FUNG INSTITUTE
for ENGINEERING LEADERSHIP
UC BERKELEY COLLEGE OF ENGINEERING

Geocoding Patent Data

Kevin Johnson
University of California, Berkeley

College of Engineering
University of California, Berkeley

Fung Technical Report No. 2013.08.07
http://www.funginstitute.berkeley.edu/sites/default/files/GeocodingPatent.pdf

August 7, 2013

The Coleman Fung Institute for Engineering Leadership, launched in January 2010, prepares engineers and scientists – from students to seasoned professionals – with the multidisciplinary skills to lead enterprises of all scales, in industry, government and the nonprofit sector.

Headquartered in UC Berkeley's College of Engineering and built on the foundation laid by the College's Center for Entrepreneurship & Technology, the Fung Institute combines leadership coursework in technology innovation and management with intensive study in an area of industry specialization. This integrated knowledge cultivates leaders who can make insightful decisions with the confidence that comes from a synthesized understanding of technological, marketplace and operational implications.

Abstract: Innovation drives economic success. By analyzing U.S. patent data, we can understand what encourages and discourages innovation. We are particularly interested in the geographic location and mobility of inventors. In order to understand this, we must know where patents were invented. Latitude and longitude provide a consistent method for understanding location, but most patent applications only include the city and country of the inventor. We determine longitude and latitude given only this information – a process known as geocoding.

I. Introduction

The United States Patent and Trademark Office (USPTO) provides XML files containing information about all of the patents granted since 1976. We have built a program to parse, clean, and geocode this information. This allows us to provide access to the patent data as a simple API. Here, we examine the process required to geocode the patent data. There are two main challenges involved in this process. First, we identify ambiguous locations, matching them consistently with real location. Second, we assign latitude and longitude information to these locations, allowing them to be easily mapped.

II. The USPTO Data

II.1 Overview

There are over 12 million locations present in the patent files provided by the USPTO. Each location is split into up to five fields, depending on what information is available: `street address`, `city`, `state`, `country`, and `zipcode`. When non-unique locations are filtered out, there are roughly 900,000 unique locations to identify. However, not all of these unique locations are relevant.

II.2 Precision

It is rare for all five fields to be present; only 6.5% of locations have any information in the `street` or `zipcode` fields. Some locations contain street-level data in the `city` data field, making it difficult to understand exactly how precise the data are. However, we are confident that the vast majority of locations are only precise to the city level.

In addition, there is relatively little value in being accurate to a street level as opposed to the city level, since most analysis takes place at a city or state level. Therefore, we disregard all all street and zipcode information when geocoding the data.

II.3 Data Errors and Ambiguities

After disregarding the `street` and `zipcode` fields, there remain roughly 350,000 unique locations to analyze. These locations are poorly formatted and difficult to interpret for many reasons.

II.3.1 Accents

Accents are represented in many different ways in the data. Often, HTML entities such as `&#x212b;` are used. However, not all representations are so straightforward. For example, all of the following strings are intended to represent an angstrom (Å): `.ANG.`, `.circle.`, `&angst;`, `dot over (A)`, and `acute over (&#x212b;)`. These must be cleaned and converted into single characters.

II.3.2 Extraneous Information

Some foreign cities contain additional information that must be identified and dealt with consistently. For example, many cities in South Korea end include the suffix "-si", which indicates that the location is a city – as opposed to a county, which ends with the suffix "-gun". These suffixes are represented in a variety of ways, and should be interpreted consistently.

II.3.3 Incorrect Data

In some cases, data is recorded incorrectly on a consistent basis. For example, locations in the United Kingdom are often recorded with the country code "`EN`," and locations in Germany can be recorded as "`DT`."

II.3.4 Mislabeled Fields

In some cases, correct data for one field may be incorrectly assigned to a different field. For example, there are seven entries for a location with a `city` field of `San Francisco` and a `country` field of `CA`. There is no city named "San Francisco" in Canada; instead, "CA" was erroneously placed into the `country` field instead of the `state` field.

This problem is especially prevalent with foreign names. The `state` and `zipcode` fields only contain information for US locations. When such information exists for foreign locations, it is added to the `city` field – either in addition to or instead of the actual city.

II.3.5 Misspellings

All manner of creative and potentially ambiguous spellings can be found within the data. For example, all 31 of the following spellings are intended to refer to the “San Francisco” in California:

- `San Francais`
- `San Francesco`
- `San Francico`
- `San Francicso`
- `San Francis`
- `San Francisc`
- `San Franciscca`
- `San Franciscio`
- `San Francisco`
- `San Francisco County`
- `San Francisco,`
- `San Francisco, both of`
- `San Franciscos`
- `San Franciscso`
- `San Francisico`
- `San Franciso`
- `San Francisoc`
- `San Francisoco`
- `San Francsco`
- `San Francsico`
- `San Francsicso`
- `San Frandisco`
- `San Franicisco`
- `San Franicsco`
- `San Franisco`
- `San Franscico`
- `San Franscisco`
- `San Fransciso`
- `San Fransico`
- `San Fransicso`
- `San Fransisco`

This is by no means an exhaustive overview of the many ways that “San Francisco” can be spelled. Identifying and correcting these misspellings is an important challenge.

II.3.6 Romanization of Foreign Names

Converting location names from languages with different alphabets is a difficult task. To use a simple example, “Geoje” in South Korea is represented in six different ways: `Geojai-si`, `Geojae-si-Gyungnam`, `Geoje`, `Geoje-si`, and `Geoji-si`. The more complex the name, the more ways it can be converted into English, and the more difficult it is to identify what the name of the city is supposed to be from a given romanization.

III. Data Cleaning

Before performing any disambiguation work, we first focus on cleaning the raw location data. After cleaning, each location consists of a comma-separated string of the format “City, state, country”.

III.1 Accents

Because the format used to identify accents is so idiosyncratic, we individually identify and replace many accent representations using a handcrafted list of replacements. In addition, we automatically convert HTML entities to their corresponding Unicode characters.

III.2 Incorrect Data

We make some corrections for consistent error patterns that are difficult for our disambiguation method to decipher automatically. Though the list of corrections is small, this will be a major area of development going forward as we learn what kinds of locations are most difficult to interpret.

III.3 Mislabeled Fields

We deal with mislabeled states by using a format for cleaned locations that does not explicitly label the `state` and `country` fields. Though this slightly increases ambiguity, our disambiguation method is capable of interpreting the information. For our purposes, is better to have slightly ambiguous data than unambiguously false and misleading data. In addition, we automatically remove house numbers and postal code information from the `city` field.

III.4 Other Changes

In addition to the above, we perform a variety of minor alterations and corrections – pruning whitespace, removing extraneous symbols, and formatting the locations into a comma-separated format.

IV. Disambiguation

After cleaning the data, approximately 280,000 unique locations remain that must be disambiguated. For this process, we consulted with Jeffrey Oldham, an engineer at Google. He used an internal location disambiguation tool and gave us the results. For each location, Google’s API returns six fields:

- `city`, the full name of the city. For example, “San Francisco” or “Paris.”

- `region`, the name or two-letter abbreviation of a state, region, or other major institutional subdivision, as appropriate for the country. For example, "CA" or "Île-de-France."
- `country`, the two-letter country code corresponding to the country. For example, "US" or "FR."
- `latitude` and `longitude`, the latitude and longitude of the precise location found, depending on how much information is available. This is accurate to a street level if that information is provided in the input.
- `confidence`, a number representing how confident the disambiguation is in its result. -1 is returned if no result is found; otherwise, it ranges from 0 to 1.

Because the latitude and longitude data provided are more precise than we want them to be, we run the results of the disambiguation through the geocoding API again, giving each location a consistent latitude and longitude.

This process is ongoing, so detailed results are not yet available. However, preliminary results suggest that we will be able to geocode more than 99% of all locations with reasonable accuracy. The most recent version of our code can be found online at Github[1].

V. Acknowledgements

We would like to thank the United States Patent and Trademark Office for providing the data required to perform this research, and the National Science Foundation for funding this research under grant 1064182. Additionally, we are very grateful to Jeffrey Oldham for his assistance in disambiguating the locations.

References

[1] Kevin Johnson. `https://github.com/Vadskye/uspto_geocoding`, 2013.

[2] USPTO and Google. `http://www.google.com/googlebooks/uspto-patents-grants-text.html`, 2013.