# Improving IP Geolocation using Query Logs

Ovidiu Dan[1,2], Vaibhav Parikh[2] and Brian D. Davison[1]

[1]Lehigh University, Bethlehem, PA, USA, {ovd209,davison}@cse.lehigh.edu

[2]Microsoft Bing, Redmond, WA, USA, vparikh@microsoft.com

## ABSTRACT

IP geolocation databases map IP addresses to their geographical locations. These databases are important for several applications such as local search engine relevance, credit card fraud protection, geotargetted advertising, and online content delivery. While they are the most popular method of geolocation, they can have low accuracy at the city level. In this paper we evaluate and improve IP geolocation databases using data collected from search engine logs. We generate a large ground-truth dataset using real time global positioning data extracted from search engine logs. We show that incorrect geolocation information can have a negative impact on implicit user metrics. Using the dataset we measure the accuracy of three state-of-the-art commercial IP geolocation databases. We then introduce a technique to improve existing geolocation databases by mining explicit locations from query logs. We show significant accuracy gains in 44 to 49 out of the top 50 countries, depending on the IP geolocation database. Finally, we validate the approach with a large scale A/B experiment that shows improvements in several user metrics.

## Categories and Subject Descriptors

H.3.m [**Information Search and Retrieval**]: Miscellaneous

## General Terms

Measurement, Experimentation, Algorithms

## Keywords

IP geolocation; geographic targeting; geotargeting; local search; contextual relevance; geographic personalization

## 1. INTRODUCTION

IP geolocation databases are widely used in online services to determine the location of users when real time location information is not available. Web search engines are an example of services which rely on this type of information. Search engines use the geographical location of users to personalize the results shown on the page. For instance, for the query "weather" search engines may display a page block with the local weather forecast based on the location of the user. Outside of the realm of search engines, IP geolocation databases are used for applications such as:

- **Content Delivery Networks** [21]: IP geolocation databases help direct users to the closest server, improving latency, throughput, and bandwidth costs.
- **Credit card fraud protection** [9]: Financial institutions use the location of online users in their fraud detection algorithms.
- **Advertising networks** [24]: IP geolocation of users can lead to better ad targeting and higher revenue.
- **Law enforcement** [27]: Locating IP addresses can help law enforcement in cybercrime investigations.
- **Location based content licensing** [25]: Streaming services such as Spotify and Netflix make use of IP geolocation to enforce geographic content restrictions.
- **E-commerce** [28]: Shopping Websites can use geolocation to apply correct taxes and shipping charges.
- **Organizations with regional offices** can use IP geolocation to direct users to their closest facility.

Although IP geolocation databases have much higher real-life coverage than methods such as Wi-Fi localization [11], they can still suffer from inaccurate information and low coverage in certain parts of the world. Accurate IP geolocation can make the difference between satisfied and dissatisfied customers. In the case of search engines, incorrect location data can lead to abandoned searches and engine switches. In other cases, such as credit card fraud protection, imprecise location information can lead to loss of revenue. In this paper, we present methods for measuring IP geolocation databases and improving their accuracy **at the city level** using query logs. More specifically, our contributions are:

1. **We describe a technique to generate IP geolocation ground truth data using real time location information.** We construct a dataset of 8.4 million IP addresses with known geographical location. The dataset is derived from a subset of search engine logs that contain real time global positioning information obtained from mobile devices. To the best of our knowledge this dataset is the largest ever reported in literature. *(Section 4)*
2. **We determine the impact of incorrect IP geolocation information on implicit user behavior.** We show that inaccurate geolocation information has a negative effect on users of search engines. Metrics such as click through rate, click success, and ad revenue are impacted. *(Section 5)*
3. **We evaluate the accuracy of three IP geolocation databases.** We measure the accuracy and the distance error of three major commercial IP geolocation databases. There are considerable differences between databases, with the best database having an

*WSDM'16,* February 22–25, 2016, San Francisco, CA, USA.

ACM 978-1-4503-3716-8/16/02 ...$15.00.
DOI: http://dx.doi.org/10.1145/2835776.2835820.

accuracy of 55% at the city level in the U.S., and the worst having an accuracy of 32.9%. *(Section 6)*

4. **We propose a preliminary method to improve IP geolocation databases using information extracted from user queries.** We show that user queries are a good source of location information. We use this information to improve an IP geolocation database by modifying the locations of certain IP ranges. The improved database can be used in a variety of other applications outside of search engines. We discuss privacy implications and steps we take to remove personally identifiable information. *(Section 7)*
5. **Finally, we run a large scale online experiment to validate our improvements to an IP geolocation database.** We conduct an A/B experiment on 1.7 million distinct search engine users and 3.2 million queries over the course of seven days, in the Mexico market. We observe improvements in several user metrics. *(Section 8)*

## 2. IP GEOLOCATION DATABASES

The most precise way to determine the location of an online service user is to use positioning systems such as GPS, Galileo, GLONASS and BeiDou [19, 13, 20]. Unfortunately, this information is only available on devices that have location capabilities. Other location detection methods include Wi-Fi positioning systems, as well as requesting users to self-report their location. In the former case, the location of the user is determined by comparing the list of wireless access points visible from the device to a list of access points with known locations. These databases suffer from lack of coverage, especially in rural areas [11]. In the latter case, having users self-report their location might be accurate in the short term, but such information can become stale in the long run if users move without updating their location.

IP geolocation databases are an alternative with higher coverage. They are typically provided by commercial vendors such as MaxMind, IP2Location, Neustar, IPligence, and Digital Element. These databases are compiled by combining data from different sources. The primary data sources are the five regional Internet registries, which manage how IP ranges and network autonomous system numbers are assigned to organizations around the world [22]. These registries maintain databases with mappings of IP ranges to the geographical location of the organizations that use these IPs. However, there are cases where large ISPs or multinational companies are assigned large IP ranges. Since these organizations deploy infrastructure over large areas, the geographic information contained in the registry databases is insufficient to obtain accurate locations from any IP address in the range. Secondary data sources can include more granular data collected from ISPs or from websites which collect user location data. In the former case, large ISPs maintain internal records of how they geographically subdivide large IP ranges provided by the Internet registries. In the latter case, the information is obtained from the logs of websites, such as weather websites, which ask the user to self-report their location.

Commercial IP geolocation providers offer their databases in a format similar to the example in Table 1. The IP ranges are specified as integers instead of the typical dotted-quad notation to make lookups faster. While most ranges cover IPv4 addresses, providers have more recently started to provide information on the IPv6 address space. However, since IPv6 adoption is currently low, we restrict our discussion to IPv4 ranges.

## 3. RELATED WORK

We divide previous research into two categories, based on the methods employed to generate IP geolocation: network delay and topology, and web mining.

Table 1: Rows from an IP geolocation database.

| StartIP | EndIP | Country | State | City |
|---|---|---|---|---|
| 16781312 | 16781567 | JP | tokyo | tokyo |
| 16789504 | 16789759 | CN | guangdong | guangzhou |
| 3147413504 | 3147413759 | MX | quintana roo | cancun |

### 3.1 Network Delay and Topology

The line of research on Network Delay relies on the observation that the delay experienced by network packets as they travel between two Internet hosts is proportional to the geographical distance between the hosts. Early work on IP geolocation by Padmanabhan and Subramanian discusses *GeoPing* [26], which uses delay measurements made from geographically distributed locations using ICMP packets to infer the coordinates of the target IP. The method has an error distance of 150 kilometers at the 25th percentile when locating 256 servers hosted in universities across the United States. *CBG* [17] is a method that uses a geometric approach using multilateration to estimate the location of a given target IP. Experimental results on 95 IP addresses in the U.S. and 42 addresses in Western Europe produce a median error distance of below 100 kilometers for the first dataset and below 25 kilometers for the second dataset, respectively. Youn et al. [33] develop a statistical method for IP geolocation based on applying kernel density estimation to delay measurements. On a dataset of 85 IP addresses, the method achieves a median error of 53 kilometers.

Network Topology Geolocation methods also use knowledge of the network structure to achieve increased accuracy. *GeoCluster*, also proposed by Padmanabhan and Subramanian [26], combines BGP routing information with sparse IPs of known locations to assign geographical locations to whole address prefixes. On a set of 256 target IPs located at universities in the United States, *GeoCluster* achieves a median error of 28 kilometers. This approach performs well with hosts located on university campuses, but it performs much worse on a larger and more realistic dataset, with errors of hundreds of kilometers. *TBG* [23], which stands for Topology-Based Geolocation, uses the *traceroute* tool to collect both the intermediate routers located between the fixed probes and target IP address, and the network round trip measurements along each intermediate link. *TBG* achieves higher accuracy than *CBG*, with a median error of 67 kilometers on a dataset consisting of hosts located at U.S. universities. *Octant* [32] is a framework where the location of the intermediate routers and target IP are constrained to specific regions based on their delays from landmarks and intermediate routers. It achieves a median accuracy of 22 miles on a set of 51 machines. *GeoTrack* [26] tries to infer the location of IP addresses from the DNS names of the routers along the *traceroute* path. It yields a median error distance of 590 km on a test dataset of 2,830 IPs.

**Both Network Delay and Network Topology methods have significant limitations.** First, all such methods require access to nodes spread throughout the globe to perform measurements. Second, geolocating a large number of IP addresses using network measurements can run into scalability issues, as each target IP address or range requires separate measurements. Third, not all networks allow ICMP pings or fully disclose their network topology. Fourth, routes on the Internet do not necessarily map to geographic distances. Fifth, previously reported mean and median errors of tens to hundreds of kilometers show that these methods cannot be used for practical applications at the city granularity. Sixth, the ground truth data for work in this area is usually limited to a few tens of IP addresses, typically located in the United States.

Our work addresses several of these limitations. We mine search engine logs by extracting explicit locations mentioned in user queries. This approach bypasses the limitations of delay measurements and network topology methods. We also use the logs to generate a ground truth dataset of 8.4 million IP addresses with known location, as determined by realtime GPS location. Our ground truth dataset is several orders of magnitude larger than the ones used in previous work. The three baselines we use generally outperform the median and mean distance error of approaches proposed in previous work. For instance, in the case of the United States, one of our baselines has a median error of only 6.9 kilometers. Finally, we show that our method improves accuracy across tens of countries, compared to prior work which usually evaluates locations in the United States, or a handful of large countries in Europe and Asia.

## 3.2 Web Mining

Web mining approaches use diverse information mined from the web. Guo et al. [18] extract locations mentioned in web pages and assign the locations to the IPs which host the content. Using an IP geolocation database as ground truth they report an agreement on the city level for 87% of the IPs. This work has two obvious problems: first, an IP geolocation database with unknown accuracy is used as ground truth; and second, the method focuses on geolocating IP addresses of servers, not of end users. Endo and Sadok [14] propose using *whois* information. Unfortunately, the evaluation section lacks a comparison against ground truth. Wang et al. [30] combine the *CBG* approach with extracting the location of web servers from the web pages that they host. Using the web servers as landmarks they are able to achieve good results, with a median error of 0.69 kilometers for the best data set, which contains 88 target IP addresses. The approach suffers from the same scalability limitations as network delay based methods. The results also depend on the density of servers in a geographical area.

Backstrom et al. [6] propose an interesting approach which relies on a user's social graph to determine their location. They derive the location of target users based on the locations of friends. Using self-reported location as ground truth they show an improvement over an unnamed IP geolocation database. For an error distance of less than 25 km, the amount of correctly classified IPs increases from 57.2% for the baseline to 67.5% for the proposed method. The authors state that this method works so long as an individual has a sufficient number of friends whose location is known, preferably more than 16.

# 4. GROUND TRUTH

Limited ground truth information has been a significant limitation in previous IP geolocation research, as discussed in Section 3. Previous work has typically used tens of IP addresses with known geographical location, mainly located in the U.S. We propose a method to generate a large scale evaluation set by aggregating real time location information from search engine logs. Using our method we have generated a ground truth set of more than 8.4 million IP addresses from across the world.

Mobile devices contain sensors for global positioning systems. Mobile applications can request access to real time location information in order to provide better results. While not all users are comfortable with sharing their location, a representative set of users agree to provide this information. We use this information from search engine logs to generate the ground truth set.

Although we collect most of the ground truth data from mobile devices, **we perform several filtering steps to ensure that most IP addresses in the ground truth set are from fixed broadband connections**. The main reason for retaining only fixed IP addresses is that here we aim to improve existing IP geolocation databases, which assume each IP address has a single location. Since here we are computing a ground truth set, it is reasonable to eliminate IP addresses which appear in multiple cities in the dataset. Second, obtaining the real location of fixed broadband connections is arguably more difficult than obtaining the location of mobile network connections, as most most wired devices such as desktop PCs do not contain sensors capable of determining real time location, while mobile devices do.

First, after aggregating all reported locations per IP address, we restrict the IP addresses to the ones where the location readings stay within a 1.6 kilometer (1 mile) radius. We then assign each IP address a centroid computed by combining these location readings in time. While this constraint does filter out some valid IP addresses, it ensures that the remaining addresses are located in a relatively fixed position. Consider a user who commutes to work in a close-by city. Although the heuristic will likely filter out their mobile IP address as cell phone tower coverage spans large geographical regions, it will retain their fixed broadband IP address, if the user connects the device to both networks. When the user reaches their home and switches to a fixed Wi-Fi connection, their device will still broadcast the real time location to the search engine, thus linking the fixed broadband connection to a fixed location which stays within a small radius. Second, we retain only the IP addresses which appear in the logs on at least three different days, with at least three different location readings. This step ensures that we filter out the devices which are only turned on for a short period of time or that broadcast the same identical stale location.

We have obtained 8.4 million IP addresses and their corresponding location by mining logs from the Bing search engine for a period of 180 days ending on October 10th, 2014. To the best of our knowledge, this is the largest ground truth set ever used in IP geolocation research. The set spans 220 countries, with a maximum of 2 million IP addresses in the United States. The top 50 countries by IP density have a mean and median size of 163,000 and 45,000 IP addresses, respectively.

# 5. IMPACT OF INCORRECT IP GEOLOCATION

We now use the ground truth dataset to analyze the impact of impressions where the location information is incorrect. These cases are nuanced as we expect the effect of incorrect information to become more apparent as the error increases. For instance, for the query *"restaurants"* a user might still click the results even if the search returns businesses from a close-by city.

We extract the impressions issued on the Bing search engine in a seven day period, across all devices. We intersect this set with our ground truth set and then we compute the distance between the ground truth location and the location from the IP geolocation database used by Bing. We compare the set of impressions where the distance between the real location and the assumed location is more than 15 km, to the impressions where the error distance is less than 15 km. We further partition the data based on queries with local intent versus all queries. By a query with local intent we mean queries with local context, such as "plumbers".

Table 2 shows the change in metrics for the queries issued from the U.S. market. Both overall and ad click through rates decrease when the location is incorrect. Prior research has shown that this metric is positively correlated with user satisfaction [16]. However, we observe that Algorithmic CTR increases for the "All queries"

Table 2: Change in metrics between queries where distance error is more than 15 km, versus when the error is less than 15 km. **Note that both result columns show the impact of incorrect locations.** The difference between the columns is that they show two subsets of the data. See text for an explanation of the two positive metrics.

| Metrics | Local Intent | All Queries |
|---|---|---|
| Overall CTR | -4.3% | -12.8% |
| Algorithmic CTR | -1.1% | +2.4% |
| Algorithmic Click Success | -6.1% | +0.2% |
| Ads CTR | -17.9% | -6.1% |
| Ads Click Success | -15.2% | -7.0% |
| Ad Revenue | -40.3% | -6.0% |
| Results statistically significant using a two-sample t-test at 1%. | | |

set. This increase can be attributed to users who avoid clicking on the answers which contain incorrect local content, and instead choose to click on algorithmic search results, which might be more general.

The table also indicates that ads click success is negatively affected. Click success occurs when the click results in a dwell time greater than or equal to 30 seconds [16]. This metric has also been correlated with user satisfaction. Therefore, since in our case the metric goes down, user satisfaction might suffer. However, in the case of algorithmic clicks we observe that there is an increase in success when the location is incorrect. This change can be caused by users that click on general algorithmic results because the local answer shows incorrect results. Finally, we can observe ad revenue decreases dramatically in both cases. Therefore, improving IP geolocation accuracy could lead to higher revenue.

Other geographic markets yield similar results, but we do not show them here due to space constraints. We conclude that when the IP geolocation data is incorrect there is a significant drop in metrics. These results reinforce our hypothesis that incorrect IP geolocation information can lead to lower user satisfaction.

## 6. EVALUATING IP GEOLOCATION DATABASES

We use the ground truth dataset described in Section 4 to evaluate three commercial IP geolocation databases, chosen based on their high self-reported accuracy. These databases represent the state of the art in the industry and, as we will show here, have higher coverage and better accuracy than most previously published research. Due to legal reasons we will use the generic names Vendor *A*, Vendor *B*, and Vendor *C* to denote them. The providers were chosen from among the companies listed in Section 2. Nevertheless, the findings are still valuable.

Table 3 presents basic statistics about these databases, generated on data available on October 10th, 2014. The IPv4 address space is 4.29 billion IP addresses, which includes 592 million addresses yet to be distributed to organizations, or reserved for other uses. Vendor *C* has the highest IP coverage at 3.6 billion, followed by the other two vendors each at 3.48 billion. The databases report countries using the ISO 3166-1 standard, which yields a maximum of 249 possible countries. Vendors *A* and *B* cover all possible countries, while Vendor *C* covers 247 of them. The last column in the table shows city level coverage. The difference between the vendors with highest and lowest city coverage is 36,357 cities. The databases have very high coverage at the IP and country level, but varying degrees of coverage at the city level.

We now turn to the ground truth data to evaluate the accuracy and error distance of the databases. Since the data is provided by different companies, it is possible that the names of locations are not consistent. To solve this problem we have used the public Bing reverse geocoding API [2], which maps coordinates to street addresses. The reverse geocoder is aware of the boundaries of cities. This reverse geocoder is the same one that we have used while generating the ground truth dataset in Section 4. Since we are interested in geolocation at the city level, we discard the street addresses. We normalize the locations in each database by assigning a unique identifier to each distinct city. Using the same reverse geocoding method on the databases and on the ground truth data ensures a level playing field in terms of location names and identifiers. One important side effect to note is that all ground truth locations within the limits of a city are converted to the coordinates of the center of that city.

Table 3: Basic statistics on the the IP geolocation databases used in this work.

| Provider | IP Coverage | Countries | Cities |
|---|---|---|---|
| Vendor **A** | 3.48B (˜94.1%) | 249 (100%) | 87,842 |
| Vendor **B** | 3.48B (˜94.1%) | 249 (100%) | 66,480 |
| Vendor **C** | 3.6B (˜97.3%) | 247 (99.2%) | 102,837 |

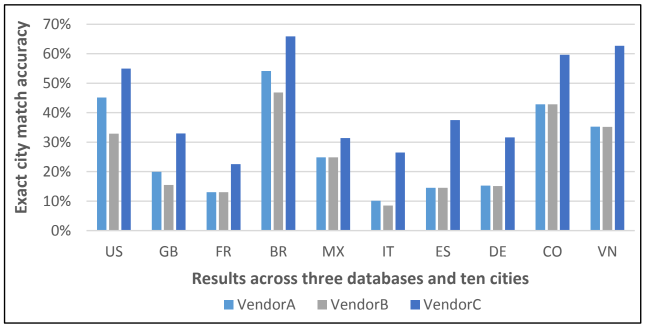

Figure 1: Accuracy at the city level for the three IP geolocation databases, across ten countries. Data points show exact city match accuracy.

We compute the accuracy of exact city matches by counting how often the location of ground truth IPs matches the location given by the databases. In Figure 1 we show the accuracy for the three databases across the top ten countries by ground truth IP density. There are two interesting findings. First, none of the databases achieve an accuracy above 70% at the city level, and in some cases have an accuracy below 10%. Second, Vendor *C* outperforms the other two providers, often significantly, across all ten countries. The same findings hold for all countries. This finding is particularly interesting as in 37.4% of the database from Vendor *C* there is no city level information. This suggests that Vendor *C* has a high accuracy but low city-level IP coverage, while the converse is true for the other two vendors. We have also found that country-level accuracy is significantly better than city level accuracy, with a median of 95.3%, 96.7%, and 89.2% in the top 50 countries for the three vendors, respectively.

Exact city level matches might not present a complete picture of the performance of IP geolocation databases. Therefore, we also compute the error distance between the cities from the ground truth IPs, and the cities returned by the geolocation databases for these IPs. The error distance is given by the distance between the centers of the cities. For a given ground truth IP, if the ground truth and IP geolocation cities are the same, the distance is zero. Figure 2 shows the cumulative error distance for the three databases, in the United

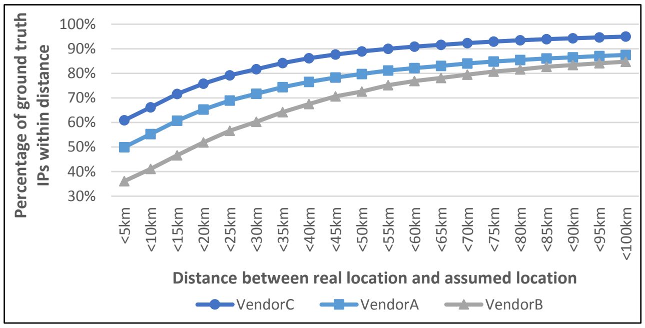

Figure 2: Cumulative error distance for the three IP geolocation databases in the United States. Data points show percentage of ground truth IPs where the real location and the IP geolocation assumed location fall within a certain distance. The distance interval on the X-axis is five kilometers.

States. The shape of the curves is similar in the other countries. The X axis shows the error distance at an interval of five kilometers and the Y axis shows the percentage of ground truth IPs with that error distance. For example, consider the three data points above the X-axis label titled *<10km*. The numbers show that the real location of 66.2%, 55.3%, and 41.1% of ground truth IPs is within 10 kilometers of the location provided by Vendors *C*, *A*, and *B*, respectively. The closer the curve is to the upper left corner, the smaller the error distance, and the better the results.

We have also performed a comparison of the accuracy and error distance results between our ground truth method, and the self-reported data provided by Vendor *C* for the United States. We have found that the exact city accuracy difference is within two percentage points, while the error distance at 10 kilometers is within seven percentage points. For the exact city match accuracy their results are within 10 percentage points of the results we have obtained in Figure 1.

# 7. IMPROVING IP GEOLOCATION

We investigate the feasibility of using location information extracted from search queries to improve IP geolocation databases **at the city level**. We define important terms used in the section, present the problem statement, go over challenges, discuss the approach we have taken, and evaluate it against the ground truth dataset.

The aim of the preliminary method presented here is not to optimize the model but rather to determine the predictive value of user queries in determining the location of IP addresses.

## 7.1 Definitions

We start by defining some terms used in this section.

- *Explicit location queries* are search engine user queries which contain locations. Examples include "weather in bellevue" and "movie showtimes in kirkland wa".
- A *target geolocation database* is a baseline IP geolocation database that we are starting from and which we are aiming to improve by combining it with locations extracted from user queries.
- *Reverse geocoding* is the process of converting a latitude and longitude pair to the corresponding readable address (continent, country, state, county, city, street, etc.). In this paper we reverse geocode locations down to the city level.
- *Primate cities* are cities which dominate the surrounding populated places economically and culturally due to their size [8].

## 7.2 Problem Statement

The examined commercial IP geolocation databases have higher accuracy, higher coverage, and lower error distance than previous research. For instance Vendor *B*, which had the worst results out of all commercial databases, has a median distance of 18 kilometers in the US, as measured by our ground truth set, while past research presented in Section 3 has a median error of 25 kilometers or more. So even the worst performing commercial database has a much better median distance than most published work. Furthermore, both total IP coverage and ground truth IP coverage are very low in previous research with only tens and up to thousands of IP addresses. Comparatively, Vendor *B* has a worldwide coverage of 3.48 billion IP addresses, and in this work we use a ground truth set of 8.4 million IP addresses.

Nevertheless, the accuracy of these commercial databases is still lacking for some practical applications and needs improvement. For example, Figure 2 shows that the database provided by Vendor *A* geolocates only 60% of the ground truth IPs within 15 kilometers of their actual location in the United States. If we were to use this database in a search engine setting, it might return incorrect location information for roughly 40% of real user IP addresses. In other countries the error distance can be even higher. If a search engine shows businesses, weather, or news for an incorrect location for 40% of the time, users might switch to a competitor.

To augment existing IP geolocation databases we must introduce new sources of information. One possible solution is to use real time location information, as we did for the ground truth data. While this approach has potential, in this work we focus instead on explicit location information extracted from user queries. **The reason why we use locations extracted from queries instead of directly using the ground truth set is the difference in IP coverage**. While our ground truth set covers 8.4 million IP addresses, the location information extracted from queries has a coverage several orders of magnitude higher. In our experiments we extract billions of location data points from queries. In future work we will investigate creating a hybrid approach which uses both methods, but in this paper we have decided to use the entirety of the ground truth dataset for testing.

Given a search engine log which includes IP addresses and user queries, and given an IP geolocation database, our task is to improve the accuracy of the geolocation database using the information in the log by correcting some of the locations in the database at the city level. The implication of starting from an existing database is that we are relying on the existing IP ranges already present in the database. The *maximum* IP range size in the three commercial databases we use is 256 addresses per range, which is a reasonably fine granularity. Furthermore, the coverage of the starting databases is very high, at more than 3.48 billion addresses each.

**We are improving existing databases instead of creating a new database from scratch for several reasons.** First, we aim to improve the existing state of the art, not replace it. Second, while the coverage of the locations we extract from user queries is high at 360 million distinct IP addresses, it is still not enough to cover all of the used IP range space. Third, as a practical manner it would be more efficient to make use of a single geolocation database which combines location information from different types of data sources. Creating an independent geolocation database from signals from GPS-equipped devices, locations extracted from queries, and other behavioral signals remains potential future work.

## 7.3 Technical Challenges

In order to extract locations from queries and combine them with an existing geolocation database, we had to overcome several technical challenges:

1. **We must extract locations from queries**. Given a query log, we must extract the explicit locations mentioned in the queries.
2. **Query logs have a large size and extracting locations from queries is computationally expensive at scale.** Our query logs contain billions of data points, which makes it difficult or impossible to store and process the data on a single machine. A secondary problem is that extracting locations from queries is computationally expensive. Therefore, we must extract queries which are most likely to contain locations.
3. **The locations in the database and the locations extracted from queries are not normalized.** We must normalize locations to bring them into a common space before we can combine and compare them directly. For instance the query "plumbers in nyc" contains the location "nyc", which cannot be matched directly with the location "40.7141667, -74.0063889" stored in the geolocation database.
4. **For a given IP range there can be multiple candidate locations extracted from the query log.** Users within a certain IP range are likely to search for a variety of different locations. We must rank these locations in order to pick the most likely candidate for the IP range.
5. **We need to compensate for the effect that primate cities have on surrounding towns.** Primate cities can skew results due to their overall popularity. Consider the example in Figure 3. Here we plot all the locations mentioned in user queries issued from a single IP range, within one month. In this example the radius of the circles depicts the number of times each location was mentioned by users. We can observe that London and Reading are both popular, but London is mentioned more often. The actual number of mentions is 2,159 for London and 1,254 for Reading. In reality, the correct location of users in this IP range is Reading. Here we observe that primate cities can skew the number of mentions due to their global popularity. Therefore, in our approach we must correct for this effect.
6. **We must combine the locations extracted from query logs with the pre-existing IP geolocation database**. For a given IP range, whenever a location extracted from queries does not match the location in the geolocation database, we must find a scoring system to choose whether to keep the original location or change it to the one extracted from user queries.
7. **The resulting combined database needs to be evaluated on a ground truth dataset and ideally tested on a production application**. Here the challenge is devising the ground truth set and determining which metrics to use. Note we have already partially covered this in Section 4. It is also worthwhile to obtain more proof of the improvements using a real production application. The challenge in this second case is determining objective measures of user satisfaction.

Since we are focused on the feasibility of using query logs to improve geolocation, wherever possible we have tried to reuse existing state-of-the-art technologies, allowing us to focus on the novel contributions instead of re-implementing solutions to already solved problems. More specifically, our contributions are focused on challenges *4*, *5*, *6*, *7*, and the second part of challenge *2*.

## 7.4 Datasets

Below is a list of the datasets we use in this section:

- **Main query log**: It contains 180 days of Bing query logs, ending on October 9th, 2014. This dataset spans hundreds of billions of queries.
- **Validation query log**: For parameter tuning we use 30 days of Bing query logs collected prior to the main query log. There is no overlap between the main query log and the validation log.
- **Baselines**: We use the three IP geolocation databases we previously described in Section 6 as baselines. We consider these databases to be the state of the art in the industry.
- **Ground truth**: The ground truth, which we described in Section 4, contains 8.4 million IP addresses with known location.

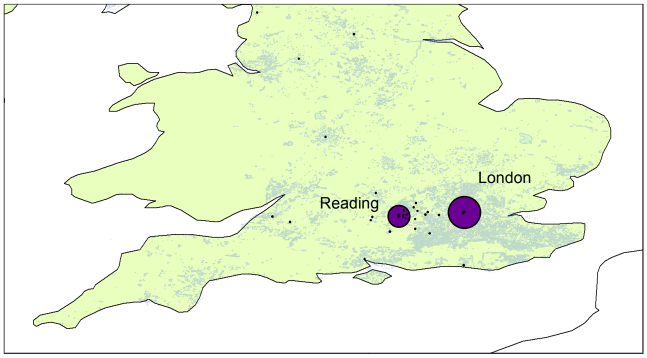

Figure 3: An example of the effect of *primate cities*. If we were to choose the location based solely on number of mentions, we would choose London. In reality, the correct location is Reading.

## 7.5 Approach

We propose improving IP geolocation databases by correcting the location of certain IP ranges using cities extracted from user queries. We begin from the assumption that when search engine users use explicit locations in their queries, in aggregate these queries reveal the users' location. This assumption may not be true at an individual user level. For example, a person might live in New York City, but they might be planning a vacation in Italy. Queries such as "restaurants in venice" might lead to the incorrect conclusion that the user lives in Italy. However, our hypothesis is that if we aggregate queries from users within the same IP range, the locations which are most mentioned will be geographically close to the users in the range. Furthermore, even at the individual level users do not search for the same non-local locations for extended periods of time. In our example, once the user has completed their research on Venice, they might resume searching for locations closer to home.

Users include explicit locations in their queries for several reasons. First, it is possible they do this because of habit or because they do not realize that search engines know their general location. Second, they might have noticed that the search engine returns incorrect location for their searches, and they are correcting it by explicitly specifying the location. For example, if for the query "weather" the answer shows the weather forecast in an incorrect city, the user is likely to click on a search result instead, switch engines, or requery using the explicit location, such as "weather in seattle". Third, they might be searching for information in a location other than their own city. The first two cases yield the true location of the user, while the last case can lead to false positives.

Figure 4 illustrates the intuition behind our approach with two examples. In each of the examples we plot the locations mentioned in queries issued from one IP range in a one month interval. The blue dots represent the distinct locations mentioned in the queries, while the shading around them shows how often each location was mentioned. The white dot reveals the real location of users in the IP range, as given by GPS information. The correct cities are Morelia in Mexico, and Florence in Italy, respectively. In both cases the city with most query mentions is also the correct one.

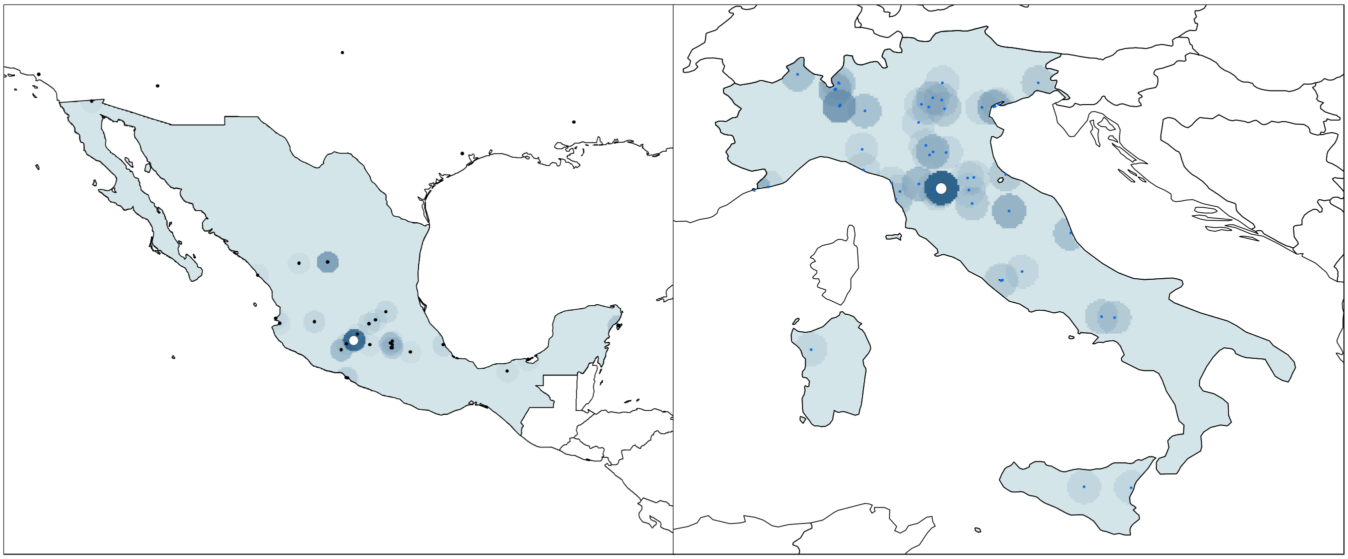

Figure 4: Examples of locations extracted from user queries. Each map shows data from a separate IP range. The individual points show the distinct geographical locations mentioned in user queries issued from the IP range. The amount of shading around them shows the frequency at which these locations were mentioned. The larger white dot shows the actual location of users in the IP range as given by realtime GPS data. The maps show that in both cases the location most frequently mentioned in user queries is also the real location of the users. The figure shows Morelia, Mexico on the left, and Florence, Italy on the right. The size of the IP ranges is 60 and 11 IP addresses, respectively.

Given search engine query logs and an IP geolocation database, our goal is to improve the accuracy of the database. Below is a summary of the steps we have taken to solve this problem.

1. **Extract queries** and corresponding IP addresses from query logs.
2. **Filter impressions**, keeping the ones likely to contain locations.
3. **Extract locations** from the queries that remain.
4. **Reverse geocode locations** extracted from queries and locations extracted from the target geolocation database.
5. **Aggregate locations** first on IP address, then on the IP ranges in the target IP geolocation database.
6. **Compute the popularity of each distinct location** to be used as a proxy for determining *primate cities*.
7. **Score the location candidates** in each IP range.
8. For each IP range where there are candidates, and where the top query location is different than the original location in the IP range, **decide whether to keep the original location** or modify it based on queries.
9. **Test the modified geolocation database** against the ground truth.

We will now provide details for each of the steps. We first extract the queries and IP addresses from the query log. We have performed this step using an implementation of SCOPE, which is a language for processing massive datasets in parallel across a distributed cluster of machines. Please refer to the work by Chaiken et al. for more details [10]. Due to the capabilities of SCOPE, this step is trivial to implement. The result is a smaller but still sizable dataset where we have removed unwanted extra metadata.

Second, we filter the impressions in the log obtained in the first step to retain only the queries which are most likely to contain explicit locations. This filtering step is required in order to reduce the number of queries we need to pass to the location extraction phase. We keep the impressions where the query has local intent, such as "plumbers in chicago", using the production Bing query classifier.

Next, we extract explicit locations from queries. To achieve this we use a method similar to the query dominant location extraction algorithm presented in Wang et al. [29]. We use the user query as input. The output contains the detected location, including the country, city, and coordinates. When no location is detected in the query, the output is empty. This results in several billion explicit location queries, issued from 360 million distinct IP addresses. Please refer to Subsection 7.8 for a discussion on reproducing these experiments using publicly available resources.

We then normalize both the locations extracted from user queries and the locations contained in the target IP geolocation database through reverse geocoding using the publicly available Bing reverse geocoding API. Given a latitude and longitude pair as input, the output is a normalized location at the city level which contains the country, state and city. Each city is assigned a unique identifier. Distinct cities with the same name receive different identifiers. Since the geocoding is performed at the city level the public Bing API achieves an excellent reverse geocoding accuracy in all 50 countries we test in this paper. We perform this step to ensure that the locations extracted from queries and the locations in the target database can be directly compared.

We aggregate the candidate locations per IP address. For each IP address where we have extracted at least one location from user queries, we count the occurrences of each distinct location. We then map these IP addresses on the IP ranges of the target geolocation database, and further aggregate locations per IP range. These two related tasks are performed in a distributed fashion. The result is a list of candidate locations for each IP range, along with their counts.

We then compute the popularity of each distinct city across all IP addresses. We use these counts as a proxy for determining *primate cities*. We name these counts *GlobalMentions*.

Next, we rank the candidate locations in each IP range to determine the top one for that range. For each location mentioned in an IP range we compute *MentionsNorm*, which stands for *normalized mentions* and is shown in Equation 1. For each IP range the candidate cities are ranked in descending order by *MentionsNorm* and the top location is retained. In the equation, *LocalMentions* is given by the number of times the current location was mentioned by users in the current IP range. *GlobalMentions* is the number of times the current location was mentioned across all IP addresses. *IPInst* counts the number of distinct IP addresses in the current IP range that have mentioned the location. Exponents $x$ and $y$ can increase or decrease the importance of *LocalMentions* over *GlobalMentions*. Using a local parameter search on the separate validation dataset we have set $x$ to be 1.5 and $y$ to be 0.5. The parameter search was performed using a step size of 0.1 and limits of 0.0 to 3.0. The exponents can account for *primate cities* by promoting smaller cities and demoting larges ones. In the previous example from Figure 3 London will now have a lower score than Reading, which is the correct choice.

$$MentionsNorm = \frac{LocalMentions^x}{GlobalMentions^y} \cdot IPInst \qquad (1)$$

In the last step, for each IP range where the city in the database does not match the location extracted from user queries, we have to decide if we have to modify the location in the database. To achieve this we introduce Equation 2. *IPInstPercentage* in the equation is the percentage of IP addresses in the IP range which mentioned the current location, where *IPInst* is defined as in the previous equation and *EndIP* and *StartIP* are the last and first IP address in the range, respectively.

$$IPInstPercentage = \frac{IPInst}{EndIP - StartIP + 1} \qquad (2)$$

To determine if we have to perform the replacement we use both equations 1 and 2. Using the same validation dataset we determine the cutoff thresholds for each equations. That is, we perform the replacement only if the top location in the IP range has *MentionsNorm* of at least 0.3 and *IPInstPercentage* of at least 5%. We have determined these thresholds using a local parameter search with a step size of 0.05 and 0.5%, respectively, a minimum of 0, and a maximum of 1 and 25%, respectively.

## 7.6 Ground Truth Evaluation

We compared the accuracy and distance error of the three commercial IP geolocation databases to the equivalent databases modified using our approach based on user queries. Figure 5 shows the improvements in exact city match accuracy for five high-traffic countries. We observe the best improvement for Vendor *B*, where for countries such as Germany, Italy, and Spain, the accuracy increases by more than 100%. Table 4 shows an evaluation summary for the top 50 countries by IP density. For Vendors *A* and *B*, the accuracy improves in 49 out of 50 countries, while for Vendor *B* the accuracy improves in 44 countries. In the few cases where accuracy decreases, it does so by less than 0.4%: in Israel for the first two vendors, and Colombia for the last one. Median and mean accuracy computed across all 50 countries show significant gains, especially for the first two vendors. For Vendor *C* the improvement is more modest, as this baseline had the highest initial accuracy. Nevertheless, for several countries such as India, Belgium, Netherlands, and Mexico the improvements are higher than 5% for Vendor *C*, with Taiwan seeing the best improvement at +74.2%.

The improvements are also apparent when we plot the cumulative error for the distance between correct and assumed locations. In Figure 6 we show the cumulative error for Vendor *B* in the United States. We do not superimpose the results for the other two vendors here to make the figure easy to understand. The shapes of the curves for the other two vendors are similar, but the improvements are less pronounced. The results for Vendor *B* show a remarkable improvement, as the percentage of ground truth IP addresses where the error is less than 5 kilometers increases from 36.1% to 58.7%.

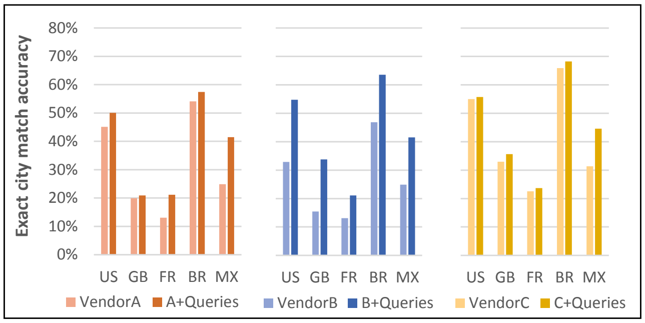

Figure 5: Improvement in accuracy for five high-traffic countries between the original IP geolocation databases and the databases modified using locations extracted from user queries.

Table 4: Summary of evaluation results across three IP geolocation databases and the top 50 countries by IP density. The results are for city level accuracy.

| Vendor: | A | B | C |
|---|---|---|---|
| **Countries Accuracy Improved** | 49 | 49 | 44 |
| **Countries Accuracy Decreased** | 1 | 1 | 6 |
| **Median Accuracy Change** | +20.7% | +32.2% | +1% |
| **Mean Accuracy Change** | +114.8% | +121.1% | +3.4% |
| **Worst Accuracy Change** | -0.2% | -0.2% | -0.4% |

## 7.7 Privacy Considerations

Our approach protects the privacy of search engine users, as query logs contain sensitive personal information. First, our method extracts locations from user queries, discarding other words in the queries. Second, all extracted locations are normalized using reverse geocoding. We retain location information only at city level granularity, although some queries initially contain locations which are precise up to the street level. Third, the location information is further aggregated at the IP range level, which combines data from individual users in that IP range into a single set of location counts. Aggregating data across a range of IP addresses makes the data less granular. These steps provide strong privacy safeguards as the output data is coarse and contains no personally identifiable information.

## 7.8 Reproducing Experiments

Most of the steps in our approach can be reproduced using public techniques and APIs. The IP geolocation databases can be obtained

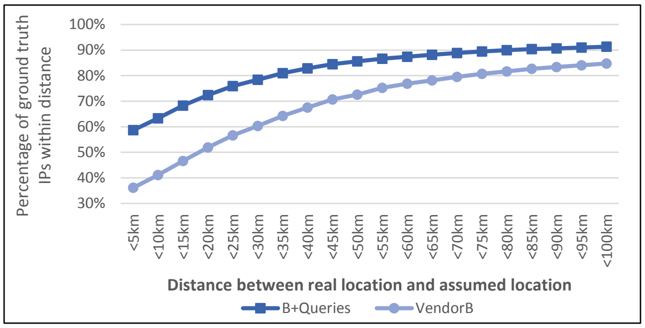

Figure 6: Cumulative error distance from ground truth to Vendor*B*'s IP geolocation database and the same database modified using locations extracted from user queries.

Table 5: Statistically significant changes in metrics for the A/B experiment carried out on the Mexico market.

| Metrics | Change | P-Value |
|---|---|---|
| **Overall Click Success** | +0.8% | 0.03 |
| **Answer Click Success** | +1.67% | 0.04 |
| **Ads CTR** | +1.57% | 0.03 |
| **Ads Click Success** | +1.57% | 0.03 |
| **Entity Pane CTR** | +1.58% | 0.03 |

from the companies previously listed in Section 2. Locations can be extracted from queries using publicly available alternatives such as Yahoo! PlaceSpotter [5] and OpenCalais [4]. Another substitute is a Named Entity Recognizer such as Stanford NER [15] or GATE ANNIE [12], combined with a geocoder such as the ones from Bing [1] or Google [3]. The same geocoders also provide functions for reverse geocoding, which would allow normalizing locations. Finally, all distributed aggregation steps can be implemented on an open source implementation of MapReduce, such as Hadoop [31].

# 8. VALIDATING IMPROVEMENTS

In order to validate our IP geolocation improvements, we have carried out an A/B experiment on the Bing search engine during the seven day period ending on November 1st, 2014. The treatment and control variants each spanned approximately 850,000 unique users and 1.6 million queries from the Mexico market, which represents a sizable percentage of that market. For the control we have used a proprietary IP geolocation database obtained by combining the databases from Vendors *A*, *B*, and *C*, and by adding other sources of information. The resulting database has higher accuracy at the city granularity than any of the individual databases from the three vendors. The geolocation database used in the treatment was generated by modifying the location of some of the IP ranges in the proprietary database, using the same data and methodology as described in Section 7. Comparing the two databases using the ground truth and methodology explained in Section 6 shows that the treatment database has 10% higher accuracy in the Mexico market than the control database. We have purposely chosen to conduct the experiment in a country where our observed improvement is neither the highest nor the lowest.

Table 5 contains the statistically significant changes in user metrics computed at the conclusion of the experiment. The first two rows show that both the overall and the answer success clicks have improved, with 0.8% and 1.67%, respectively. By answers we mean the special page blocks, such as restaurant and cinema listings, which are visually different than the algorithmic results. An improvement in the success metrics indicates that users are more likely to be satisfied with the results once they click on a link. The table also shows an improvement in advertising click through rate and click success. These changes can lead to higher advertising revenue. Finally, the metrics show that the click through rate on the Entity Pane has increased by 1.58%. By Entity Pane we mean the right side of the screen, also known as a Knowledge Graph in the context of the Google search engine, which presents rich contextual information such as details about businesses and maps with close-by restaurants. Since a large percentage of such contextual information is based on the location of the user, the increase in engagement could have been caused by displaying more relevant local information.

Finally, we studied the impact of the experiment on the local answer by focusing on the subset of page views from the IP ranges modified by our method. Compared to the control we see that the overall local answer coverage has decreased by 1.65%, but the click through rate on individual items has increased by 75.5%. These results, which are are statistically significant with P-value at 0.04 and 0.03 respectively, show that while the local answer is shown slightly less often, the engagement with the answer is much higher. The improvement in engagement suggests that the local results might be more accurate in the treatment due to higher geolocation accuracy.

# 9. DISCUSSION AND IMPLICATIONS

Our work is the first to propose improving IP geolocation databases using search engine logs. We claim that mining search engine logs is a natural choice for this task, as these logs centralize a great deal of location information from diverse and geographically dispersed users. This approach has several advantages: it does not rely on network delay measurements as previous research, it can scale to cover any country, and it generally leads to accuracy measurements which are higher than previous work. Furthermore, real time GPS location extracted from the logs can be used to generate large scale ground truth data.

Bennett et al. [7] demonstrated that search results can be improved by incorporating location-based features into the ranking function. Here we have shown that incorrect location can impact user experience negatively. Therefore, IP geolocation databases with higher accuracy can result in improvements in location-based personalization. Outside the realm of search engines our work has implications in fields such as credit card fraud protection. We have shown that IP geolocation databases have relatively high accuracy and low error distance in large countries, such as the United States. However, we have also seen that the accuracy is lower in smaller countries or countries with lower Internet penetration. Increasing the accuracy of IP geolocation is crucial to combating credit card fraud in countries such as Ukraine, and Malaysia [9].

# 10. CONCLUSIONS AND FUTURE WORK

In this paper we have generated a large scale ground truth IP geolocation dataset, we have studied the impact that incorrect IP geolocation has on search engine user metrics, and we have proposed a scalable approach to improving IP geolocation at the city level, using search engine logs. We have generated a large ground truth dataset, containing millions of IP addresses with known location, by mining GPS locations from search engine logs. Our ground truth dataset is the largest and most geographically diverse in the open literature, and the first to use GPS location from devices. We have demonstrated that IP geolocation information is vital to search engines by showing how user metrics degrade when the user location is incorrect. We have evaluated three state-of-the-art commercial IP geolocation databases, showing that their accuracy is generally higher than most previously published research efforts. Starting from the three baselines, we have demonstrated an approach to improve IP geolocation databases using explicit locations extracted from query logs. We have validated the IP geolocation improvement process by carrying out a large scale A/B test on the Bing search engine. The results are promising, with statistically significant improvements in several user metrics, such as click success and ads click through rate.

There are several directions for future work. One possible improvement could result from combining data from different sources. For instance, instead of strictly using locations from user queries one could combine them with the real time location we have used for our ground truth, or with social network information as proposed by Backstrom et al. [6]. Another promising approach would be to analyze the implicit behavior of users. Mining search result clicks in search engine logs or toolbar logs could reveal preferences towards local websites, such as *The Kirkland Reporter*. It could also be worthwhile to study the intent of local queries at a finer granularity. For example, we might want to see if queries for airplane tickets leaving from a certain location are more likely to show the real location of a user, compared to vacation related queries, such as "points of interest in vienna".

## 11. REFERENCES

[1] Bing Geocoding API. http://msdn.microsoft.com/en-us/library/ff701711.aspx, (accessed July 17, 2015).

[2] Bing Reverse Geocoding API. http://msdn.microsoft.com/en-us/library/ff701710.aspx, (accessed July 17, 2015).

[3] Google Geocoding API. https://developers.google.com/maps/documentation/geocoding/, (accessed July 17, 2015).

[4] OpenCalais. http://www.opencalais.com/, (accessed July 17, 2015).

[5] Yahoo! PlaceSpotter. https://developer.yahoo.com/boss/geo/docs/key-concepts.html, (accessed July 17, 2015).

[6] L. Backstrom, E. Sun, and C. Marlow. Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *WWW 2010*, pages 61–70, Raleigh, North Carolina, USA, 2010. ACM.

[7] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and Using Location Metadata to Personalize Web Search. In *SIGIR 2011*, pages 135–144, Beijing, China, 2011. ACM.

[8] B. J. L. Berry. City Size Distributions and Economic Development. *Economic Development and Cultural Change*, 9(4):573–588, 1961.

[9] T. P. Bhatla, V. Prabhu, and A. Dua. Understanding credit card frauds. *Cards business review*, 1(6), 2003.

[10] R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou. Scope: Easy and efficient parallel processing of massive data sets. *Proc. VLDB Endow.*, 1(2):1265–1276, Aug. 2008.

[11] Y.-C. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm. Accuracy Characterization for Metropolitan-scale Wi-Fi Localization. In *MobiSys 2005*, pages 233–245, Seattle, Washington, 2005. ACM.

[12] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *ACL 2002*, pages 168–175, 2002.

[13] A. El-Rabbany. *Introduction to GPS: The Global Positioning System*. Artech House mobile communications series. Artech House, 2002.

[14] P. Endo and D. Sadok. Whois Based Geolocation: A Strategy to Geolocate Internet Hosts. In *AINA 2010*, pages 408–413, April 2010.

[15] J. R. Finkel, T. Grenager, and C. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL 2005*, pages 363–370, Ann Arbor, Michigan, 2005.

[16] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating Implicit Measures to Improve Web Search. *ACM Transactions on Information Systems*, 23(2):147–168, Apr. 2005.

[17] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-Based Geolocation of Internet Hosts. *IEEE/ACM Transactions on Networking*, 14(6):1219–1232, Dec 2006.

[18] C. Guo, Y. Liu, W. Shen, H. Wang, Q. Yu, and Y. Zhang. Mining the Web and the Internet for Accurate IP Address Geolocations. In *INFOCOM 2009*, pages 2841–2845, April 2009.

[19] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. *Global Positioning System: Theory and Practice*. Springer, 1993.

[20] B. Hofmann-Wellenhof, H. Lichtenegger, and E. Wasle. *GNSS – Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and more*. Springer, 2007.

[21] C. Huang, D. Maltz, J. Li, and A. Greenberg. Public DNS system and Global Traffic Management. In *INFOCOM 2011*, pages 2615–2623, April 2011.

[22] K. Hubbard, M. Kosters, D. Conrad, D. Karrenberg, and J. Postel. Internet Registry IP Allocation Guidelines. Technical report, United States, 1996.

[23] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe. Towards IP Geolocation Using Delay and Topology Measurements. In *IMC 2006*, pages 71–84, Rio de Janeriro, Brazil, 2006. ACM.

[24] B. Kölmel and S. Alexakis. Location based advertising. In *First International Conference on Mobile Business*, Athens, Greece, 2002.

[25] L. MacVittie. Geolocation and Application Delivery. https://f5.com/resources/white-papers/geolocation-and-application-delivery, 2012 (accessed November 28, 2015).

[26] V. N. Padmanabhan and L. Subramanian. An Investigation of Geographic Mapping Techniques for Internet Hosts. In *SIGCOMM 2001*, pages 173–185, San Diego, California, USA, 2001. ACM.

[27] C. A. Shue, N. Paul, and C. R. Taylor. From an IP Address to a Street Address: Using Wireless Signals to Locate a Target. In *WOOT 2013*, Washington, D.C., 2013. USENIX.

[28] D. J. B. Svantesson. E-Commerce Tax: How The Taxman Brought Geography To The 'Borderless' Internet. *Revenue Law Journal*, 17(1):11, 2007.

[29] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting Dominant Locations from Search Queries. In *SIGIR 2015*, pages 424–431, Salvador, Brazil, 2005. ACM.

[30] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang. Towards Street-level Client-independent IP Geolocation. In *NSDI 2011*, pages 365–379, Berkeley, CA, USA, 2011. USENIX.

[31] T. White. *Hadoop: The Definitive Guide*. O'Reilly and Associates Series. O'Reilly, 2012.

[32] B. Wong, I. Stoyanov, and E. G. Sirer. Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts. In *NSDI 2007*, pages 23–23, Berkeley, CA, USA, 2007. USENIX Association.

[33] I. Youn, B. Mark, and D. Richards. Statistical Geolocation of Internet Hosts. In *ICCCN 2009*, pages 1–6, Aug 2009.