ATLANTIS PRESS

Advances in Computer Science Research, volume 44
3rd International Conference on Wireless Communication and Sensor Network (WCSN 2016)

Method for IP Geolocation Based on Path Similarity

Lian-Xing Ren
Network Engineering Department, Electronic Engineering Institute, Hefei, China
E-mail: 18655193001@163.com

***Abstract*-The ability to accurately determine the geographic location of an arbitrary IP address has potential in many applications. This paper presented a novel algorithm for IP geolocating based on network measurement. Considering the features of China Internet's distribution, using trace route tools, we build database about the routing path to landmarks. Based on the database, this paper introduced GeoPath to geolocate a target quickly by path similarity. A set of landmarks with known location were tested by the proposed method and experimental results demonstrate that the method is validity and practical.**

Keywords-IPGeolocation; measurement; delay; traceroute; path similarity; landmark

I. INTRODUCTION

The problem of locating the geographical location of an Internet host is an important research that is currently addressed by many research groups. IP geolocation is the process of locating an Internet host or device that has an IP address. IP geolocation has many important applications to targeted advertising, restriction of content delivery, automatic selection of language to display web site content and Internet frauds, etc.

Nevertheless, determining geographical location of internet hosts by a single IP address poses many challenges, since there is no direct relationship between the IP address of a host and its geographical location. Many geolocation services are based on databases which store organizational information of IP address and geography. The accuracy of these services is usually insufficient due to the lack of reliable information. Furthermore, the databases are often proprietary and manually updated, their consistency and accuracy are questionable at best.

To overstep the uncertainty of techniques based on database, many techniques attempt to locate an host by network measurements. Many automatic IP address geolocation techniques based on landmarks and active delay measurement have been proposed in recent years.

The measurement-based geolocation of a particular target t generally consists of four steps:

- *Select a set of landmarks;*
- *Measure delay from target t to each landmark;*
- *Convert delays to distances;*
- *Use distances and known landmark locations to form a geolocation estimate for target t.*

II. RELATED WORK

IP Geolocation techniques can be classified into three different categories: database-based, measurement-based, and data-mining-based .In this section, we discuss three typical geoloation schemes.

A. GeoPing

In GeoPing[1], a set of hosts with known geographical location, called active landmark or vantage points, perform netwok measurements by transmitting ICMP ping packets between each other and to the passive landmarks. The index set foe all landmarks is given by $L = L_a \cup L_p$ .Let d_{ij} denote the measured delay between l_i and l_j ,where $l_i \in L_a$, $l_j \in L_p \cup L_a$ .Then the location of the target is defined as l_k ,where

$$k = \arg\min_{j \in L} \{ \sum_{i \in La} (d_{ij} - d_{it})^2 \}$$

Therefore, the accuracy of GeoPing depends on the "closest" landmarks. When the target is relatively far from the "closest" landmarks, the estimation error will be significant.

B. Constraint-Based Geolocation

The literature [2] proposed an approach to IP geolocation based on an idea from the field of landmarks, called Constraint-Based Geolocation(CBG), which infers the geographic location of Internet hosts using multilateration with distance constraints. Multilateration refers to the process of estimating a position using a sufficient number of distances to some fixed points, thus establishing a continuous space of answers instead of a discrete one. The geographic distances from the landmarks to the target host have to be estimated based on delay measurements between these vantage points. Then the relationship of delay and distance is obtained.The paper draws a straight line L so that all (x, y) points just above the L that is called "bestline". The

slope of bestline k is the upper limit factor of the ratio of delay to distance.

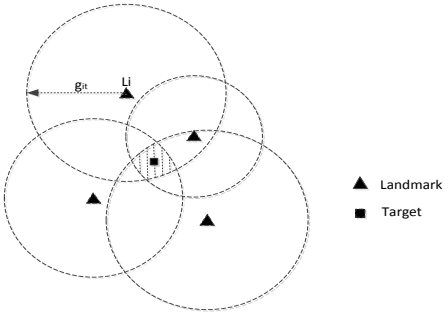

Figure 1. Principle of CBG

When positioning the target T, the delay between each VP point and the target host, $d_{i\tau}$ is measured. According to k and $d_{i\tau}$, the time delay is converted to geographical distance $g_{i\tau}$. Then, make a circle C_i, with VP_i as the center, with a radius of $g_{i\tau}$ and the target should be inside of C_i. By making several circles, the area of the target will be smaller and smaller.

C. IP geolocation by database

This is the current application of a wide range of geolocation methods, such as QQ ChunZhen[3] database,IP138[4] and MaxMind[4]. They collect relevant IP information and establish the mapping table between IP and location. The disadvantage of database is that the accuracy is difficult to determine.

In addition, TBG[10] improves on these earlier techniques by leveraging network topology, along with measurements of network delay, to constrain host position. The paper[11] introduce a detailed path-latency model to be able to determine the overall propagation delays along the network paths more accurately. Dima Feldman and Yuval Shavitt [13]try to collect large scale PoP level maps using traceroute measurement to locate the targets. But the methods all need a large number of distributed vantage points to provide enough information for geolocation.

III. IP-GEO BY PATH

A. China'S Internet

China's Internet[12] consists mostly of several large ISPs, such as CHINANET, UNINET and CMNET. The major ISPs' structure is hierarchical. China's Internet is very different from the Internet in US and Europe and has its unique features: the number of AS is not a lot, relatively simple structure. Therefore, the topology of China's Internet depends on large ISPs' interior structure. Nevertheless, China's Internet lacks the infrastructure and resources that are essential for large-scale Internet measurement studies, such as PlanetLab[6], iPlane[7] and Rocketfuel[8]. For example, China has few PlanetLab nodes and looking glass servers. That is we do not have plenty of landmarks for network measurement.

China's major ISPs are always divided in three parts: core layer, convergence layer and access layer .The routers' interfaces are assigned to different municipalities or provincial capitals. The routers in convergence layer belong to different cities and the routers in access layer lay in different cities or country.

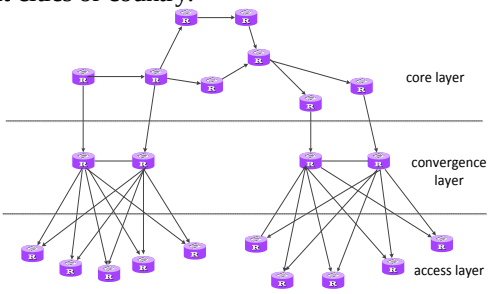

Figure 2. Structure of china's major ISP.

Suppose that ICMP packets are transmitted from s to τ .The path can be devided into three parts: local network, backbone network and target network. For example, we launch a traceroute to a target host and gain path information as follow.

TABLE I. AN EXAMPLE OF TRACEROUTE PATH

Hop	delay(RTT)	IP address
1	1ms	60.166.208.1
2	1ms	61.190.244.237
3	1ms	202.102.206.245
4	25ms	202.97.66.86
5	18ms	120.36.77.94
6	25ms	202.109.195.254
7	17ms	61.154.9.68

The difference of the first three RTTs is not huge. So we can estimate that the first three hops are in local network. But the RTT in the forth hop's delay increase rapidly. We conclude that the forth hop is not in local network. By querying in ChunZhen and MaxMind databases, the 4th hop IP address is located in ChinaNet backbone network.

There is small difference between The 6th RTT and the 7th RTT .So we estimate the last two hops is lay in the target network. The result of geolocation by ChunZhen and MaxMind database has confirmed our belief.

It must be noted that, rapid increase of delay perhaps happens one or two times in a traceroute measurement because of ISP's structure.

The rapid increase of delay(RID)can defined as below: the phenomenon of delay increase rapidly because of geographical span. According to the previous detection data, when the difference between two adjacent hops is equal or greater than 10ms, it is considered generally that there is RID. In other words, the two adjacent routers are not in the same place.

As mentioned above, the major ISPs are hierarchical in China's Internet. The core layer of ISPs is distributed mainly in BeiJing, ShangHai and GuangZhou, which connects the access points in every capital cities. The capital cities lay in the convergence layer and the regional cities lay in the access layer. In the same way, the routers in the convergence converge the routers in one area

According to the RID, we can discover the link between core layer and convergence layer. With the help of the relationship between "distance" and "delay", we also restrict the location to a specific area such as province and region. But the distance of link in a small region can't be estimated precisely because of its geographical span. Therefore, we should accomplish the rest of work with the aid of other information in network topology.

Some routers and hosts don't answer the ICMP requests, which makes it difficult to accomplish the traceroute measurements. This is the cause of incomplete traceroute paths. Many researches abandon the incomplete traceroute paths because they can't accurately reflect the delay information. We launched a traceroute to 2564 websites which distributed in a number of cities and gain 235 incomplete paths. We located the last hops incomplete paths which traceroute can reach by ChunZhen and MaxMind. With further research, we find that 211 last hops' are in the same place with their target respectively. The ratio reaches 89.7%. This shows that many incomplete traceroute have reached the target network. Consequently, the incomplete traceroute can be useful in the IP geolocation.

B. Trceroute Path

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

This article launch traceroute measurements from one source(HeFei, Anhui province)to various target hosts. The different targets, which are close to each other geographically, have similar paths because of China Internet's feature. For example, we send a set of IMCP packets to host A and host B respectively and gain their traceroute paths, p_A and p_B .

TABLE II. TRACEROUTE PATH TO TARGET A

Hop	delay(RTT)	IP address
1	1ms	60. 166. 208. 1
2	1ms	61. 190. 244. 237
3	***	* * *
4	19ms	202. 97. 42. 178
5	29ms	220. 162. 234. 246
6	26ms	61. 131. 101. 206
7	29ms	61. 131. 39. 242
8	25ms	61. 131. 37. 239

TABLE III. TRACEROUTE PATH TO TARGET B

Hop	delay(RTT)	IP address
1	1ms	60. 166. 208. 1
2	1ms	61. 190. 244. 237
3	5ms	202. 102. 206. 29
4	24ms	202. 97. 66. 82
5	22ms	220. 162. 234. 2
6	31ms	61. 131. 101. 42
7	***	***

As shown in the table, p_A is complete, while p_B is incomplete. By comparing p_A to p_B , the similarity of path is obvious. The two paths consist of local network hops(1,2,3),backbone network hops(4)and target network hops(5 and latter) .It is important that both paths contain the same network segment 61.131.101.0/24.Furthermore,the segment 61.131.101.0/24 is in the target network. There are only 256 IP addresses in a /24 network segment. So 61.131.101.206 and 61.131.101.42 are very likely located in the same place.

Through geoloacting by ChunZhen and MaxMind database, we find that both IP addresses are in SanMing city, FuJian province. Therefore, if the position of a landmark is known, and the position of target is unknown, then according to similarity of paths, it can be inferred that the orientation of A and B is close geographical.

C. Path Similarity

Let s is source, $L=\{l_1,l_2,.....,l_n\}$ is the set of landmarks, τ is the target ti be located. The path from s to l_i is denoted as $P_{sli}=(h_1\rightarrow h_2\rightarrow\ldots\ldots\rightarrow h_k)$ and hi is the i-th hop in the path.

Let $h_p\in Psli$ and $h_q\in Pslj$ .If h_q and h_p are in the same segment and both p and q are larger than the threshold value .Then they have a common intersection point and are similar.

Considering there are only 256 IP addresses in a /24 segment, the different IP addresses in the same /24 segment should be located in the same place. That is to say, /24 segment is the minimal granularity in geolocation. Therefore, when hp and hq are in the same /24 segment, it can be identified that two paths have a intersection point.

In this paper, the source is regular throughout the measurements, which makes it possible that there are some intersection points close the local network. To avoid disturbing the geolocation, we should set a threshold T to ensure that the intersection points not be in the side of local network. We set the value of threshold to 4 in accordance with specific conditions.

D. IP Geolocation Space

Based on the above indicators, the paper gives the basic element of the IP geolocation space:

- *Source:* s *the source of detection;*
- *Target:* $T=\{\tau_1,\tau_2,.....,\tau_n\}$ *the targets to be located;*
- *Path:* $P_{s\tau}=(h_{\tau 1},h_{\tau 2},.....,h_{\tau m})$ *, Where m is the number of valid hops;*
- *Landmarks set:* $L=\{l_1,l_2,.....,l_p\}$ *, Where p is the number of landmarks;*
- *Paths set of Landmark:* $P_L=\{P_{sl_1},P_{sl_2},.....,P_{sl_p}\}$ *,where* P_{sli} *is the traceroute path from* s *to the landmark* l_i *;*

Traceroute detection in the mainland of China can be completed within 15 hops. Therefore, the landmark path node information can be expressed as:

$$H_{n\times 15}=\begin{pmatrix} h_{11} & h_{12}\cdots & h_{1k_1}...0 \\ h_{21} & h_{22}\cdots & h_{2k_2}...0 \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nk_n}...0) \end{pmatrix} \quad 1\le k_i\le 15$$

, where h_{ij} is the j-th hop in the path to landmark l_i ;

E. Geolocation Algorithm

Based on the above analysis, we propose a method of IP geolocation based on path similarity. First of all, we launch traceroute detection to the selected passive landmark to get the corresponding path information,and constantly expand the passive landmarks set, to establish a more complete path information. Nextly, we can also learn from the experience of GeoPing and locate the target in the "nearest" landmark by path similarity.

It has been found that the government website is generally suitable as passive landmark nodes through analysis. Because the vast majority of governments and their websites in the same place. Furthermore, government websites distribute widely and evenly, which can guarantee the landmark density (landmarks density) as appropriate. We collect goverment websites of different cities in China as passive landmarks set with the benefit of crawler program. Then through traceroute detection, the path information is obtained, which is used as the initial set of landmarks' paths.

On the basis of establishing the set of landmarks' paths, this paper gives the main steps of geolocation algorithm as follows:

Let path of the target to be located is $P_{sr}=(h_{r1},h_{r2},.....,h_{rm})$, where m is the number of valid hops:

- Let h_{pq} in P_{sr} ,if there exists $hij\in H$ that h_{ij} and h_{pq} are located in the same /24 segment,then Add P_i to an alternative set of landmarks $L^{(1)}$,to (3); else, to (2);
- If $j>threshold$, then $j--$,return (1);
- if $\left|L^{(1)}\right|=0$,failure;
- if $\left|L^{(1)}\right|=1$,then the landmrak which hasthe only element is the result of geolocation;
- if $\left|L^{(1)}\right|\ge 1$,the landmark which has maximum mask matching hop is the result of geolocation.

IV. EXPERIMENTS RESULTS

Currently measurement-based approaches generally use the Distance Error as the standard of evaluation. But there are no such large scale distributed measurement infrastructures as PlanetLab or iPlane in China. Therefore, this paper selects a set of websites with known geographical location as the experimental object, by comparing the differences between actual position and located position, to verify the effectiveness of the algorithm.

The effect of IP geolocation is generally measured from two aspects of coverage and precision. Coverage is achieved by measuring the null response rate. Accuracy is achieved by measuring the Distance Error.

TABLE IV. NON-REPLY RATIO

Methods	*Sites number*	*Located Sites number*	*Non-reply ratio*
GeoPath	***357***	***319***	***10.64%***
ChunZhen	***357***	***357***	***0***
MaxMind	***357***	***302***	***15.4%***

It can be seen from the table above, due to the widely distributed software application of Tencent, ChunZhen's response rate is the highest. MaxMind is the foreign service abroad and therefore China's regional coverage rate is a lot less. Although the GeoPath algorithm in single point only rely on the network detection conditions, its coverage rate is close to 90%.

In order to ensure the site geography distribution widely, we collected 510 local colleges and local goverment website for test. The 357 sites which are consistent with the results of the web location and the ChunZhen database are selected because of the phenomenon of server trusteeship. We use the above 357 sites as a test set. Due to the network jitter, each time the delay data acquisition may be different. So the method of measuring the minimum value at different time period is used to determine the final measurement result. Meanwhile, the landmarks are located by the MaxMind database. The experimental results are shown in fig 3:

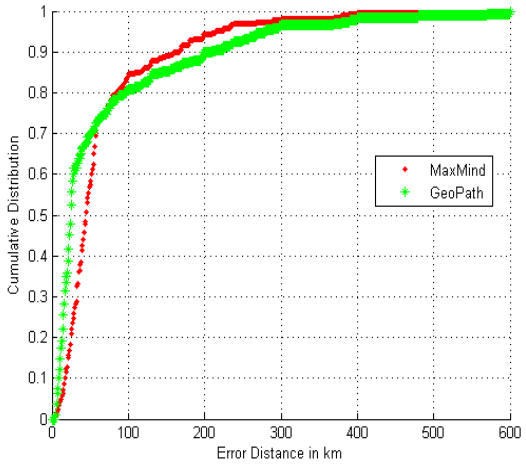

Figure 3. Cumulative probability of error distance.

TABLE V. ACCURACY WITH DIFFERENT METHODS

Accuracy with different methods			
settings	***Mean error***	***Median error***	***Maximum error***
GeoPath	69.19	25	1200
MaxMind	70.22	45	678

For a set of target paths P_T that have been verified, we can merge it with an initial set of landmark paths P_L into a new set $P_{NEW} = P_L \bigcup P_T$ and further improve the coverage of paths set.

V. RELATIVE RESEARCH COMPARISON

The proposed method is compared with several methods mentioned above. The paper compare the methods in several aspects: the amount of processing data, the need for a priori knowledge and the amount of measurement points.

TABLE VI. COMPARISON OF SEVERAL METHODS

Methods	*a large number of VP nodes*
ChunZhen	*Unknown*
MaxMind	*Unknown*
IP2Geo	*Y*
CBG	*Y*
GeoPath	*N*

As is shown in chart, compared with the current common positioning methods, the method proposed in this paper is more practical and feasible, mainly in the following aspects:

- GeoPath only need a source to complete the detection, greatly reducing the difficulty of geolocation. Current methods such as CBG,IP2Geo,Structon , require hundreds or even thousands of distributed vantage points to achieve the measurement.
- Relative to the construction of the IP database, GeoPath to collect the amount of data to be much smaller. The number of IP records in GeoPath is not more than 30000,while the ChunZehn database has 446615 IP records.(October 25, 2016).
- GeoPath will not take up too much memory due to the processing of the data is less, while the requirements of CPU is not high.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented GeoPath , a new approach that use traceroute measurement together with path similarity, for IP address to geolocation mapping. By the single-point measurement, this method obtains the path information to different passive landmarks. According to the similarity comparison between the target path and the known landmark, the orientation of the target is inferred. The experimental results show that the method is effective. In addition, compared with other measurement-based methods, this approach can significantly reduce the implementation difficulty. GeoPath can expand the path set by the new landmark measurement, so as to further enhance the coverage and accuracy of geolocation.

REFERENCES

[1] Padmanabhan, V.N., Subramanian, “An investigation of geographic mapping techniques for internet hosts”. SIG-COMM Comput. Commun. Rev. 31, 173-185 (2001).

[2] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, “Constraint-based geolocation of Internet hosts,” IEEE/ACM Trans. Net., vol. 14, no. 6,pp. 1219 – 1232, 2006.

[3] QQWry. http://www.cz88.net/

[4] IP138. http://www.ip138.com/J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[5] Maxmind. http://dev.maxmind.com/geoip/legacy/geolite

[6] PlanetLab. http://www.planet-lab.org, 2007.

[7] iPlane Project. http://iplane.cs.washington.edu/

[8] N. Spring, R. Mahajan, and D. Wetherall. "Measuring ISP topologies with Rocketfuel". In ACM SIGCOMM, pages 133–145, 2002.

[9] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang. “Mining the web and the Internet for accurate IP address geolocations.” Proc. of INFOCOM'09, Rio de Janeiro, Brazil, 2009. Structon

[10] Katz-Bassett, E., John, J.P., Krishnamurthy, A., Wetherall,D., Anderson, T., Chawathe, Y. “Towards ip geolocation us-ing delay and topology measurements.” In: Proceedings of the 6th ACM SIGCOMM conference on Internet mea-surement, IMC '06, pp. 71-84. ACM, New York, NY, USA(2006).

[11] S.Laki, P.Matray,and G. Vattay. Spotter: "A model based active geolocation service." In IEEE INFOCOM 2011,Shanghai, China, 2011.

[12] Y. Tian, R. Dey, Y. Liu, and K. W. Ross.“China’s Internet: Topology mapping and geolocating,” University of Sci. & Tech. of China, Tech.Rep., 2011,

[13] Y. Shavitt and N. Zilberman. “A structural approach for PoP geo-location.” Proc. of INFOCOM Workshop on Network Science for Communications (NetSciCom), San Diego, CA, 2010.