Mapping of Large Internet Regions using IP-Geo Location Scheme

D.V. Rajendrababu  
M.Tech, Department Of Computer Science  
QIS College of Engineering and Technology, Vengamukkapalem, Ongole – Prakasam

A. Srinivasa Reddy  
M.Tech, (Ph.D), Associate Professor, Department Of Computer Science  
QIS College of Engineering and Technology, Vengamukkapalem, Ongole – Prakasam

Abstract-  
The Global Positioning System plays a vital role in mapping any part or any place in the world. In spite of these services there is flaws, in which some of the locations are not identified which are moderately connected to the spatial regions. Know geographical and locations becoming necessity more than facility nowadays. Most IP-geolocation mapping schemes 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 large region of moderately connected Internet and find the delay-distance correlation is weak. But we discover a probable rule—with high probability the shortest delay comes from the closest distance. Based on closest-shortest rule, we develop a simple and novel IP-geo location mapping scheme for moderately connected Internet regions, called Geo Get. In Geo Get, we take a large number of web servers as passive landmarks and map a targeted client to the geo location 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 geo location of a targeted client, finally to city level. The evaluation results show that when probing about 100 landmarks, Geo Get correctly maps 35.4 percent clients to city level, which outperforms current schemes such as Geo Lim and Geo Ping by 270 and 239 percent, respectively, and the median error distance in Geo Get is around 120 km, outperforming Geo Lim and Geo Ping by 37 and 70 percent, respectively.

I. Introduction  
The problem of accurately locating the geographical location of an Internet host, referred to as host geo location, has many useful practical applications. Some examples of such location-aware applications include showing targeted advertisements on web sites, automatic selection of a language to display web site content, web content delivery based on region, credit card fraud detection, and load Balancing and resource allocation between Internet hosts. Today’s applications will benefit from or be enabled by knowing the geographical locations (or geo locations) of Internet hosts. Such locality-aware applications include local weather forecast, the choice of language to display on web pages, 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 [10]. Traditional IP-geo location mapping schemes [5], [6], are primarily delay-measurement based on the targeted client. Traditional IP-geo location mapping schemes are primarily delay-measurement based on the targeted client.

The delays from a targeted client to the landmarks are measured, and the targeted client is mapped to a geo location 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 land mark. The strong correlation has been verified in some regions of the Internet, such as North America and Western
Europe [3], [4]. But as pointed out in the literature [4], 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-geo location mapping scheme for moderately connected Internet regions, called Geo Get. In Geo Get, we map the targeted client to the geo location of the landmark that has the shortest delay. We take a large number of web servers with wide coverage and known geo locations 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 are targeted. client, down to city level. In practice, Geo Get can be deployed in combination with a certain locality-aware application such that the application can easily obtain the geo locations of their clients. We implement Geo Get in the moderately connected Internet region we study (China). In the implementation, we collect a large number of web servers and choose about 40,000 of them as passive landmarks, whose geo locations 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-on-demand (VOD) service, and attract more than 5,000 clients from diverse geo locations to visit and participate during our measurement interval. The evaluation results show that when probing about 100 landmarks, Geo Get accurately maps 35.4 percent targeted clients to city level, which outperforms existing schemes such as Geo Lim [5] and Geo Ping [3] by 270 and 239 percent, respectively, and the median error distance in terms of city in Geo Get is around 120 km, outperforming Geo Lim and Geo Ping 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-Geo location 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 Geo Get, 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 closest shortest rule and the mapping of the particular geo location maintains the accuracy of Geo Get is still not very high. However, we go a large step toward developing a better IP-Geo location 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-geo location mapping, and most of them take delay-measurement approach [8]. In this approach, there are landmarks with known geo locations, and the networking delays between a targeted client and landmarks are measured. The geo location of the targeted client is inferred from the measured results. In what follows, we introduce some representative schemes taking this approach, including Geo Ping [3], Geo Lim [5], TBG [6], and Octant [7].

In Geo Ping [3], 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 geo location 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 Geo Ping
depends on strong delay-distance correlation, since it maps the similarity of vectors in distance dimension to that in delay dimensions the similarity of vectors in distance dimension to that in delay dimensions here provided are where delay-distance correlation is weak, this mapping between delay dimension and distance dimension will introduce large error.

Using location hints from the DNS names of the routers along the path, the locations of the routers are determined. Of the routers whose locations are known, the closest one to the target is selected, and its location is chosen as the target location. The accuracy of the technique depends on the distance from the target to the nearest router of known location. Next, Geo Ping works on the assumption that hosts that are geographically close have similar network delays with respect to other fixed hosts. By comparing the ping times to the target from a set of landmarks or probe machines with the ping times to a set of nodes at known locations, Geo Ping estimates the target location to be the same as that of the node with known location having the most similar ping values. The approach, Geo Cluster, is a database lookup technique which groups IP addresses to clusters based on geographical proximity. This information is combined with the user registration records from web based services such as e-mail services. This technique suffers from the general problems related to database lookup-based approaches, such as reliability, scalability and maintainability issues and it uses data sources unavailable for public access. Recent data-mining based approach [5] is similar to Geo Cluster except that it uses publicly available web pages instead of Proprietary data sources in order to extract geo location. They find that such strong correlation holds at least for richly connected Internet regions as North America. But for Internet regions are associated with the addresses are associated with their IP addresses. Then, this mapping information goes through multi-stage inference processes in order to improve the accuracy and coverage of its IP geo location repository of different IP segments. Finally, those IP segments that are not covered in the first two steps, are mapped with the location of the access router with the help of trace route tool. The accuracy of Structon implement the Internet depends heavily on the accuracy of extracted geographical mapping information. Moreover, with Structon [5], it is harder to get accuracy more than in the granularity of city level. Constraint-Based Geo location (CBG) uses ping times from landmarks as a measure of latency. For each landmark a maximum distance bound for a given latency is derived using distance-to-latency relationships observed between landmarks. During geo location the observed latencies from landmarks to the target are used to draw circles centered at each landmark based on the maximum distance bounds derived earlier. The target is assumed to reside in the convex region resulting from the intersection of circles, and the target location is estimate das the cancroids of this convex region. This technique requires the target to be geographically well surrounded by landmarks. Similar to CBG, Topology-based Geo location (TBG) [2] computes the possible location of the target as a convex region. In TBG, the maximum distance bound is obtained based on the maximum transmission speed of packets in fiber which gives a conservative estimate of the possible region. This region is further refined using inter-router latencies along the path from the target to the landmark, obtained from the trace route command. The final target location is obtained required geo location is from through a global optimization that minimizes average position error for the target and the routers. A more recently proposed measurement-based technique for geo location is Octant [3]. In contrast to other constraint based approaches that only limit the area where the target may be located, Octant also identifies areas where the target may not be located based on observed latencies (referred to as negative constraint). Octant expresses such information by considering two circles corresponding to the maximum and minimum distances from each landmark to the target which constrains the possible geographical area where the target may be located. Each landmark fits a convex hull to all of its delay-to-distance data points with other landmarks. Upper and lower facets of the convex hull correspond to the maximum and minimum distance bounds. Different weights are assigned to different geographical areas based on the number of intersections (higher weights assigned to larger numbers of intersections). The final estimated
region is the union of all regions, where the weight exceeds a desired weight or the region size exceeds a selected threshold. A Monte-Carlo algorithm is applied to pick the best single point location from the final estimated regions. These estimated regions in Octant often end up being disconnected parts. Octant uses geographical and demographical constraints to improve the localization accuracy beyond its measurement-only solution.

III. Client-dependent IP-geolocation systems

GPS-based geo location Global Positioning System (GPS) devices, that have been embedded into billions of mobile phones and computers at now a days, could precisely providing User location where it is exactly in the GPS technology differs from our Geo location strategy in the sense that it is a ‘client-side’ geo location and the approach, which means that the server does not now where the location has user is, unless the user explicitly reports his information back to the server. Cell tower and Wi-Fi-based geo location. Google My Location and Skyhook introduced their cell tower-based and Wi-Fi –based geo location approaches. In particular, the cell tower based geo location offers users estimated locations by triangulating from cell towers surrounding users, while the Wi-Fi-based geo location uses Wi-Fi access point information instead of cell towers. Specifically, every tower or Wi-Fi access point has a unique.

Identification and footprint. To find a user’s approximate location, such methods calculate user’s position relative to the unique identifications and footprints of nearby cell towers or Wi-Fi access points. However, there are many devices (IPs) bound with wired network on the Internet. Such wireless geo location methods are necessarily incapable of geo locating these IPs, while on the other hand to our method does not require any precondition on the end devices and IPs.

IV. Delay-Distance Relationship

There are many researches and discussions on delay distance relationship. To find that delay-distance correlation is strong within North America and within Western Europe, but weak for the entire Internet as a whole. As presented in the section above, most previous work on IP-geo location mapping is based on the assumption of a strong delay-distance correlation.

In this section, we investigate the delay-distance relationship from a large data set collected in a particular region of the particular locations in the the Internet, China, which is the world’s largest country in terms of the the number of Internet users and the second largest in terms of the size of IP address space. To be consistent with prior work, we use round-trip delay (RT Delay) as the delay measurement in this paper.

Data Set

Our data set is composed of (delay, distance) pairs collected between 240 probing hosts and 6,000 webserver landmarks. Each (delay, distance) pair is unique for a (probing host, landmark) pair. The probing hosts and landmarks come from diverse geo locations in China. The values of distance and delay are obtained as follows: First, based on the known geo locations of the probing hosts and landmarks, we calculate the geographical distances of the given (probing host, landmark) pairs using Vincenty’s formula, which assumes that the figure of the Earth is an oblate spheroid, and accordingly is more accurate than models assuming a spherical Earth. Second, the delays of (probing host, landmark) pairs are measured using HTTP request.

Landmark Selection (LMS)

Given so many landmarks in Geo Get, 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 geo location 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 geo locations, 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.
V. CONCLUSION
In this paper, we explore the delay-distance relationship in China, which are the world’s argest 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 closest shortest rule holds with high probability. For IP-Geo location mapping in moderately connected Internet regions, we develop Geo Get. Geo Get adopts closest-shortest rule as the mapping principle, and does not depend on delay distance correlation as prior work. Geo Get takes use of a large number of web servers as passive landmarks. Java-Script code is embedded in web pages 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 geo location of a targeted client, first to area-level and then to city-level. We have implemented Geo Get, and the evaluation results shows that the mapping accuracy of Geo Get significantly outperforms traditional IP-Geo location schemes such as Geo Lim and Geo Ping.

REFERENCES


