

City-level Geolocation Algorithm of Network Entities Based on Landmark Clustering

Guang Zhu*, Xiangyang Luo*, Fenlin Liu*, and Fan Zhao*

*State Key Laboratory of Mathematical Engineering and Advanced Computing,

*Zhengzhou Science and Technology Institute, Zhengzhou 450001, China;

mail_guang@foxmail.com, corresponding author:luoxy_ieu@sina.com, liufenlin@vip.sina.com, zhaofan_123@yeah.net

Abstract—For the existing network entity geolocation algorithms based on the delay measurement are usually suffer from the inflation and jitter of the network delay, a new algorithm of city-level geolocation based on landmark clustering is proposed in this paper. According to the feature of the delays tend to be similar among the neighboring network entities generally, city-level network entities geolocation algorithm based on landmark clustering is given. In the proposed algorithm, the city-level landmarks are divided into several clusters based on the “delay similarity” principle to filters out the high-reliability landmarks for improving the precision of city-level geolocation algorithm, the minimal relative delay between cluster and target is calculated based on the minimal relative delay, and the region of this cluster is taken as the estimation of target’s location. The rationality of the delay clustering, feasibility of the proposed algorithm, and experimental results show that the proposed algorithm reduces effectively the delay influence on the IP geolocation algorithms, and the geolocation accuracy is higher than the classic GeoPing algorithm.

Keywords—network entity geolocation, landmark clustering, network delay, network measurement, city-level geolocation

I. INTRODUCTION

The network entity geolocation is the process of determining geographical location from an IP address [1]. For an IP address, a unique identifier for a network entity on the Internet generally, the network entity geolocation also is known as IP geolocation. The geographical location of IPs can be efficient to provide various personalized and targeted services for network users. For example, the commercial corporations can push targeted-advertisings and provide Location Based Services according to geographical locations of IPs; government departments can push local weather, and natural disaster warning for specified regions; the individuals also can judge credit card fraud, cope with junk mail and improve P2P (Peer to Peer) network performance [2][3], etc. A new and convenient way of cybercrime is provided by the Internet with the rapid development of the Internet, meanwhile, the detection and combat of cybercrime activities become complex increasingly. How to determine the geographical location of criminal and attacker is an important part of network detection work.

The original methods about network entity geolocation are GeoTrack, GeoPing, and GeoCluster, which were proposed in 2001 [4]. After more than a decade of development, the existing geolocation technologies mainly include Database-Based IP geolocation approaches, Measurement-Based approaches, and some other approaches.

Database-Based IP geolocation approaches: in this way, the geographical location of an IP is queried by the Internet public IP databases, such as the widely used Whois [5], MaxMind [6], IP2Location [7], Quova [8], etc. However, the information of databases is generally incomplete, outdated, or even false, which can be guaranteed the city-level granularity only [9]. What’s more, data source and collection technology of these databases are mostly private. The authenticity of the data and the reliability of the collection technology are difficult to evaluate. Thereby, the geographical location of an IP queried by this approach is low reliability and low accuracy [10].

Measurement-Based approaches: for this type of approaches, the delay or topology is measured. The delay and topology between target and landmarks (the geographical location are known) are measured, and the relationship between the target and landmarks are acquired to estimate the location of target, such as the classic GeoPing algorithm [4], learning-based approach [11], Spotter [12], CBG (Constraint-Based Geolocation) [13], SLG (Street-Level Geolocation) [14], and Octant algorithm [15], etc.

For the geolocation approaches, city-level geolocation is the first and crucial step of high-accuracy geolocation algorithms (such as SLG algorithm). For a target IP, a rough location estimation (city-level granularity) is conducted in the first step, and the target is analyzed and processed further to achieve a more accurate location estimation (street-level granularity). The latest city-level geolocation algorithms include learning-based approach [16] and city-level geolocation based on the routing feature [17]. The former addresses the machine-learning classification problem based on the Naive Bayes model with the lower computational complexity, and the latter is suitable for hierarchical network architecture. From the feature of hierarchical architecture, identifying routers are acquired from a large number of landmark samples according to the decision tree learning algorithm, and the identifying routers are taken as feature to judge whether a target network path contains the identifying routers, and then, the city-level location of target can be estimated.

The classic GeoPing algorithm and learning-based geolocation approach are based on the delay measurement. Delay-Based IP geolocation technology is one of reliable approaches to determine geographical location of an IP [18]. However, the network congestion, load balancing, heterogeneous network, poor performance of network equipment and other factors could give rise to delay inflation, network jitter, etc., which affects the accuracy of Delay-Based geolocation algorithms. Therefore, how to reduce the influence caused by the above situations on geolocation results, and how to describe accurately the

average delay for a region are crucial to Delay-Based network entity geolocation algorithms.

In order to improve the accuracy of Delay-Based network entity geolocation algorithms, an algorithm of city-level geolocation based on landmark clustering (GBLC) is proposed in this paper. According to the “delay similarity” principle, the delays tend to be similar among the neighboring network entities generally, the city-level landmarks are divided into several clusters to filter out the high-reliability landmarks; and the minimal relative delay between cluster and target is calculated based on the minimal relative delay; next the region of this cluster is taken as the estimation of target’s location.

II. CITY-LEVEL GEOLOCATION ALGORITHM OF NETWORK ENTITIES BASED ON LANDMARK CLUSTERING

To improve the above situations in Delay-Based network entity geolocation algorithms and the reliability of the delay measurement, an algorithm of city-level geolocation based on landmark clustering is proposed in this paper. The proposed algorithm is based on clustering, ADVs of the specified regions are calculated according to the clustering algorithm, which are more close to the average delay feature for a specified region. And then the “minimal relative delay” is taken as principle to estimate the city-level geographical location of target IP. This algorithm is expected to reduce delay inflation, delay jitter, and other cases caused by the network congestion, load balancing, heterogeneous network, poor performance of network equipment and other factors.

The proposed algorithm includes clustering part and geolocation part, the main flow is shown in Figure 2. In the aspect of clustering: a number of VPs are deployed at different places, each VP sends packets to measure the Round Trip Time (RTT) between itself and all the landmarks. According to the delays measured by each VP, the DV is built for each landmark. And then, all the landmarks with same region are divided into different groups. Next, for the same region, DVs are clustered into multiple clusters according to clustering algorithm and the maximal cluster is selected to calculate its centroid, and used to describe ADV of the region. In the aspect of geolocation part: a number of VPs measure the RTTs from themselves to target, and the target DV is built. Then, According to the ADV of each region, RDV is calculated from target to each region. Finally, the geographical location of region with the minimal relative delay is taken as the final location estimation of target.

The major steps of GBLC algorithm are shown below.

Input: target IP;

Output: city-level estimation location of target IP;

Step1. Selecting city-level landmarks: under the precondition of uniform distribution and under the same Internet Service Provider (ISP), landmarks, for multiple candidate regions, are selected as the reference nodes of geolocation algorithm.

Step2. Building landmarks DV: according to the requirements of geolocation accuracy, m VPs are deployed at different places, each VP sends the packets to measure RTTs from itself to all the landmarks, namely, $d_{i,1}, d_{i,2}, d_{i,3}, \dots, d_{i,m}$, where i stands for the i -th landmark, m means m VPs. And $DV = (d_{i,1}, d_{i,2}, d_{i,3}, \dots, d_{i,m})$ is built for each landmark.

Step3. RTTs clustering: for all the DVs of reference nodes, which take the region as a unit to divide into different groups. And DVs with same group are clustered according to K -Means algorithm, where K is

an experience threshold, determined by lots of experiments and combine with the user requirements.

Step4. Calculating ADV: for each region, the maximal cluster is selected to calculate its centroid, and its centroid is taken as the average delay vector $ADV = (delay_1, delay_2, \dots, delay_m)$, of the region, shown as (1), where $MaxCluster()$ is used for selecting the maximal cluster, and $Centroid()$ is used for calculating the centroid, m means m VPs, $delay_m$ means the average delay from the m -th VP to the region.

$$ADV = Centroid(MaxCluster(cluster_1, \dots, cluster_k)) \\ = (delay_1, delay_2, \dots, delay_m) \quad (1)$$

Step5. Building target DV: same as Step3, RTTs of target are measured and Delay Vector DV_{target} is built according to the measurement results of multiple VPs, shown as (2), where i stands for the i -th VP, $target$ means target, $GetTargetRTT()$ is used for measuring RTT of target, and DV_{target} is the DV of target.

$$DV_{target} = GetTargetRTT(probes, target) \\ = (d'_{i,1}, d'_{i,2}, \dots, d'_{i,m}) \quad (2)$$

Step6. Calculating RDV: target is the passive network entity, in general, and cannot send the packets. Therefore, we adopt the relative delay to calculate RTT between target and cluster. For each region, we combine ADV and (3) to calculate the RDVs between target and each clusters, where i stands for the i -th VP.

$$RDV = (|delay_1 - d'_{i,1}|, \dots, |delay_m - d'_{i,m}|) \quad (3)$$

Step7. Calculating the minimal RDV: the relative delays from target to each cluster are calculated according to ADV and (4), where i stands for the i -th VP, $min()$ is used for selecting the cluster with the minimal relative delay. And the minimal RDV is selected. Finally, the geographical location of this region with the minimal relative delay is taken as the final location estimation of target.

$$RDV = \min(\frac{1}{m}(|delay_1 - d'_{i,1}| + \dots + |delay_m - d'_{i,m}|)) \quad (4)$$

III. RATIONALITY ANALYSIS ON THE DELAY CLUSTERING AND THE FEASIBILITY OF GBLC ALGORITHM

For Delay-Based geolocation algorithms, generally, the reliability of the delay measurement is one of the important factors to the geolocation accuracy. However, for the real Internet, network congestion, load balancing, heterogeneous network, poor performance of network equipment and other factors can always influence the delay measurement, such as delay inflation, delay jitter, and other cases, which could affect the accuracy of geolocation result.

The clustering is based on similarity, and divides data objects into different clusters, which have high similarities in the same cluster [19]. Reference [4][20] have revealed the “delay similarity” phenomenon, namely, the delay of network entities with the similar geographical location distribution often have similarity.

To verify the rationality of landmark clustering, 130, 130, and 90 city-level landmarks under the same ISP (China Telecom) in the cities of Hebi, Xuchang, and Zhengzhou are selected, respectively. Two VPs in Zhengzhou and Qingdao are deployed. RTTs are measured from VPs to all of the landmarks, these landmarks are clustered by K -Means algorithm, and set $K = 2$. Taking Hebi, Xuchang, and Zhengzhou as the examples, and the clustering results are shown in Figure 1.

We can see from Figure 1, when the number of VPs is two, DVs can be divided into two clusters by K-Means algorithm, where the green part is the maximal cluster and the red part is the smaller clusters in clustering result. The RTTs of the maximal cluster is more suitable to describe the delay feature of the region accurately. This suggest that there is a strong correlation in RTTs from VPs to landmark set. We can screen out the delays with relative concentration and stabilization as the average delay vector of a region. The differences and abnormal delays can be removed to reduce influence on accuracy of Delay-Based geolocation algorithms.

Therefore, the accuracy of Delay-Based geolocation algorithms can be improved. Furthermore, due to the ISP lines and routing configuration, etc., landmark set, VPs, and targets are selected under the same ISP. Under the same ISP, RTTs from VPs to landmark set are suitable to measure the average delay of a region. In conclusion, an algorithm of city-level geolocation based on landmark clustering is feasible.

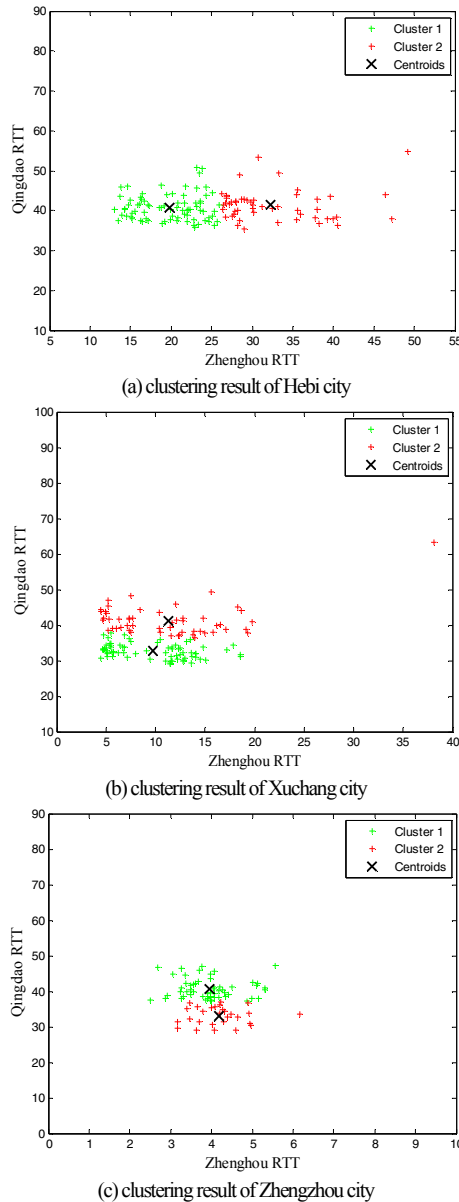


Figure 1. K-Means clustering results. (a), (b), and (c) stand for the clustering results of Hebi, Xuchang, and Zhengzhou.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Three VPs are deployed in Zhengzhou, Qingdao and Beijing, which belong to the same ISP (China Telecom). The geolocation results of the proposed algorithm are compared with the classic GeoPing algorithm.

A. Landmarks clustering

130, 130, and 90 city-level landmarks are distributed in Hebi, Xuchang, and Zhengzhou, respectively, which belong to China Telecom.

Landmarks are clustered by three VPs, respectively, and the centroid of the maximal cluster is calculated. And the centroid is taken as the region's average delay vector ADV, the related data is shown in Table I, where M means the number of VPs.

TABLE I. THE CALCULATION OF ADV FOR HEBI, XUCHANG, ZHENGZHOU CITIES

City	Landmark	Target	Centroid / ADV
Hebi	130	30	(19.84,40.83,22.49)
Xuchang	130	30	(10.69,35.86,15.49)
Zhengzhou	90	30	(4.14,38.31,14.70)

B. Comparison Of Geolocation Results

After ADVs of the above cities are determined, Hebi, Xuchang, and Zhengzhou city, 90 reliable IP addresses with known geographical location are selected as the input of GeoPing and GBLC algorithm to be geolocated, respectively. The geolocation results is shown in Table II.

TABLE II. GEOLOCATION RESULTS COMPARISON WITH GEOPING[4]

Algorithm	Target	Correct Number	Accuracy
GeoPing	90	75	83.3%
GBLC	90	82	91.1%

The Table II shows that: 1) when the number of VPs is 3, the geolocation results correct number of GeoPing is 75, and the correct number of GBLC is 82, which improves 7.8% compared with GeoPing. In conclusion, the proposed algorithm adopts clustering, regional average delay vector, the minimal relative delay, etc. to reduce delay jitter and other cases, which improves the geolocation accuracy at some extent. Compared with the traditional Measurement-Based delays, GBLC algorithm process them further, and a more reliably delays are acquired as the input of city-level geolocation algorithms. Therefore, the accuracy of geolocation results is higher.

V. CONCLUSION

In order to improve the accuracy of network entity city-level geolocation algorithms, an algorithm of city-level geolocation based on landmark clustering is proposed in this paper. According to the feature which the delays tend to be similar among the neighboring network entities generally, city-level landmarks are clustered into multiple clusters to filter out the high-reliability landmarks for improving the accuracy of Delay-Based geolocation algorithms. Experimental results based on 350 landmarks and 90 targets in 3 cities show, for different quantity of VPs, compared with the classic GeoPing algorithm, the proposed algorithm has a better geolocation accuracy. Next, we will

study the other factors influencing the landmarks clustering, such as the population density, network scale, to improve the reliability of city-level geolocation further.

VI. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No.61379151, No.61274189, No.61572052, No.U1536102), the Excellent Youth Foundation of Henan Province of China (No. 144100510001) and the Foundation of Science and Technology on Information Assurance Laboratory (No. KJ-14-108).

REFERENCES

- [1] Muir J A, and Oorschot P C V, "Internet geolocation: Evasion and counterevasion," ACM Computing Surveys, vol. 42, no. 1, pp. 5515-5528, 2009.
- [2] Liu H, Zhang Y, and Zhou Y, et al., "Mining checkins from location-sharing services for client-independent IP geolocation," Proceedings of the 33rd IEEE Conference on Computer Communications (INFOCOM), pp. 619-627, 2014.
- [3] Guo C, Liu Y, and Shen W, et al., "Mining the Web and the Internet for Accurate IP Address Geolocations," Proceedings of the 28th IEEE Conference on Computer Communications (INFOCOM), pp. 2841-2845, 2009.
- [4] Padmanabhan V N, and Lakshminarayanan S, "An investigation of geographic mapping techniques for Internet hosts," ACM SIGCOMM Computer Communication Review, vol. 31, no. 4, pp. 173-185, 2001.
- [5] Whois, <http://www.whois.com/>.
- [6] MaxMind, <https://www.maxmind.com/>.
- [7] IP2Location, <http://www.ip2location.com/>.
- [8] Qouva, <http://www.quova.com/>.
- [9] Poese I, Uhlig S, and Kaafar M A, et al., "IP geolocation databases: Unreliable?" ACM SIGCOMM Computer Communication Review, vol. 41, no. 2, pp.53-56, 2011.
- [10] Siwipersad S S, Gueye B, and Uhlig S, "Assessing the geographic resolution of exhaustive tabulation for geolocating internet hosts," Proceedings of the 9th Passive and active network measurement (PAM), pp. 11-20, 2008.
- [11] B. Eriksson, P. Barford, and J. Sommers, et al., "A learning-based approach for IP geolocation," Proceedings of the 11th International Conference on Passive and Active Measurement (PAM), pp. 171-180, 2010.
- [12] Matray, Haga, and Sebok, et al., "Spotter: A model based active geolocation service," Proceedings of the 30th IEEE Conference on Computer Communications (INFOCOM), pp. 3173-3181, 2011.
- [13] Gueye B, Ziviani A, and Crovella M, et al., "Constraint-based geolocation of internet hosts," IEEE/ACM Transactions on Networking, vol. 14, no. 4, pp. 1219-1232, 2006.
- [14] Wang Y, Burgener D, and Flores M, et al., "Towards Street-Level Client-Independent IP Geolocation," Proceedings of the 8th Conference on Networked Systems Design and Implementation (NSDI), pp. 27-40, 2011.
- [15] Wong B, Stoyanov I, and Sirer E G, "Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts," Proceedings of 4th Conference on Networked Systems Design and Implementation (NSDI), pp. 313-326, 2007.

- [16] B. Eriksson, P. Barford, and J. Sommers, et al., "A learning-based approach for IP geolocation," Proceedings of the 11th International Conference on Passive and Active Measurement (PAM), pp. 171-180, 2010.
- [17] Zhao F, Song Y, and Liu F, et al., "City-Level Geolocation Based on Routing Feature," Proceedings of the 29th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 414-419, 2015.
- [18] Taylor J, Devlin J, and Curran K, "Bringing location to IP addresses with IP geolocation," Journal of Emerging Technologies in Web Intelligence, vol. 4, no. 3, pp. 273-277, 2012.
- [19] Han J, and Kamber M, "Data mining concept and technology," Second Edition, Elsevier Inc., pp. 251-263, 2007.
- [20] Ziviani A, Fdida S, and De Rezende J F, et al., "Similarity models for Internet host location," Proceedings of the 11th IEEE International Conference on Networks (ICON), pp.81-86, 2003.



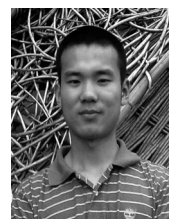
ZHU Guang was born in Henan Province, China, in 1990. ZHU got the B.S. degree from Henan University of Technology, Zhengzhou, China, in 2013. He is currently a M.S. candidate in The State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou Information Science and Technology Institute. His research interest includes landmark mining of network nties, IP geolocation and information Security.



LUO Xiang-Yang was born in Hubei Province, China, 1978. Luo received the B.S. degree, the M.S. degree and the Ph.D. degree from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2001, 2004 and 2010, respectively. He is the author or co-author of more than 70 refereed international journal and conference papers. He is also a guest editor for "International Journal of Internet" and "Multimedia Tools and Applications". LUO is currently an associate professor of Zhengzhou Science and Technology Institute. His research interest is Networking and Information Security.



LIU Fen-Lin was born in in Jiangsu Province, China, 1964. He received his B.S. from Zhengzhou Institute of Information Science and Technology in 1986, M.S. from Harbin Institute of Technology in 1992, and Ph.D. from the Eastnorth University in 1998. Now, he is a professor of Zhengzhou Information Science and Technology Institute. His research interests include Networking and Information Security.



ZHAO Fan, born in 1989, M. S. candidate. He is currently a M.S. candidate in The State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou Science and Technology Institute. His research interests focus on network security.