A Feature Clustering-Based IP Localization Algorithm

Yong Gan, Yifan Wang and Dongwei Jia

1 Zhengzhou Institute of Engineering and Technology, Zhengzhou, Henan, China
2 School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, China
Email: ganyong@zzuli.edu.cn; 925053758@qq.com; zzulijdw@163.com

Abstract. The geographic location of network IP is an important foundation of location-based services, which plays an increasingly important role in people’s daily life and work. However, the existing IP location methods based on database query, network measurement and machine learning often have some problems, such as poor timeliness, low reliability, difficult feature modeling and so on, or need a large amount of user data as support, which is often only suitable for the more ideal network environment, so the positioning accuracy and applicability are far from the actual needs. To solve this problem, this paper proposes an IP location method based on clustering. The initial IP address of the classifier is located by training. Combined with the method of IP address database matching, the accurate location of IP address is finally realized. Experimental results show the effectiveness of the method.

Keywords. IP geolocation; clustering; support vector machine.

1. Introduction
With the rapid development of the Internet, network applications have penetrated into all aspects of people’s daily life, work and study. Network entity location is an important foundation in many web applications of location-based services. The process of determining the geographic location of a network entity in the real world is called network entity location [1]. Since network entities can usually be identified by their IPs, the process of network entity localization can often determine the geographic location of a target IP at a certain granularity by the IP address of the network entity [2], and thus can also be referred to as IP localization. IP location technology is widely used in people’s daily life, network security and network management. The location results usually include country name, region name, latitude and longitude, time zone, etc. [3]. In terms of network security, IP positioning technology can be used to determine the location of cyber criminals, trace the source of illegal information spread on the network, effectively stop the rapid spread of network rumors, provide technical support for national anti-terrorism stability, and determine network fraud through IP positioning; in terms of network management, IP positioning technology can effectively prevent the spread of network rumors, and through IP positioning can determine network fault servers, effectively avoid problems such as network server paralysis caused by network congestion, and identify network fraud spam sources through IP location to improve network management performance. In addition, IP location technology can also be used for user access control, responding to or denying requests based on the user’s IP address. IP addresses can also be used to recommend the location of friends in some social networking software. Municipal location is used to determine the location of users based on their IPs and push video content from their webcasting platforms [4-6]. Therefore, the study of IP location technology has important application value.
At present, some researches on IP localization techniques have been conducted at home and abroad. Internationally, some important research institutes, such as the Institute of Information Science at the University of Southern California, USA, have conducted research related to network topology localization. A team led by Steve Uhlig at the University of London worked on the reliability and accuracy of locating databases and published a paper on database localization in ACM SIGCOMM [7]. The Technicolor Research Laboratory at Boston University has conducted a lot of research in machine learning localization and achieved a series of results [8]. Some domestic research institutions, such as Tsinghua University, Beijing University of Posts and Telecommunications, National University of Defense Technology, and Zhengzhou University, have also achieved a series of results in the field of IP localization technology [9-11]. Although some research results have been achieved in the research of IP positioning technology, the precision and accuracy of IP positioning are still difficult to meet the needs of high positioning due to the complex network environment, the inability to accurately measure the network topology, the difficulty to obtain a large number of high-precision landmarks [12], and the inaccuracy of the existing IP geolocation database information. Therefore, further research on IP city-level positioning technology is of great academic value to improve the accuracy and reliability of positioning results.

IP location techniques generally use a given IP address to locate a target node (also known as IP positioning) and are usually implemented based on database queries, network measurements, machine learning and other methods. In IP localization techniques, the following basic elements are usually included.

(1) Node to be located: mainly refers to various types of hosts, servers, etc. in the network.
(2) Probes: hosts or servers that send packets.
(3) Landmarks: hosts or routers whose geographic locations are known, mainly including active and passive landmarks.
(4) Network entity: refers to the gateway, router, etc. in the network.
(5) Network topology: refers to the connection layout structure of the network.
(6) Target network: to be defined as the network in which the node is located.
(7) Time delay: the time required to detect the transmission of messages from one node to another in the network.
(8) Location server: used to send IP query messages [13-14].

The main problems of the existing methods are: (1) the results of the IP location method based on database query are unreliable. The database query-based IP location method mainly relies on the accurate geographic location information in the database to locate the IP. first, since the IP geographic location information in the database can only be accurate to the city level, while some databases can only be accurate to the national level, the location granularity of the method is greatly affected by the granularity of the information in the database [15]; (2) the measurement error of the network delay is large. In the actual network environment, the accuracy of network delay measurement is affected by the presence of conditions such as router queuing delay, network congestion, and delay jitter. In some cases, the network delay cannot even be measured, so the delay-based localization algorithm can be greatly affected [16]. (3) The integrity of the network topology is difficult to guarantee. In topology-based localization algorithms, the blockage of firewalls or the existence of anonymous routing make the localization algorithm unable to detect the paths and thus cannot obtain the complete topological path information [17-20].

To address these shortcomings, this paper proposes a clustering-based IP localization algorithm. The learning-based IP localization algorithm detects the multidimensional features of the target IP. Based on the idea of feature clustering, a support vector machine algorithm is introduced to establish a classification model to improve the prediction accuracy of the target IP geographic location.
2. Classification-Based IP Location Algorithm

2.1. General Introduction of the Method
The title is set 17-point Times Bold, flush left, unjustified. The first letter of the title should be capitalized with the rest in lower case. It should not be indented. Leave 28 mm of space above the title and 10 mm after the title.

The current IP localization algorithms mainly include IP address database, network measurement, and machine learning. In the network measurement process, the target needs to be detected from multiple detection sources. The delay and hop count information obtained during the detection process is used as the specific measurement parameters in the network measurement process. Using the measurement parameters, the geographic location of the target IP is calculated using CBG or SLG algorithms.

To address the problem of low efficiency of IP localization algorithm in network measurement mode, a feature clustering-based IP localization algorithm is proposed in order to make full use of historical measurement data. Finally, based on the idea of multidimensional support vector machine, the target localization accuracy is further improved. The specific workflow of the algorithm is shown in figure 1. For the IP to be located, after inputting it into the classifier to get the IP city-level address, the street-level IP location is achieved by using the street IP address division database within the city for matching.

![Figure 1. Framework of IP positioning algorithm based on support vector machine.](image)

2.2. The Specific Implementation of the Classification Model
Without considering the influence of other complicating factors, if two IPs to be tested are very close to each other geographically, the detection paths from the same detection source are approximately the same and the measured parameters should be very close. In other words, IPs from the same city have a high similarity.

For each IP node, the probe data obtained from different probe sources are used as feature vectors, including delay and hop count information, etc. A known IP set \( S = \{ S_1, S_2, \ldots, S_n \} \), each IP can be \( V = \{ V_1, V_2, \ldots, V_n \} \) represented by vector containing \( m \) feature attributes, among them \( V \) the data obtained from different detection sources. Assume that the IPs in the IP set \( S \) belong to \( k \) different classes (cities), \( S \) can be divided into \( \{ C_1, C_2, \ldots, C_k \} \) a total of \( k \) subsets. Support vector machine based
IP localization algorithm is to use existing IP features and category data to train a classification model to determine the category to which an IP of unknown category belongs based on its feature vector.

Support vector machine (SVM) is a dichotomous model. Its basic idea is to find a hyperplane to segment the sample. The principle of segmentation is to maximize the interval and finally transform it into a convex quadratic programming problem to solve it. By using kernel function technology, nonlinear classification ability can be obtained through sample learning. For nonlinear classification problems, the classification model can be expressed as

\[ f(x) = w^T \phi(x) + b \]

where \( \phi(x) \) represents \( x \) corresponding eigenvectors. The classification model can be solved by minimizing functions:

\[
\min \frac{1}{2} |w|^2, \text{s.t. } y_i(w^T \phi(x_i) + b) \geq 1 (i = 1, 2, \cdots, m)
\]

The dual problem of the above minimization problem is:

\[
\max \sum_{j=1}^{m} a_j - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j \phi(x_i)^T \phi(x_j), \text{s.t. } \sum_{i=1}^{m} a_i y_i = 0, a_i \geq 0
\]

After using Gaussian kernel function to solve the problem:

\[
f(x) = w^T \phi(x) + b = \sum_{i=1}^{m} a_i y_i \phi(x_i) \phi(x_j) + b = \sum_{i=1}^{m} a_i y_i K(x_i, x_j) + b
\]

The Gao Si kernel function can be written as follows:

\[
K(x_i, x_j) = \exp \left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right)
\]

3. Experiment and Results Analysis

The experiment selected 1022 city effective landmarks. Beijing and Shanghai are selected as detection sources to detect flag points and obtain delay and hop information. The measurement process is affected by the state of the network, and it is easy to produce a large number of errors. At the same time, some data are missing, direct use of such data in the prediction model can make stability worse. Therefore, data must be cleaned and filled before model training. For outliers in raw data, the average or multiple numbers in the same class of data can be taken. In this experiment, the outliers of the average time delay and the hops are crowdsourced.

In order to evaluate the influence of the number of detection sources on the positioning performance, based on the above two detection sources, this paper compares the classifier training using one of the two detection sources separately, and evaluates the accuracy. The evaluation results are shown below. It can be seen from the results of the graph that the support vector machine can achieve better performance than other classifiers no matter which probe source. At the same time, the classifier trained with IP data features of two detection sources is better than any single data source.

After cleaning the experimental data, it is divided into training set and test set, in which training set accounts for 70 and test set accounts for 30. on the training set, the algorithm based on naive bayes, decision tree and support vector machine is used to train, three corresponding classifiers are obtained, and the prediction accuracy is evaluated on the test set.

From table 1, we can see that in the IP location algorithm based on feature similarity, the machine learning algorithm based on support vector machine algorithm has high positioning accuracy. The performance of naive Bayes algorithm is the worst. This is because the domestic network belongs to the hierarchical structure, in which the size and distance of the network delay is not a simple linear relationship. Naive Bayesian algorithm is more suitable for simple network architecture, but it cannot
deal with complex hierarchy. The decision tree algorithm often performs well on the training data, but the actual effect on the test data is generally under the influence of fitting factors.

### Table 1. Accuracy of algorithms.

| Algorithm         | Accuracy |
|-------------------|----------|
| NBM               | 47.1%    |
| Decision Tree Model| 80.4%    |
| SVM               | 86.9%    |

The confusion matrix obtained by the SVM algorithm and the location of each city in the database is as follows. The confusion matrix can most intuitively respond to the situation of classification. The right side shows the corresponding division code of each letter. The number of classified correctly on diagonal line is higher in figure 2.

![SVM confusion matrix of classification in Henan Province by IP database.](image)

### 4. Conclusion

A street landmark acquisition method based on feature clustering based SVM classifier is proposed to make up for the shortcomings of structural structure. This method applies SVM classifier to server IP classification and realizes the association between server IP and organization name. The experimental results show that this method can obtain more reliable street horizontal landmarks in a short time. The median error of geographical location is reduced by using the landmarks added in geographical location. Finding a better classifier to classify IP, finding a better way to realize the link between domain name and agency name is the focus of future work. Proposed a feature clustering-based IP city-level localization algorithm that makes full use of the potential value of historical IP detection data. The IP localization algorithm based on feature clustering bypasses the tedious process of calculating population density one-dimensional distribution based on machine learning IP localization algorithm. According to the idea of feature clustering, support vector machine algorithm is introduced to establish classification model. Finally, the accuracy and inverse fitting ability of target IP position prediction are improved by using address matching method.

### Acknowledgments

This work is supported by the Nation Nature Science Foundation of China (NSFC), (NO. 61572445).
References
[1] Muir J A and Oorschot P C V 2009 Internet geolocation: evasion and counter evasion ACM Computing Surveys (CSUR’09) 42 (1) 4.
[2] Boyes H, Hallaq B, Cunningham J and Watson T 2018 The industrial internet of things (IIoT): An analysis framework Computers in Industry 101 1-12.
[3] Seo H, Lee K, Yasukawa S, Peng Y and Sartori P 2016 LTE evolution for vehicle-to-everything services IEEE Communications Magazine 54 (6) 22-28.
[4] Costa C, Pasluosta C, Eskofier B, Silva D and Righi R 2018 Internet of Health Things: Toward intelligent vital signs monitoring in hospital wards Artificial Intelligence in Medicine 89 61-69.
[5] Engineer A, Sternberg E M and Najafi B 2018 Designing interiors to mitigate physical and cognitive deficits related to aging and to promote longevity in older adults: A review Gerontology 64 (6) 612-622.
[6] Li Q, Feng X, Wang H and Sun L 2018 Understanding the usage of industrial control system devices on the internet IEEE Internet of Things Journal 5 (3) 2178-2189.
[7] Gill P, Ganjali Y, Wong W and Lie D 2010 Dude, where’s that IP?: Circumventing measurement-based IP geolocation Proc. Usenix Secur. Symp. Washington (DC, USA) p 16.
[8] Katz-Bassett E, John J P, Krishnamurthy A, Wetherall D and Anderson T 2006 Towards IP geolocation using delay and topology measurements Proc. 6th ACM SIGCOMM Conf. Internet Meas. 71-84.
[9] Chen J, Liu F, Luo X, Zhao F and Zhu G 2016 A landmark calibration based IP geolocation approach EURASIP J. Inf. Secur. 4 1-11.
[10] Liu W, Luo X, Liu Y, Liu J, Liu M, et al. 2018 Localization algorithm of indoor Wi-Fi access points based on signal strength relative relationship and region division Computers, Materials & Continua 55 (1) 71-93.
[11] Ma Y, Luo X, Li X, Bao Z and Zhang Y 2018 Selection of rich model steganalysis features based on decision rough set α-positive region reduction IEEE Transactions on Circuits and Systems for Video Technology 1.
[12] Zhu G, Luo X, Liu F and Chen J 2015 An algorithm of city-level landmark mining based on internet forum 18th IEEE International Conference on Network-Based Information Systems pp 294-301.
[13] Zhang Y, Qin C, Zhang W, Liu F and Luo X 2018 On the fault-tolerant performance for a class of robust image steganography Signal Processing 146 99-111.
[14] Li R, Sun Y, Hu J, Ma T and Luo X 2018 Street-level landmark evaluation based on nearest routers Journal on Security & Communication Networks 2018 (2) 1-12.
[15] Gueye B, Ziviani A, Crovella M and Fdida S 2006 Constraint-Based Geolocation of Internet Hosts IEEE/ACM trans. Networking 14 (6).
[16] Jiang H, Liu Y and Matthews J N 2006 IP geolocation estimation using neural networks with stable landmarks Proceedings of the IEEE International Conference on Computer Communications Workshops pp 170-175.
[17] Denisiko D and Hoffman M M 2018 Classification and interaction in random forests Proceedings of the National Academy of Sciences pp 1690-1692.
[18] Dietterich T G 2000 Ensemble methods in machine learning Proc International Workshop on Multiple Classifier Systems pp 1-15.
[19] Breiman L 2001 Random forests Machine Learning 45 (1) 5-32.
[20] Lu F, Shi Z, Gu L, et al. 2019 An adaptive multi-level caching strategy for distributed database system Future Generation Computer Systems 97 61-68.