Building IP Geolocation Database from Online Used Market Articles

Hyunsu Mun, Youngseok Lee
Data Network Laboratory
Chungnam National University, Daejeon, Korea
munhyunsu@cnu.ac.kr
One morning in the past

- When the weather is curious
  - Enter Weather on Internet search site!
One morning in the past

- When the weather is curious
  - Enter Weather on Internet search site!
One morning in the past

- When the weather is curious:
  - Enter Weather on Internet search site!

However, we need to search again to get the weather of our area.
One morning nowadays

- When the weather is curious
  - Enter Weather on Internet search site like past!
One morning nowadays

- When the weather is curious
  - Enter
One morning nowadays

- When the weather is curious
  - Enter your location based search

Location based search results appear
IP Geolocation

- Geolocation: Identifying the actual geographical location of objects (or analogy)
- IP Geolocation: Using IP address to identify geographical location (or analogy)
  - When searching for Weather, use IP address to show location based results
  - A method to discover the location of a client when a resource request is received through the network.
The Accuracy of IP Geolocation

- IP geolocation data has an accuracy of geolocation
- Commercial companies such as MaxMind and IP2Location reveal their accuracy in database descriptions:
  - 87% accurate on a city level for the US within a 50km radius
The Low Accuracy in Korea

- Commercial database failed to ensure city-level accuracy
- Measuring the accuracy of commercial IP geolocation database is measured in U.S.
- The accuracy of commercial IP geolocation in Korea does not show as stated in the product description

Korea IP address distribution sequence
1) IANA(Internet Assigned Numbers Authority): ICANN
2) RIR(Regional Internet Registry): APNIC
3) NIR(National Internet Registry): KISA

The Low Accuracy in Korea

- Commercial database failed to ensure city-level accuracy [1]
- Measuring the accuracy of commercial IP geolocation database is measured in U.S.
- The accuracy of commercial IP geolocation in Korea does not show as stated in the product description
  - Korea IP address distribution sequence
    1) IANA (Internet Assigned Numbers Authority): ICANN
    2) RIR (Regional Internet Registry): APNIC
    3) NIR (National Internet Registry): KISA

The IP Geolocation Error in Korea

- IP address of Daejeon, Korea is mapped to a different location by databases
  - Real Location: Wired ethernet
  - MaxMind, IP2Location: Commercial version (Not lite version)
  - WHOIS: KISA WHOIS API
To Better IP Geolocation

- City-level IP geolocation based on PoP-level network analysis [2]
- IP geolocation using landmarks [3]
- Building IP geolocation database with a crowd-sourcing [4]
  - Using broadband performance test tool data


The Ruliweb Article

- Create an average of 300 posts
- Ruliweb postings are required to reveal the location (Province-level) and IP address
- IP address disclosure to prevent fraud user
- Seller to reveal the district-level location in the body area – about 45% disclosure
The Key Challenges

- Broad region coverage
  - The IP address assigned to the user of all districts
  - Not IP block coverage
- Continuous update
  - Flexibility to the change of ISP’s policy
Building IP Geolocation Database

Online used market

Crawl posts

HTML Document

Parse HTML

Header

IP address, Location1 (Metropolitan city)

Body

Location2 (District)

Extract
IP address and Location name

Group by /26 prefix

IP address1, District name1
IP address2, District name1
IP address3, District name2
IP address4, District name3

Filter out if portion of majority city < 50%

IP prefix1, District name1
IP prefix1, District name1
IP prefix1, District name2
IP prefix2, District name3

APNOM
Building IP Geolocation Database

Crawling

Header
IP address, Location1 (Metropolitan city)

Body
Location2 (District)

Extract
IP address and Location name

Group by /26 prefix
IP address1, District name1
IP address2, District name1
IP address3, District name2
IP address4, District name3

Filter out if portion of majority city < 50%
IP prefix1, District name1
IP prefix1, District name2
IP prefix2, District name3

APNOM
Building IP Geolocation Database

Crawling

HTML Document

Parsing

The bag of word for District

Extract
IP address and Location name

Group by /26 prefix

Filter out if portion of majority city < 50%

IP address1, District name1
IP address2, District name1
IP address3, District name2
IP address4, District name3

IP prefix1, District name1
IP prefix1, District name1
IP prefix1, District name2
IP prefix2, District name3
Building IP Geolocation Data

1. Crawling
2. Parsing
3. Extracting
4. Group by /26 prefix
5. Filter out if portion of majority city < 50%

The bag of word for District

Header
IP address, Location1 (Metropolitan city)

Body
Location2 (District)

IP address1, District name1
IP address2, District name1
IP address3, District name2
IP address4, District name3

IP prefix1, District name1
IP prefix1, District name1
IP prefix1, District name2
IP prefix2, District name3

IP prefix1, District name1
IP prefix2, District name3
Building IP Geolocation Data

Crawling

HTML Document

Parsing

The bag of word for District

Header

IP address, Location1 (Metropolitan city)

Body

Location2 (District)

Extracting

IP address, Location name

Grouping

IP address1, District name1
IP address2, District name1
IP address3, District name2
IP address4, District name3

Group by /26 prefix

IP prefix1, District name1
IP prefix1, District name1
IP prefix1, District name1
IP prefix1, District name2
IP prefix2, District name2
IP prefix2, District name3

Filter out if portion of majority city < 50%

IP prefix1, District name1
IP prefix2, District name1
IP prefix2, District name3
Building IP Geolocation Database

- Crawling
- Parsing
- Extracting
- Grouping
- Filtering

Online used market

HTML Document

The bag of word for District

Header
IP address, Location1 (Metropolitan city)

Body
Location2 (District)

IP address

Extraction
IP address name

Grouping
IP address1, District name1
IP address2, District name1
IP address3, District name2
IP address4, District name3

Filtering
IP prefix1, District name1
IP prefix1, District name1
IP prefix1, District name1
IP prefix2, District name2
IP prefix2, District name3

IP prefix1, District name1
IP prefix2, District name3

APNOM
The Data Summary

- 2015. 01. ~ 2017. 05.
- 195,937 Posts
  - 11.7 Gbytes
  - 34,740 Sellers
- 71,449 IP addresses
  - 44,916 /26 IP Prefixes

Ruliweb renew
# The IP Block (/26) Ratio per Databases

<table>
<thead>
<tr>
<th>Metropolitan city or Province</th>
<th>WHOIS</th>
<th>MaxMind</th>
<th>IP2Location</th>
<th>Ruliweb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyeonggi</td>
<td>44.01</td>
<td>0.90</td>
<td>3.84</td>
<td>17.05</td>
</tr>
<tr>
<td>Seoul</td>
<td>53.39</td>
<td>96.23</td>
<td>91.67</td>
<td>43.38</td>
</tr>
<tr>
<td>Busan</td>
<td>0.12</td>
<td>0.20</td>
<td>0.83</td>
<td>7.51</td>
</tr>
<tr>
<td>Gyeongnam</td>
<td>0.16</td>
<td>0.04</td>
<td>0.09</td>
<td>3.36</td>
</tr>
<tr>
<td>Incheon</td>
<td>0.41</td>
<td>0.44</td>
<td>1.16</td>
<td>5.50</td>
</tr>
<tr>
<td>Gyeongbuk</td>
<td>0.19</td>
<td>0.09</td>
<td>0.26</td>
<td>2.36</td>
</tr>
<tr>
<td>Daegu</td>
<td>0.18</td>
<td>0.44</td>
<td>0.25</td>
<td>5.83</td>
</tr>
<tr>
<td>Chungnam</td>
<td>0.03</td>
<td>0.08</td>
<td>0.03</td>
<td>1.81</td>
</tr>
<tr>
<td>Jeonbuk</td>
<td>0.07</td>
<td>0.21</td>
<td>0.07</td>
<td>1.89</td>
</tr>
<tr>
<td>Jeonnam</td>
<td>0.06</td>
<td>0.14</td>
<td>1.50</td>
<td>0.95</td>
</tr>
<tr>
<td>Chungbuk</td>
<td>0.00</td>
<td>0.12</td>
<td>0.01</td>
<td>1.62</td>
</tr>
<tr>
<td>Daejeon</td>
<td>0.99</td>
<td>0.24</td>
<td>0.12</td>
<td>2.91</td>
</tr>
<tr>
<td>Gwangju</td>
<td>0.16</td>
<td>0.44</td>
<td>0.02</td>
<td>1.89</td>
</tr>
<tr>
<td>Gangwon</td>
<td>0.00</td>
<td>0.20</td>
<td>0.06</td>
<td>1.40</td>
</tr>
<tr>
<td>Ulsan</td>
<td>0.13</td>
<td>0.14</td>
<td>0.05</td>
<td>1.85</td>
</tr>
<tr>
<td>Jeju</td>
<td>0.10</td>
<td>0.09</td>
<td>0.05</td>
<td>0.70</td>
</tr>
</tbody>
</table>

| The number of IP Blocks      | 1755976 | 1744364 | 1743633 | 44916 |
The IP Block (/26) Ratio per Databases

<table>
<thead>
<tr>
<th>Metropolitan city or Province</th>
<th>WHOIS</th>
<th>MaxMind</th>
<th>IP2Location</th>
<th>Ruliweb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyeonggi</td>
<td></td>
<td></td>
<td></td>
<td>17.05</td>
</tr>
<tr>
<td>Seoul</td>
<td></td>
<td></td>
<td></td>
<td>43.38</td>
</tr>
<tr>
<td>Busan</td>
<td></td>
<td></td>
<td></td>
<td>7.51</td>
</tr>
<tr>
<td>Gyeongnam</td>
<td></td>
<td></td>
<td></td>
<td>3.36</td>
</tr>
<tr>
<td>Incheon</td>
<td></td>
<td></td>
<td></td>
<td>5.50</td>
</tr>
<tr>
<td>Gyeongbuk</td>
<td></td>
<td></td>
<td></td>
<td>2.36</td>
</tr>
<tr>
<td>Daegu</td>
<td></td>
<td></td>
<td></td>
<td>5.83</td>
</tr>
<tr>
<td>Chungnam</td>
<td></td>
<td></td>
<td></td>
<td>1.81</td>
</tr>
<tr>
<td>Jeonbuk</td>
<td></td>
<td></td>
<td></td>
<td>1.89</td>
</tr>
<tr>
<td>Jeonnam</td>
<td></td>
<td></td>
<td></td>
<td>1.89</td>
</tr>
<tr>
<td>Chungbuk</td>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td>Daejeon</td>
<td></td>
<td></td>
<td></td>
<td>1.62</td>
</tr>
<tr>
<td>Gwangju</td>
<td></td>
<td></td>
<td></td>
<td>2.91</td>
</tr>
<tr>
<td>Gangwon</td>
<td></td>
<td></td>
<td></td>
<td>1.40</td>
</tr>
<tr>
<td>Ulsan</td>
<td></td>
<td></td>
<td></td>
<td>1.85</td>
</tr>
<tr>
<td>Jeju</td>
<td></td>
<td></td>
<td></td>
<td>0.70</td>
</tr>
<tr>
<td>The number of IP Blocks</td>
<td>1755976</td>
<td>1744364</td>
<td>1743633</td>
<td>44916</td>
</tr>
</tbody>
</table>

The number of IP Blocks:

- WHOIS: 1755976
- MaxMind: 1744364
- IP2Location: 1743633
- Ruliweb: 44916
The Accuracy Of The Crowd-Sourcing DB.

- The ground-truth data is hard to acquire
- Comparison of commercial and crowd-sourcing IP geolocation database
The Accuracy Of The Crowd-Sourcing DB.

- The ground-truth data is hard to acquire
- Comparison of commercial and crowd-sourcing IP geolocation database

39% difference less than 10 km
The Accuracy Of The Crowd-Sourcing DB.

- The ground-truth data is hard to acquire
- Comparison of commercial and crowd-sourcing IP geolocation database

39% difference over than 50 km

39% difference less than 10 km
The Distance from Seoul

- 80~90% of commercial database point to Seoul (Capital city of Korea)
- Seoul: The location of the ISP assigned the IP address
  - The reasons for why Seoul is so many in database
The Distance from Seoul

- 80~90% of commercial database point to Seoul (Capital city of Korea)
- Seoul: The location of the ISP assigned the IP address
  - The reasons for why Seoul is so many in database

60% of crowd-sourcing database point to non-Seoul (Gyeonggi) location
# Province/District-level IP Prefix (/26)

<table>
<thead>
<tr>
<th>Metropolitan city or Province</th>
<th>Province-level IP prefix # (A)</th>
<th>District-level IP prefix # (B)</th>
<th>B/A (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyeonggi</td>
<td>8584</td>
<td>4434</td>
<td>51.65</td>
</tr>
<tr>
<td>Seoul</td>
<td>21579</td>
<td>6490</td>
<td>30.08</td>
</tr>
<tr>
<td>Busan</td>
<td>3710</td>
<td>2498</td>
<td>67.33</td>
</tr>
<tr>
<td>Gyeongnam</td>
<td>1693</td>
<td>993</td>
<td>58.65</td>
</tr>
<tr>
<td>Incheon</td>
<td>2750</td>
<td>1976</td>
<td>71.85</td>
</tr>
<tr>
<td>Gyeongbuk</td>
<td>1153</td>
<td>684</td>
<td>59.32</td>
</tr>
<tr>
<td>Daegu</td>
<td>2888</td>
<td>1817</td>
<td>62.92</td>
</tr>
<tr>
<td>Chungnam</td>
<td>917</td>
<td>458</td>
<td>49.95</td>
</tr>
<tr>
<td>Jeonbuk</td>
<td>957</td>
<td>593</td>
<td>61.96</td>
</tr>
<tr>
<td>Jeonnam</td>
<td>480</td>
<td>137</td>
<td>28.54</td>
</tr>
<tr>
<td>Chungbuk</td>
<td>815</td>
<td>496</td>
<td>60.86</td>
</tr>
<tr>
<td>Daejeon</td>
<td>1475</td>
<td>930</td>
<td>63.05</td>
</tr>
<tr>
<td>Gwangju</td>
<td>944</td>
<td>581</td>
<td>61.55</td>
</tr>
<tr>
<td>Gangwon</td>
<td>686</td>
<td>331</td>
<td>48.25</td>
</tr>
<tr>
<td>Ulsan</td>
<td>939</td>
<td>434</td>
<td>46.22</td>
</tr>
<tr>
<td>Jeju</td>
<td>346</td>
<td>10</td>
<td>2.89</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>49916</strong></td>
<td><strong>22862</strong></td>
<td><strong>45.80</strong></td>
</tr>
</tbody>
</table>
## Province/District-level IP Prefix (/26)

<table>
<thead>
<tr>
<th>Metropolitan city or Province</th>
<th>Province-level IP prefix # (A)</th>
<th>District-level IP prefix # (B)</th>
<th>B/A (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyeonggi</td>
<td>8584</td>
<td>4434</td>
<td>51.65</td>
</tr>
<tr>
<td>Seoul</td>
<td></td>
<td>6490</td>
<td>30.08</td>
</tr>
<tr>
<td>Busan</td>
<td></td>
<td>2498</td>
<td>67.33</td>
</tr>
<tr>
<td>Gyeongnam</td>
<td></td>
<td>993</td>
<td>58.65</td>
</tr>
<tr>
<td>Incheon</td>
<td></td>
<td>926</td>
<td>71.85</td>
</tr>
<tr>
<td>Gyeongbuk</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daegu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chungnam</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jeonbuk</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jeonnam</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chungbuk</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daejeon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gwangju</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gangwon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ulsan</td>
<td>939</td>
<td>434</td>
<td>46.22</td>
</tr>
<tr>
<td>Jeju</td>
<td>346</td>
<td>10</td>
<td>2.89</td>
</tr>
<tr>
<td>Total</td>
<td>49916</td>
<td>22862</td>
<td>45.80</td>
</tr>
</tbody>
</table>

Can find 45% /26 IP prefixes in district name within metropolitan city or province.
New/Updated IP Prefix (/26) ~ 29 months

APNOMS 09/27/17
New/Updated IP Prefix (/26) ~ 29 months

Added an average of 1720 IP Prefix (/26) per month
New/Updated IP Prefix (/26) ~ 29 months

Added an average of 1720 IP Prefix (/26) per month

21% of post updates IP Prefix (/26) per month
Limitation

- Lack of ground-truth: Can not calculate the exact criteria of IP geolocation
  - Verification by comparing to commercial database
- Data volume limits: Less data accumulated in Ruliweb
- Limitations of application: Few community reveal IP address and location
  - Policies that do not apply in other online used markets (or community)
  - Difficult to use in mobile apps: NAT
Conclusion

- Create IP geolocation database using the Korea console gaming community data
- Although only 2.25% of the Korean IP blocks can be covered, 40% are mapped to areas other than the metropolitan area and have extensive coverage.
- Compared with commercial IP geolocation database, 40% /26 prefix is more than 50km difference
- Need to develop services using enhanced IP geolocation performance
Thank you.

Source Code:
https://github.com/munhyunsu/UsedMarketAnalysis