IJDCST @Dec, Issue- V-3, I-1, SW-41
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

A Novel Scheme to Map Moderately Connected Internet Regions

Amerineni Tejaswini[1], Shaik. Gouse John[2]

[1]M.Tech (CS), Nimra College of Engineering and Technology, A.P., India.

[2]Asst. Professor, Dept. of Computer Science & Engineering, Nimra College of Engineering and Technology, A.P., India.

Abstract— Knowing geographical locations becoming necessity more than facility nowadays. Most IP-geolocation mapping schemes [14], [16], [17], [18] take delay-measurement approach, based on the assumption of a strong correlation between networking delay and geographical distance between the targeted client and the landmarks. In this paper, however, we investigate a large region of moderately connected Internet and find the delay-distance correlation is weak. But we discover a more probable rule—with high probability the shortest delay comes from the closest distance. Based on this closest-shortest rule, we develop a simple and novel IP-geolocation mapping scheme for moderately connected Internet regions, called GeoGet. In GeoGet, we take a large number of webservers as passive landmarks and map a targeted client to the geolocation of the landmark that has the shortest delay. We further use JavaScript at targeted clients to generate HTTP/Get probing for delay measurement. To control the measurement cost, we adopt a multistep probing method to refine the geolocation of a targeted client, finally to city level. The evaluation results show that when probing about 100 landmarks, GeoGet correctly maps 35.4 percent clients to city level, which outperforms current schemes such as GeoLim [16] and GeoPing [14] by 270 and 239 percent, respectively, and the median error distance in GeoGet is around 120 km, outperforming GeoLim and GeoPing by 37 and 70 percent, respectively.

Keywords — IP geolocation, GeoGet, moderately connected Internet

I. INTRODUCTION

Today's applications will benefit from or be enabled by knowing the geographical locations (or geolocations) of Internet hosts. Such locality-aware applications include local weather forecast, the choice of language to display on webpages, targeted advertisement, page hit account in different places, restricted content delivery according to local policies, etc. Locality-aware peer selection will also help P2P applications in bringing better user experience as well as reducing networking traffic [1], [4].

Traditional IP-geolocation mapping schemes [5], [16], [8], [9] are primarily delay-measurement based. In these schemes, there are a number of landmarks with known geolocations. The delays from a targeted client to the landmarks are measured, and the targeted client is mapped to a geolocation inferred from the measured delays. However, most of the schemes are based on the assumption of a linear correlation between networking delay and the physical distance between targeted client and landmark.

The strong correlation has been verified in some regions of the Internet, such as North America and Western Europe [5], [6]. But as pointed out in the literature [6], the Internet connectivity around the world is very complex, and such strong correlation may not hold for the Internet everywhere.

In this paper, we investigate the delay-distance relationship in a particular large region of the Internet (China), where the Internet connectivity is moderate. The data set contains hundreds of thousands of (delay, distance) pairs collected from thousands of widely spread hosts. We have two observations from the data set. First, the linearity between the delay and distance in this region of Internet is positive but very weak. Second, with high probability the shortest delay comes from the closest distance, and we call this phenomenon the "closest-shortest" rule.

Based on the observations, we develop a simple yet novel IP-geolocation mapping scheme for moderately connected Internet regions, called GeoGet. In GeoGet, we map the targeted client to the geolocation of the landmark that has the shortest delay. We take a large number of webservers with wide coverage and known geolocations as passive landmarks, which eliminates the deploying cost of active landmarks. We further use JavaScript at targeted clients to generate HTTP/Get probing for delay measurement, eliminating the need to install client-side software. To control the measurement cost, we step-by-step refine the geolocation of a targeted

client, down to city level. In practice, GeoGet can be deployed in combination with a certain locality-aware application such that the application can easily obtain the geolocations of their clients.

We implement GeoGet in the moderately connected Internet region we study (China). In the implementation, we collect a large number of webservers and choose about 40,000 of them as passive landmarks, whose geolocations can be accurately obtained. The passive landmarks cover the entire region we are interested. We deploy a coordination server in combination of a website providing video-ondemand (VOD) service, and attract more than 5,000 clients from diverse geolocations to visit and participate during our measurement interval. The evaluation results show that when probing about 100 landmarks, GeoGet accurately maps 35.4 percent targeted clients to city level, which outperforms existing schemes such as GeoLim [7] and GeoPing [5] by 270 and 239 percent, respectively, and the median error distance in terms of city in GeoGet is around 120 km, outperforming GeoLim and GeoPing by about 37 and 70 percent, respectively.

The contributions of this paper are twofold. First, by studying a large data set, we show that most of the traditional IP-Geolocation mapping schemes cannot work well for moderately connected Internet regions, since the linear delay-distance correlation is weak in this kind of Internet regions. Second, based on the measurement results (MR), we develop and implement GeoGet, which uses the closest-shortest rule and works much better than traditional schemes in moderately connected Internet regions. We acknowledge that we are not the first to apply the closestshortest rule and the mapping accuracy of GeoGet is still not very high. However, we go a large step toward developing a better IP-Geolocation system for moderately connected Internet regions. We believe the accuracy will improve significantly if probing more landmarks.

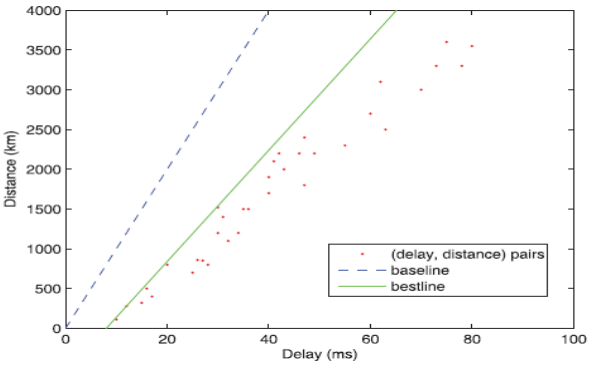
II. Related Work

Delay-measurement approach. Various schemes have been proposed for IP-geolocation mapping, and most of them take delay-measurement approach [10], [11], [12]. In this approach, there are landmarks with known geolocations, and the networking delays between a targeted client and landmarks are measured. The geolocation of the targeted client is inferred from the measured results. In what follows, we introduce some representative schemes taking this approach, including GeoPing [5], GeoLim [7], TBG [8], and Octant [9].

In GeoPing [5], there are a number of landmarks and probing hosts (in practice, the landmarks and probing hosts are usually overlapped and thus the landmarks are active landmarks). Each probing host uses ICMP probing to measure its delays to a targeted client as well as all the landmarks. As a result, every landmark and the targeted client get a delay vector to all the probing hosts. Then, the geolocation of the targeted client is mapped to the location of the landmark whose delay vector has the shortest euclidean distance with that of the targeted client. Therefore, the mapping accuracy of GeoPing depends on strong delay-distance correlation, since it maps the similarity of vectors in distance dimension to that in delay dimension.

They find that such strong correlation holds at least for richly connected Internet regions such as North America. But for Internet regions where delay-distance correlation is weak, this mapping between delay dimension and distance dimension will introduce large error.

GeoLim [7] uses distance constrains based on measured delays to geolocalize a targeted client. Each landmark first measures its delays to the other landmarks, and fits a bestline tightly above all the (delay, distance) pairs measured, as shown in Fig. 1. There is also a baseline, which is drawn by the ideal digital transmitting speed in fiber (2/3 of the light speed), and certainly it lies above the bestline. Given



the delay measured from a landmark to the targeted client, the landmark extracts the distance from the delay value based on the bestline, and draws a circle with its own geolocation as the center and the extracted distance as the radius. If all the circles drawn by the landmarks intersect to a region, the centroid of the region is regarded as the geolocation of the targeted client. In fact, GeoLim also assumes a

Fig 1 Illustration of GeoLim

moderate or strong delay-distance correlation. Otherwise, the extracted distance based on the bestline will be overly skewed compared with the actual distance, and consequently the mapping accuracy will degrade.

Katz-Bassett et al. [8] argue that the assumption on strong delay-distance correlation is unreliable when the delay (distance) is large. They propose to use network topology information to improve the mapping accuracy when there is no landmark with short delay (distance), and they call the scheme TBG. With traceroute tool, they first find the routers along the path from a deployed landmark to a targeted client and then use delay measurement to geolocalize the intermediate routers as well as the targeted client. TBG uses the maximum transmission speed of packets in fiber to calculate the distance constraint from the measured delay, and relies on global optimization to minimize the average error distance for the routers and targeted client. However, similar to GeoLim, when the delay-distance correlation is weak, the extracted distance from a measured delay value will be much overestimated. In addition, the global optimization may introduce extra errors for deciding the geolocation of the targeted client in an effort to reduce the errors to geolocalize the intermediate routers.

Wong et. al. [9] bring forward Octant, which maps a targeted client to a geolocation region by use of not only positive constraints (where the targeted client might lie), but also negative constraints (where the targeted client cannot lie). The positive constraints indicate the upper bound distance of the targeted client, while the negative constraints indicate the lower bound distance. They formulate the IP-geolocation mapping problem as one errorminimizing constraint satisfaction, and solve the constraint system geometrically to yield the geolocation of the targeted client. Fig. 2 shows the convex hull to compute the upper bound distance and lower bound distance given the delay from a landmark to the targeted client. However, based on the data set, we study in the Internet regions where delaydistance correlation is weak, the empty lower right region in Fig. 2 does not exist. Octant also depends on delay-distance correlation to get reliable distance constraints from a measured delay.

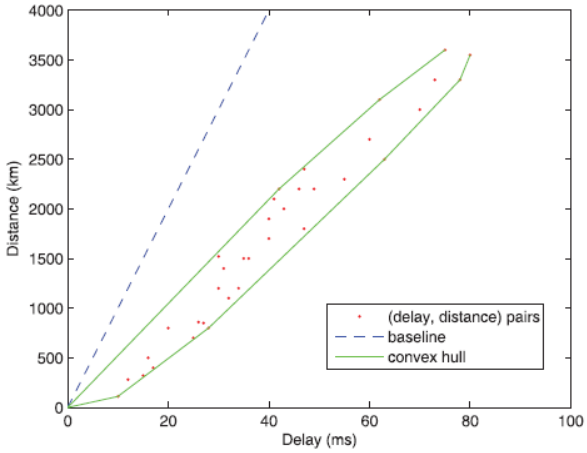
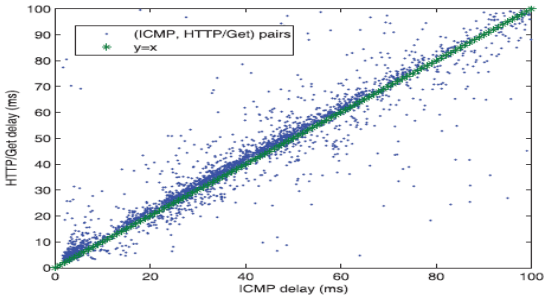


Fig 2 Illustration of Octant

Probably, the most related work with our work is [13]. Though bearing different design goals, the two works take similar approach, i.e., using webservers as landmarks and mapping the geolocation target to the closest landmark. However, our work differs from [13] in two aspects. First, we validate the weak linearity between delay and distance by a large data set from a moderately connected Internet region. Second, study in [13] requires Traceroute to infer the closest landmark but GeoGet uses client-side javascript, because the Ping/Traceroute commands are usually prohibited by many intermediate routers or webservers. In one sentence, our work for the first time focuses on the IPGeolocation accuracy in moderately connected Internet regions, including measuring and mining the real data, validating the hidden relationship between delay and distance, and developing the real system.



III. Proposed Work

A. Design Goals

GeoGet is designed specifically for moderately connected Internet regions, and it has the following design goals:

1. Mapping an IP address to a city-level geolocation with small error distance.
2. No need to install client-side software for delay measurement.
3. Controlling the measurement cost for a targeted client.

B. Using Webservers as Passive Landmarks (LM)

Based on the analysis in the previous section, the closestshortest rule is more applicable than delay-distance correlation for moderately-connected Internet regions. Therefore, in GeoGet, we map a targeted client to the same city as the landmark which has the shortest delay. To cover targeted clients from diverse geolocations, GeoGet requires landmarks in all possible cities. In addition, as shown in Section 3, if we have more landmarks to probe, the closest-shortest rule holds better, and thus the mapping result will be more accurate. For this reason, it is desirable that we have multiple landmarks in a city. More landmarks will bring additional advantages too. First, the measurement load can be shared among landmarks; Second, the single-point failure can be avoided. However, it is very difficult to actively deploy such a large number of landmarks with wide coverage. Our solution in GeoGet is to use webservers as passive landmarks. Given the popularity of web applications, there are a large number of webservers and their geolocations cover almost every city. Using webservers as passive landmarks totally eliminates the deployment and maintenance costs for active landmarks.

C. HTTP/Get Probing Using JavaScript at Targeted Clients

Since we use webservers as passive landmarks, the delay probing needs to be initiated from the client side. To avoid installing any client-side software, we use JavaScript to generate HTTP/Get probing at the targeted clients to measure the delays to the selected webserver landmarks. The JavaScript is stored at a webserver that a locality-aware application employs. When a client uses this service, it will automatically download and execute the JavaScript. The only requirement for the clients is that they have web browser installed and the browser supports JavaScript. The requirement can be easily met by all the desktop and laptop computers to date.

Fig 3 Delay Comparison between ICMP probing and HTTP/GET probing

When executing the JavaScript code, the targeted client visits a nonexisting image in a certain webserver by HTTP/ Get request and records the delay. The HTTP/Get request is sent multiple times and the minimum delay is assumed as the measured delay to the webserver. To bypass the possible web caches, each time the targeted client request for different nonexisting images.

We should make sure that networking delay is the dominant part for the delay measured by HTTP/Get probing. In other words, the server processing delay for HTTP/Get request should be quite small compared with networking delay.

To verify this, we have compared HTTP/Get probing with ICMP probing, by measuring the delays to the same set of webservers. Each webserver was probed 10 times, and the minimum value was chosen as the measured delay. We totally measured 8,000 webservers. Almost all the webservers responded to the HTTP/Get probing, but only 2,876 webservers responded to ICMP probings. It validates our arguments that ICMP probing is prohibited in many routers and servers. Fig. 3 shows the (ICMP, HTTP/Get) delay airs, each for a webserver.

D. Landmark Selection (LMS)

Given so many landmarks in GeoGet, the measurement cost is too high if a targeted client is to probe all landmarks. To control the measurement cost, it is desirable if we can select a subset of all the landmarks for a targeted client.

We adopt a two-step probing method to refine the geolocation of a targeted client. The first step is area-level probing, and the second step is city-level probing. All cities in the entire region are separated to a few numbers of areas according to their geolocations, and there is a center city in each area. In area-level probing, a number of landmarks from the center cities are selected for the targeted client. A controlled number of areas with shortest delays after area-level probing are chosen to enter city-level probing, in which the landmarks from each city of the chosen areas are selected. In this way, a targeted client does not need to probe landmarks from all cities.

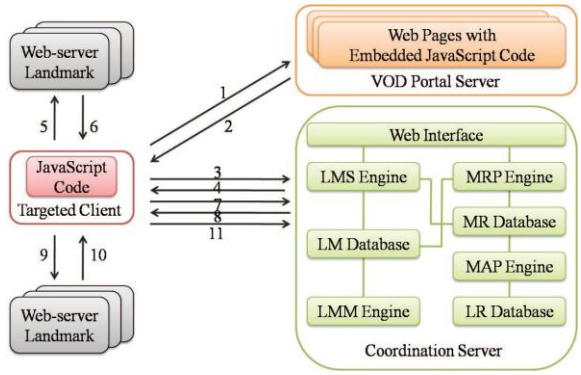


Fig 4 Implementation of GeoGet

IV. Conclusion

In this paper, we explore the delay-distance relationship in China, which are the world's largest country in the number of Internet users and the second largest in the size of IP address space. We find that the linearity between delay and distance is positive but very weak. However, the closestshortest rule holds with high probability. For IP-Geolocation mapping in moderately connected Internet regions, we develop GeoGet. GeoGet adopts closest-shortest rule as the mapping principle, and does not depend on delaydistance correlation as prior work. GeoGet takes use of a large number of webservers as passive landmarks. Java- Script code is embedded in webpages of locality-aware applications for clients to execute when visiting the site. The delay measurement can thus be carried on at targeted clients using HTTP/Get probing generated by JavaScript, without any client-side software installation. Further, we adopt a two-step probing method to refine the geolocation of a targeted client, first to area-level and then to city-level. We have implemented GeoGet, and the evaluation results shows that the mapping accuracy of GeoGet significantly outperforms traditional IP-Geolocation schemes such as GeoLim and GeoPing.

References

[1] H. Xie et al., "P4P: Provider Portal for (P2P) Applications," Proc.ACM SIGCOMM '08, 2008.

[4] K. Xu et al., "LBMP: A Logarithm-Barrier-Based Multipath Protocol for Internet Traffic Management," IEEE Trans. Parallel and Distributed Systems, vol. 22, no. 3, pp. 476-488, Mar. 2011.

[5] V. Padmanabhan and L. Subramanian, "An Investigation of Geographic Mapping Techniques for Internet Hosts," Proc. ACM SIGCOMM '01, 2001.

[6] A. Ziviani et al. "Improving the Accuracy of Measurement-Based Geographic Location of Internet Hosts," Computer Networks, vol. 47, no. 4, pp. 503-523, 2005.

[7] B. Gueye et al., "Constraint-Based Geolocation of Internet Hosts," Proc. ACM Internet Measurement Conf. (IMC '04), 2004.

[8] E. Katz-Bassett et al. "Towards IP Geolocation Using Delay and Topology Measurements," Proc. ACM Internet Measurement Conf. (IMC '06), 2006.

[9] B. Wong, I. Stoyanov, and E. Sirer, "Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts," Proc. USENIX Conf. Networked Systems Design and Implementation (NSDI '07), 2007.

[10] M. Arif, S. Karunasekera, and S. Kulkarni, "GeoWeight: Internet Host Geolocation Based on a Probability Model for Latency Measurements," Proc. 33rd Australasian Conf. Computer Science (ACSC '10), 2010.

[11] B. Gueye, S. Uhlig, and S. Fdida, "Investigating the Imprecision of IP Block-Based Geolocation," Proc. Int'l Conf. Passive and Active Network Measurement (PAM '07), 2007.

[12] B. Gueye et al., "Leveraging Buffering Delay Estimation for Geolocation of Internet Hosts," Proc. Int'l IFIP-TC6 Conf. Networking Technologies, Services, and Protocols (Networking '06), 2006.

[13] Y. Wang et al., "Towards Street-Level Client-Independent IP Geolocation," Proc. USENIX Conf. Networked Systems Design and Implementation (NSDI '11), 2011.