IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.
www.ijiset.com
ISSN 2348 – 7968

A Conceptual Approach to Compute the Geolocation of IP Addresses at the National Level Based on Machine Learning

Ebot Ebot Enaw[1],Djoursoubo Pagou Prosper[2]

[1]Department of Computer Science, National Advanced School of Engineering,

Yaounde, P.O Box 8390, Cameroon,

[2]Department of Computer Science, National Advanced School of Engineering,

Yaounde, P.O Box 8390, Cameroon,

Abstract

Over the past couple of years, our society has become more dependent on Internet making it a major driver of economy growth. However, with the development of the Internet, new threats have emerged in the cyberspace called cybercrimes which cause severe security issues for companies, governments and even individuals. In order to prosecute cybercriminals, law enforcement have to identify and geolocate IP addresses, at the origin of the crimes.

Many approaches have been developed to compute the geolocation of an IP address however their precision is not sharp enough especially in environments where NAT (Network Adress Translation) is used.

Therefore, in a bid to develop an efficient system to compute the geographic location of an IP address with a view to locating cybercriminals, in this paper, we designed an approach based on Geographic Information System and artificial intelligence's concepts like anomaly detection, inference and machine learning.

Keywords*: IP geolocation, GIS, machine learning, expert system.*

1. Introduction

With the rapid growth of the Internet, and its ubiquity in every aspects of our daily life, mastering the geographical distribution as well as the trends of Internet development is a major challenge for governments as well as Internet Service Providers for economic, strategic and security purposes. Over the past years, with the increasing interest in location-aware applications, some algorithms and approaches have been developed to calculate the geographic position of an IP address with a good precision. So far, these solutions have not been successful in providing the geographic position of IP addresses in African countries since ISPs use NAT (Network Address Translation) to allocate a single public IP address to an entire city for scarcity purposes.

In a bid to address and tackle these shortcomings, new approaches need to be developed to compute the geolocation of an IP with a certain accuracy in an environment where many users spread across a wide region, share the same public IP address through NAT. These approaches most certainly need intelligent algorithms that can adapt to different situations, evolve, infer on data and learn from past experiences.

The aim of this paper is to present a conceptual approach to compute the geolocation of an IP address with a good accuracy and infer the geographical distribution of Internet users across a country in an effort to facilitate the tracking of cybercriminals.

2. Related work

Some research has been done on topics related to this issue namely [1] that presents existing IP geolocation methods with their advantages and shortcomings. It classifies the methods into two main categories: delay based and topology-aware. Delay based methods encompass all algorithms that estimate the distance between two IP addresses based on the delay of echo packets between the IP addresses. Methods that fall in this category include GeoPing and CBG. Topology aware methods address the issue of circuitous path that hinder delay-based methods by reconstructing the entire network and identifying intermediate routers and their link delay so as to provide a "better estimate" of the distance to the target IP address. Methods that fall in this category include TBG and OCTANT.

[2] presents the Structon approach for IP geolocation. This approach is made up of three main steps: the first step consists of extracting geographic information (ZIP code, city, telephone area code) from webpages in order to get possible web server locations using information clustering. The second step consists of leveraging information from WHOIS and BGP routing table and developing several heuristics so as to improve the coverage and accuracy of their method. In the third step traceroute data is used to improve the coverage of their database. By mining 500 million web pages collected in china, they observed that the accuracy of their algorithm was around 87.4% at city level and 93.5% at province level.

[3] describes the topology based geolocation (TBG) method. This method which is based on landmarks

ISSN 2348 – 7968

disseminated around the world, consists of first determining the network topology (intermediate routers and delays) using traceroute issued from landmarks, then using end-to-end delay, infers per-hop latency and topology, and finally computing the geolocation of the target with a constraint-based optimization technique.

[4] presents a geolocation approach based on a naïve Bayes estimation method. This approach consists first of designing a model to compute the geolocation of an IP address based on parameter such as the hop count and the delay from landmarks to the target and then train the model with some measures including hop count and latency from landmark to targets with known geolocation in order to improve the model and after the training, use that machine learning to compute the geolocation of a target IP address.

[5] presents a system for assessing the geolocation of Internet users in a CDMA network. Their approach consists of identifying all the base stations with which the target mobile station is communicating as well as the characteristics (delay time) of the links between the mobile station and the base stations, then based on a mathematical equation identify three base stations among them and finally apply the TDOA (Time Difference of Arrival) method with these base stations to compute the geolocation of the mobile station in question.

3. Research problem

Given the ever-growing impact of the Internet on the global economy and the surge in the development of location-aware applications, IP geolocation algorithms with very good precision are more than ever indispensable in critical situations including; the mastery of the geographical distribution of Internet penetration rates as well as the location of cybercriminals. Some geolocation approaches have been developed namely CBG and TBG but they fail to compute IP geolocation with a good precision especially in regions where ISPs assign a single IP address to clients disseminated in a wide region through NAT. In this specific context these approaches will provide as output an estimation of the geographic location of the router that operates the NAT (which is usually located within the ISP's facilities) instead of that of the user. This might be explained by the fact that these approaches address the problem of IP geolocation in a global/worldwide perspective and thus provide results that are accurate at a country or city level. However this level of accuracy is neither enough for law enforcement to locate a cybercriminal nor for government to evaluate the geographic dispersion of Internet users across the country. In order to address this issue, in this paper, we propose a novel approach based on artificial intelligence and machine learning that will enable the development of intelligent algorithms that will learn the Internet traffic trends of specific areas in a country so as to compute the geolocation of IP address with a good accuracy.

4. ISP network and IP geolocation

As depicted in the figure below, the ISP network is usually made up of three main layers:

- Core: This layer serves as the backbone of the network and it provides the interconnection with other ISPs using BGP protocol. Therefore it should provide high bandwidth link with redundancy.
- Distribution: This layer is the central point of the architecture as it aggregates all end-user connections, process them before forwarding them to the core layer. The types of processing this layer provides include Qos policy enforcement, access control, authorization, accounting, application control, routing, traffic filtering and address translation ;
- Access: This layer connects end users terminals to the network. Usually ISP setup Point of Presence (POP) which is constituted of switches and base stations disseminated across different coverage zone to provide access to end-users terminals.

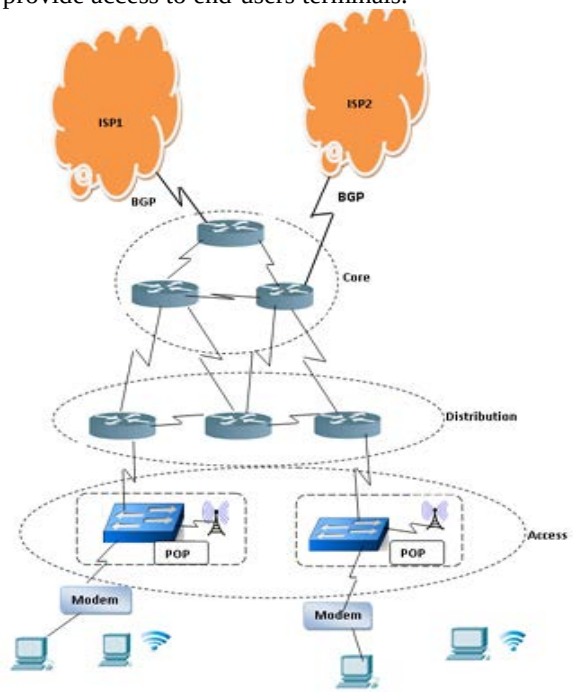

Figure 1: Typical ISP architecture

Usually, when a client attempts to connect, his request is directed to the AAA server that authenticates its credentials, authorizes the access to services the user has subscribed to and assigns him an IP address that might be private or public. In case of a private IP address, the NAT server that is usually a router located in the distribution layer performs the translation of the private IP address into a public one.

Traditional approaches of IP geolocation proceed as follows: landmarks servers disseminated across the globe issue pings to IP addresses of known location and the equations linking the delays to the distance is inferred from the results so that when the geolocation of a new IP address will be requested, these landmarks will issue

ISSN 2348 – 7968

ping to that IP address and compute the distance from the landmarks to that targeted IP address using the equation derived previously, then circles centered at landmarks and with radius equal to the distance calculated previously are drawn and their intersections depict the most probable zone where the target IP address might be located.

In the case of many clients disseminated in a wide area, sharing the same public IP address through NAT, this approach might lead to errors therefore a new approach that analyzes and classify the traffic characteristic is highly needed.

5. Machine learning

In recent years, with the digitization of almost every domain (Healthcare, finance, education, transport, etc) and the development of Internet of things, a huge volume of data are being generated by devices such as sensors and the need to analyze these data is becoming very crucial. Machine learning algorithms enable data to be fetched in order to spot the underlying distribution and trend as well as the relationship between variables so as to predict or infer on them. It finds application in a variety of areas including pattern and speech recognition, fraud detection, market analysis, intrusion detection and malware analysis, etc.

Machine learning is aimed at either building a predictive model to make predictions on the future based on past experiences data or building a descriptive model that can help gain a deep understanding of data trend. To this end, a first version of the model has to be developed then it has to be trained with data so as to improve it. Learning can be of two types supervised and unsupervised. Supervised training consists of feeding the model with input and its corresponding output ; while in the unsupervised training, only input data are provided.

Many approaches of machine learning have been developed namely:

- Bayesian: This approach is based on the bayes formulae P(A/B)= P (B/A) x P(A) / P(B). It consists of first defining a model that expresses our knowledge, specify prior probability of this model's parameters, then gather some sample data and compute the posterior probability distributions for the aforementioned parameters which will be used to make decisions or predictions ;

- Neural network: This technique is inspired by the brain system. A neural network can be defined as a combination of multiple neurons. A neuron combines several input with their respective weight and a bias parameter to deliver an output. During the learning phase the training pair (input and its corresponding output) is fed to the network. The initial values of the connection weights are set randomly to small numbers. The network derives its responses and compares them with the desired ones. If there is an error, the system adjusts (increases or decreases) the weights by a small amount in the direction identified by the predefined learning rule until the error begins to increase or is reduced to an acceptable level. Then, the weights are frozen alone and the training is completed ;

- Clustering based technique: This technique consists of assigning a set of observations to subsets (called clusters) so that observations in the same cluster are similar in some sense. It is based on some algorithms like the K-nearest which consists of grouping objects based on attributes/features into K number of group/clusters by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Though natively this technique falls under the unsupervised learning category, semi-supervised clustering techniques also exist.

6. Our Solution

6.1. Methodology

In an effort to provide an efficient solution for the geolocation of IP addresses with a better accuracy, we propose a new approach based on the methodology described below:

1. Setup a server within the network of every ISP in a city and record their GPS coordinates. These servers will be used as landmarks in our system and in so doing the latency and the hop paths that link them to a target IP address will be used to provide a good approximation of the GPS coordinate of that IP address ;

2. Define the target precision named λ for our system and for every town, divide it in circled zones of radius λ ;

3. Develop a client module that will determine the network hops and the delays of the paths that link the target IP address to the landmarks and upload these data to the application server ;

4. Develop a module called topology-builder that, based on the data provided by the client module, will build the network topology of the network that links the target IP address to the landmarks ;

5. Develop a module that will display the network topology that links the target IP address to the

ISSN 2348 – 7968

landmarks in a map using GIS (Geographic Information System) concepts ;

6. Develop a module called Analyzer that using a machine learning algorithm, will try to infer the geographic position of an IP address based on the network topology and the delays that link the target IP address to the landmarks ;

7. Train the Analyzer module with data gathered through the client module from clients of known geolocation and grouped by zone defined in step 2, so as to improve the model used by the analyzer module;

8. Develop a module that can be embedded in popular websites within a country as well as captive portal of ISP and that will assess the geolocation of every visitor in an effort to compute the geographic distribution of Internet users in the country ;

6.2. Description of the system modules

With regard to the methodology presented in the previous section, our framework is made up of five (05) main components namely client module, topology builder, GIS, Analyzer and Distribution processor which are depicted in the figure below. These modules will be described in subsequent sections.

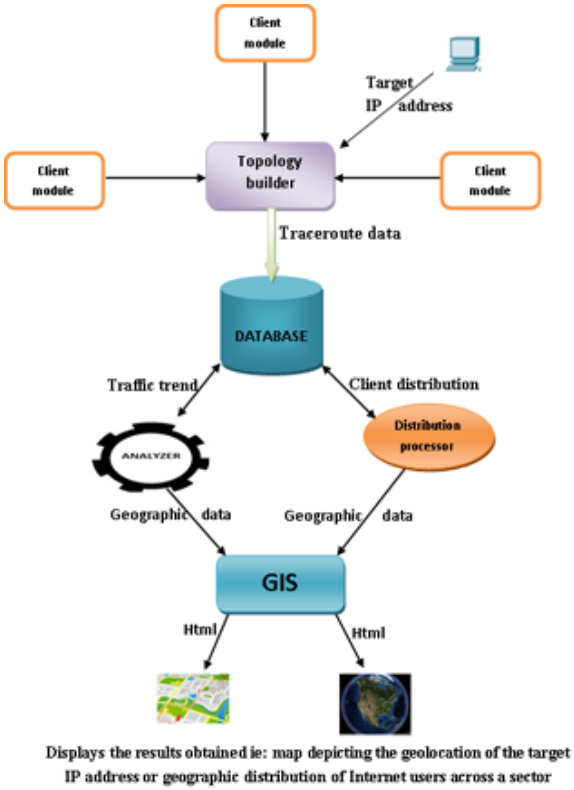

Figure 2: Architecture of our system

6.2.1 Client module

This module is aimed at assessing the topology as well as the delays of the network paths that links an IP address to the landmarks located within the city of the client.

In this light, it carries out the following activities:

- Determine the IP address of the landmarks located in the same city as the client ;
- Issue traceroute command to the landmark's IP addresses identified in the previous step, records the results and sends them to the topology builder module ;
- Request the head of a popular website from a client, record the time needed to process the request and sends it to the application server. This data is particularly useful because usually ISP configure specific policies to restrict icmp traffic, we need to work with one more protocol namely http to spot the real traffic trend ;

In order to carry out these activities especially the usage of traceroute command, the module needs to access system routine. Given that javascript or vbscript cannot access system routines, we decided to implement this module in a java applet. We choose java because it is interoperable with all browsers.

6.2.2 Topology builder

This component collects traceroute results sent by the client module and analyzes them so as to build the network topology that links the target IP address to the different landmarks. Given that the traceroute command provides the IP addresses of the different routers on the path linking the target IP address to the landmarks as well as the delays of each link, this module stores these data in the database and builds a graph with nodes and vertices representing the intermediate routers and their interconnections respectively. The value of the vertices linking two nodes will be the delay of the connection linking these nodes. This module has a web interface through which the administrator of the platform can input the target IP address to geolocate.

Every time this module receives data from a client module, it updates the topology.

6.2.3 GIS

This module which is based on google map provides a graphical representation of the network. It displays the landmarks in their respective locations as well as the

ISSN 2348 – 7968

network topology that links the landmarks to the target IP address. After computing the geolocation of a target IP address, the most probable zone where the target IP address is supposed to be is highlighted. This module also leverages some google map features by providing the itinerary to follow as well as the transport means to use in order to move to the target zone.

6.2.4 Analyzer

This component assesses data collected by the topology-builder module in an effort to provide a good estimate of the geographic location of an IP address. The module is first trained using data gathered through the client module from clients of known geolocation and grouped by zone defined in step 2 of our methodology. Our approach relies on the assumption that the behavior of the network is usually different from one ISP to another, and depends on the area as well as the time. Time is a critical parameter because during the day (8a.m to 5p.m), the connection from residential areas might be fluid because people are away at work or at school as such the network is not saturated whereas in the evening it will not be the case since people are going back home. Geographic area is also an important factor especially for ISPs that offer wireless access to their networks since electromagnetic waves propagation are influenced by obstacles (tree, wall, mountain, etc.), the way they propagate in an area depends on the inherent characteristics of that area. With regard to the aforementioned assumptions, the training phase is carried out per ISP, per zone and per time interval in order to spot the way an ISP's network behaves in a specific zone in a specific timeframe. The training phase is conducted as follows: a captive portal containing the client module is setup in every ISP so that when a client connects to the Internet, the client module applet is run. Data provided by the client modules are sent to the topology module which in turn processes and sends them to the analyzer module. For every ISP and for each timeframe, after this process has been repeated for every zone, using the K-nearest algorithm, the analyzer module without taking into account the zone where each data came from, will then try to identify similar trends in terms of topology and delays and group them in clusters. These clusters will depict the traffic trend. Then subsequent model based on the combination of time series analysis and neural network will be applied to the data in order to spot and predict the network behavior for every ISP in each region and for each time frame.

Based on the assumption that terminals connected to the same ISP and located in the same area share similar traffic trends, when a target IP address is submitted to the system for geolocation, its traffic trend (topology and delay to landmarks) provided by the topology module will be compared with the various clusters identified previously. The target IP address is located in the closest cluster (cluster with traffic trend closest to that of the target IP address):

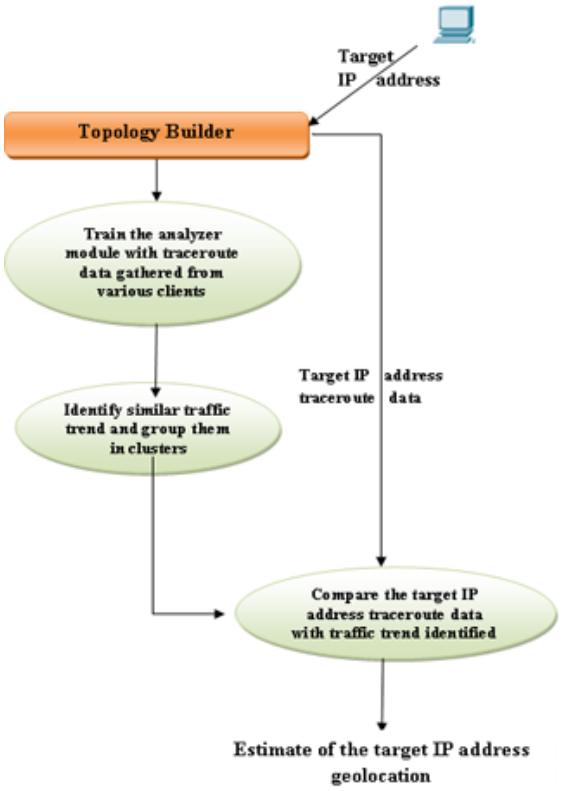

Figure 3: analyzer module

6.2.5 Internet Distribution

This component is responsible for computing the geographical distribution of Internet users across a country. To compute this distribution, the client applet has to be embedded in popular websites so that when clients visit these websites, the applet will be run and automatically it will send data to the topology-builder and subsequently to the analyzer module. The clusters identified by the analyzer module will be matched with the geographic distribution of the population so as to infer the distribution of Internet penetration rate across the country.

7. Conclusion and future work

Due to the surge in cybercrimes and coupled with the fact that location aware applications are growing in popularity, the need for algorithms capable of geolocating IP address with a good precision is more than ever a requirement.

Several approaches of IP geolocation have been developed namely TBG, CBG but they don't take into

ISSN 2348 – 7968

account some parameters including the fact that in some ISP, especially those of developing country, many users disseminated across a wide area, are bound to the same public IP address through NAT. In this context the existing approaches of IP geolocation will output the geolocation of network equipment that operates the NAT instead of the geolocation of the target user.

Therefore, in this paper we propose a new approach that leverages machine learning concepts to analyze topology as well as path delays in order to infer the geolocation of an IP address. Our approach first consists of training the model with some data (network path to landmark, delays to landmarks) gathered from client of known geolocation at specific time frame so as to spot the way the network of a particular ISP behaves at specific time frame in a particular zone. After the behavior of ISP's network has been processed for every zone, the training phase is terminated. In order to calculate the geolocation of an IP address its footprint (delay and network path to the landmarks) will be compared to the set of behaviors gathered during the training phase and its geolocation will be mapped to the zone corresponding to the behavior that resemble it the most. The innovation of this approach is twofold: first the continuous learning process as every time the geolocation of an IP address is requested the model is updated, secondly this approach takes into consideration the time frame which refines the accuracy of our model based on time series analysis.

Future work can include the development of a prototype of a system that will implement this approach so as to evaluate its pragmatism, applicability as well as its computational complexity.

References

[1] Jayaprabha Bendale , Prof. J. Ratanaraj Kumar, "Review of different IP geolocation methods and concepts " in *International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 436-440*, 2014.

[2] Chuanxiong Guo, Yunxin Liu, Wenchao Shen†, Helen J.Wang, Qing Yu, Yongguang Zhang "Mining the web and the Internet for accurate IP address geolocations" *INFOCOM 2009, IEEE, 2841-2845, 2009.*

[3] Ethan KatzBassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, Yatin Chawathe, "Towards IP geolocation using delay and topology measurement" Internet Measurement Conference - IMC , pp. 71-84, 2006.

[4] Brian Eriksson, Paul Barford, Joel Sommers*y*, and Robert Nowak, "A learning based approach for IP geolocation " in Proceedings of 11th international conference on Passive and active measurement, pp. 171-180, 2010.

[5] Ebot Ebot Enaw, Djoursoubo Pagou Prosper, "A system for assessing the geolocation of Internet users in a CDMA-EVDO network", IJACT Volume 3 Issue 4 pp 46-52, 2014

Biography

Dr. EBOT EBOT ENAW obtained his B.Eng hons degree from Liverpool University in Electronic Engineering in 1989. He later obtained an M.Eng degree in Telecommunication Engineering from The University of Manchester England in 1991. He returned home where he was recruited in the University of Yaounde I, as an assistant lecturer. He pursued his university studies and obtained a PhD in Computer Sciences from the National Advanced School of Engineering of the University of Yaounde I, where he is currently a senior lecturer. His area of specialization include: computer network security, cryptography and formal specification and verification; theorem proving and model checking. He has published many research articles in peer-reviewed international journals. In 2006 he was appointed Director General of the National Agency for Information and Communication Technologies Cameroon, a position he occupies till date. Major activities of the agency include amongst others: securing the Cameroon cyberspace through three key services: Computer Incidents Response Team (CIRT), Public Key Infrastructure (PKI) and Computer Security Audits.
Dr.EBOT EBOT ENAW may be reached at ebotenaw@yahoo.com

DJOURSOUBO PAGOU Prosper obtained his Master degree in Computer science engineering from the National Advanced School of Engineering of the University of Yaounde I in 2009. He holds several certifications in networking and cybersecurity namely CCNA, CCNP, CEH, ECSA. In 2013, he was appointed subdirector of the National Computer Incidents Response Team (CIRT) of Cameroon, a position that he occupies till date.