based on shortest relative delay in [14], and the "nearest common router" is the algorithm proposed in this paper.

5.2 Probe and Target IP are Located in Different City

In this experiment, the data set is composed of 7 landmarks (distributed in Shanghai city, China), and all of them are given in the Table 4.

Table 4. Seven landmarks of Shanghai

IP	Latitude	Longitude
218.78.244.151	31.279583	121.557675
218.78.244.158	31.277654	121.511699
218.78.245.138	31.218469	121.50256
218.78.246.47	31.311994	121.547414
218.78.245.126	31.273125	121.552464
218.78.246.131	31.248195	121.467288
218.78.246.188	31.275733	121.538078

Take the top four of Table 4 as target IPs, and the last three as landmarks used to calculate the conversion coefficient and geolocalize the nearest common router 218.78.244.253. The 4 solutions of equation (3) are [1.3921, -1.3921, 1.2856, -1.2856] respectively,

take four positive solutions [1.3921, 1.2856] into equations (4-6), and the corresponding geographic locations of the nearest common router are (31.2844, 121.53225) and (31.2661, 121.53165), then maps the four target IPs to this two location. The geolocation error of proposed algorithm is as shown in Table 5 (the unit of geolocation error is kilometer).

Table 5. Geolocation error of proposed algorithm for Shanghai targets

Target IP	δ=1.3921	δ=1.2856
218.78.244.151	2.48	2.90
218.78.244.158	2.09	2.29
218.78.245.138	7.84	5.98
218.78.246.47	3.40	5.32

The geolocation error of Table 5 shows that, when the probe and target IP are located in different cities, the proposed algorithm is still able to achieve the location estimation for target IP, and the smallest error could be reached 2.09km. Because at present, there are only have a limited number of landmarks in cities apart from Zhengzhou, the cumulative error is not analyzed here. In addition, it is worth analyzing the statistical results of geolocation error when the probe and target IP are located in different cities.

6 Conclusion

In this paper, we investigate the delay-distance relationship and the relative delay-distance in a smaller region of China, and find that in the weakly connected real Internet region: The correlation of delay and distance is very weak, as well as the relative delay and distance; There are accumulated error in estimation of relative delay; In this case, the shortest relative delay is

not coming from the shortest distance. In consideration of the real network, the target IP entities are usually distributed around the last hops router, and generally this router is the possible center of the target IP's location, an IP location estimation algorithm using the nearest common router is proposed in this paper. This algorithm calculates the conversion coefficient of delay and distance for each nearest common router using the law of cosines, obtains the geographic location of the nearest common router based on the delay (between landmarks and this router) and multilateration, and then takes the above geographic location as the location estimation of target IP. Both of algorithm analysis and experimental results show that, compare with the geolocation algorithm based on the shortest relative delay, the proposed algorithm could eliminate the accumulated error caused by relative delay, reduce the average and maximum error.

In the next work, we will focus on establishing more rules on calculating the conversion coefficient and candidate landmarks selection strategy while mapping the nearest common router. In addition, it is worth analyzing the statistical results of geolocation error when the probe and target IP are located in different cities.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (no. 61379151, 61572052 and U1636219), and the Outstanding Youth Foundation of Henan Province of China (no. 144100510001).

Conflict of Interests

All authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] 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.

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

[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, December, 2016.

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

[5] S. Liu, F. Liu, F. Zhao, L. Chai, X. Luo, IP City-level Geolocation Based on the PoP-level Network Topology Analysis, *Proceedings of IEEE Information Communication and Management (ICICM)*, Hatfield, UK, 2016, pp. 109-114.

[6] W. Jinxia, X. Xiaoyan, Y. Min, Z. Tianning, IP Geolocation Technology Research Based on Network Measurement, *Proceedings of IEEE Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, Harbin, China, 2016, pp. 892-897.

[7] Y. Shavitt, N. Zilberman, A Geolocation Databases Study, *IEEE Journal on Selected Areas in Communications*, Vol. 29, No. 10, pp. 2044-2056, December, 2011.

[8] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, B. Gueye, IP Geolocation Databases: Unreliable?, *ACM SIGCOMM Computer Communication Review*, Vol. 41, No. 2, pp. 53-56, April, 2011.

[9] P. T. Endo, D. Sadok, Whois Based Geolocation: A Strategy to Geolocate Internet Hosts, *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, Perth, Australia, 2010, pp. 408-413.

[10] N. Spring, R. Mahajan, D. Wetherall, Measuring ISP Topologies with Rocketfuel, *Proceedings of ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, Pittsburgh, PA, 2002, pp. 133-145.

[11] M. Zhang, Y. Ruan, V. Pai, J. Rexford, How DNS Misnaming Distorts Internet Topology Mapping, *Proceedings of USENIX Annual Technical Conference*, Boston, MA, 2006, pp. 369-374.

[12] B. Gueye, A. Ziviani, M. Crovella, S. Fdida, Constraint-based Geolocation of Internet Hosts, *IEEE/ACM Transactions on Networking*, Vol. 14, No. 6, pp. 1219-1232, December, 2006.

[13] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, Y. Chawathe, Towards IP Geolocation using Delay and Topology Measurements, *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, Rio de Janeiro, Brazi, 2006, pp. 71-84.

[14] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, C. Huang, Towards Street-level Client-independent IP Geolocation, *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation(NSDI)*, Boston, MA, 2011, pp. 365-379.

[15] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, Y. Zhang, IP-geolocation Mapping for Moderately Connected Internet Regions, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 24, No. 2, pp. 381-391, February, 2013.

[16] M. Li, X. Luo, W. Shi, L. Chai, City-level IP Geolocation based on Network Topology Community Detection, *Proceedings of IEEE Information Networking (ICOIN)*, Da Nang, Vietnam, 2017, pp. 578-583.

[17] S. Ding, X. Luo, D. Ye, F. Liu, Delay-distance Correlation Study for IP Geolocation, *Wuhan University Journal of Natural Sciences*, Vol. 22, No. 2, pp. 157-164, April, 2017.

[18] I. Youn, B. L. Mark, D. Richards, Statistical Geolocation of Internet Hosts, *Proceedings of the 18th IEEE International*

Conference on Computer Communications and Networks (ICCCN), San Francisco, CA, 2009, pp. 1-6.

[19] M. J. Arif, S. Karunasekera, S. Kulkarni, A. Gunatilaka, B. Ristic, Internet Host Geolocation Using Maximum Likelihood Estimation Technique, *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications(AINA)*, Perth, Australia, 2010, pp. 422-429.

[20] S. Laki, P. Mátray, P. Hága, T. Sebők, I. Csabai, G. Vattary, Spotter: A Model based Active Geolocation Service, *Proceedings of the 30th IEEE International Conference on Computer Communications(INFOCOM)*, Shanghai, China, 2011, pp. 3173-3181.

[21] B. Eriksson, P. Barford, J. Sommers, R. Nowak, A Learning-based Approach for IP Geolocation, *Proceedings of the 11th International Conference on Passive and Active Measurement*, Zurich, Switzerland, 2010, pp. 171-180.

[22] J. Chen, F. Liu, F. Zhao, G. Zhu, S. Ding, A SC-Vivaldi Network Coordinate System Based Method for IP Geolocation, *Journal of Internet Technology*, Vol. 17, No. 1, pp. 119-127, January, 2016.

[23] D. Cicalese, D. Joumblatt, D. Rossi, M. O. Buob, J. Augé, T. Friedman, A Fistful of Pings: Accurate and Lightweight Anycast Enumeration and Geolocation, *Proceedings of IEEE Conference on Computer Communications(ICCC)*, Kowloon, Hong Kong, 2015, pp. 2776-2784.

[24] O. Dan, V. Parikh, B. D. Davison, Improving IP Geolocation using Query Logs, *Proceedings of ACM International Conference on Web Search and Data Mining(WSDM)*, San Francisco, CA, 2016, pp. 347-356.

[25] B. Wong, I. Stoyanov, E. G. Sirer, Octant: A Comprehensive Framework for the Geolocation of Internet Hosts, *Proceedings of the 4th USENIX Conference on Networked Systems Design and Implementation(NSDI)*, Cambridge, MA, 2007, pp. 23-36.

[26] T. Vincenty, Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations, *Survey Review*, Vol. 23, No. 176, pp. 88-93, April, 1975.

Biographies

Jing-ning Chen was born in 1985. She received the B.S. degree, M.S. degree and Ph.D. from Zhengzhou Science and Technology Institute, Zhengzhou, China, in 2008 and 2011, respectively. Her primary interest lie in network entity geolocation.

Fen-lin Liu was born in 1964. He received his B.S. from Zhengzhou Institute of Science and Technology in 1986, M.S. from Harbin Institute of Technology in 1992, and Ph.D. from the East North University in 1998. Now, he is a professor of Zhengzhou Institute of Science and Technology. His research interests lie in network and information security.

Ya-feng Shi was born in 1976. He received his B.S. from Kashgar University and M.S. from Tianjin University. Now, he is a lecturer of mathematics and statistics of Kashgar University. His main research interests lie in set theory and mathematical basis.

Xiangyang Luo is a Professor at Zhengzhou Science and Technology Institute and the State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests lie in multimedia security and cyberspace surveying and mapping. He is the author or co-author of more than 100 refereed international journal and conference papers. He obtained the support of the National Natural Science Foundation of China and the National Key R&D Program of China.