Address Geocoding using Street Profiles for Local Search

Michael Peterman
Yellow Pages
1751 Rue Richardson
Montreal, Canada
michael.peterman@pj.ca

Omar Benomar
Yellow Pages
1751 Rue Richardson
Montreal, Canada
omar.benomar@pj.ca

Hacene Mechedou
Yellow Pages
1751 Rue Richardson
Montreal, Canada
hacene.mechedou@pj.ca

Felix-Herve Bachand
Yellow Pages
1751 Rue Richardson
Montreal, Canada
felixHerve.bachand@pj.ca

ABSTRACT

Geocoding is the process of converting addresses to geocoordinates. It is widely used in several fields such as public health to monitor socioeconomic inequalities for example or in Geographical Information Systems (GIS) to be able to use with its provided features. In this work, we describe a method to create an address geocoder from a free and open government street lines data source. The address geocoder transforms a street address into a location typically measured in latitude-longitude coordinates. The address geocoder is used in a search engine to relate spatial data to search results and improve accuracy.

Keywords

Geocoding; Street Profiles; Local Search

1. INTRODUCTION

Addresses are often used to convey geographical locations in searches. They are entered by users in a textual format. Search engines must resolve the user input and provide relevant results according to the address entered by the user. To do that, the search engine geocodes the address and uses the resulting geocoordinates to fetch relevant documents. Geocoding is not as simple as just mapping a geo-coordinate to an address. It involves several steps summarized by Zandbergen[2] which include parsing the input address, standardizing abbreviations, assigning each address element to a category, searching the reference data, assigning a score to each potential candidate, filtering out candidates based on minimum score, and delivering the best match. Geocoding quality is major concern when using address geocoders and is directly related to search results relevancy.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890484.

2. APPROACH

We use open data containing street segments in the form of geocoded points and having attributes such as length, start and last civic address, etc. This information is processed to create street profiles which are objects representing streets with civic numbers attached to them. The street profiles are then indexed by the search engine and used to geocode the user entered address.

2.1 National Road Network

The National Road Network (NRN) [1] is a free and open geospatial data source compiled and released by Natural Resources Canada. It comprises a database of poly-line objects representing road center-lines and their associated attributes in the Shape-file format. There is a unique poly-line entity for each segment of a road between intersections. In addition, divided roads with a median have a separate segment for the left and right side of the road. In addition to highways and city streets, NRN data includes many lane-ways and unnamed roads in remote areas, though these do not usually have any attributes. Private roads such as driveways and parking lots are not included. The attributes are rich and include information such as city, province, street name, street type (Avenue, Road, Highway) street direction (East/West), presence or absence of odd or even civic address, independent left and right-side civic address ranges, direction in which address number increases, and the first and last civic numbers of the road segment. The NRN data set includes $2.4M$ road segments.

2.2 Street Profiles

The raw data from NRN is not consumable by the search and need to be processed. To prepare the data to be used in a geocoding service, we eliminate unneeded data and aggregated attributes in the road segments to create a street profile, which can be consumed by a search indexer (SolR). The street profile is an ordered list of latitude-longitude coordinates and an interpolated civic number for each coordinate. For the profile, we use a subset of the coordinates in the NRN street segment, because the original coordinates have more detail than needed. An approximate spacing of 30m provides an adequate approximation of the original road segment, and the algorithm increases point density

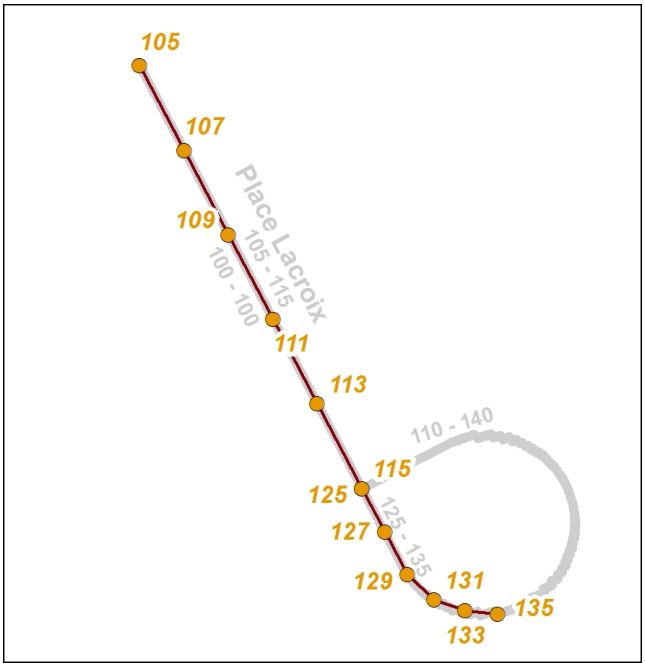

Figure 1: Place Lacroix street profile.

for sharply-curving road segments. Figure 1 shows an example of a street profile for Place Lacroix in Brossard, QC, Canada. The civic numbers represented in the figure are the endpoints of the NRN road segments.

We create a separate profile for the odd and even sides of each street, because the range of odd and even address numbers may not be overlapping. First, we select all road segments in an area, e.g, the city of Brossard, QC, Canada, which has 535 unique named roads in NRN. Each road name is processed separately, but there may be many road segments for a particular name.

We begin by building a graph using the subset of road segments of a street regardless of the civic numbers. Each end point of a segment represents graph vertex and the graph edges are the road segments. An edge is connected to another based on the distance between their endpoints. The graph is used to find the path of road segments for a chosen range of civic numbers (odd or even). Each segment has only the first and last civic numbers, so we attach civic numbers to it by interpolating according to its length and the coordinates in the original poly-line entity. We choose a start point by finding a vertex with degree of 1, i.e, a starting road segment. Then, we select the segment containing the desired civic address range (odd or even) and mark it as visited. We then find the nearest not visited road segment and continue the process until all segments with odd/even addresses have been processed. This creates a single path traversing all of road segments. In case of multiple potential start points for a complex road, we create all possible paths from each start point, and choose the shortest one, which provides the most direct path through the road segments. We generated about $495K$ street profile from the NRN data.

2.3 Geocoding Service

The street profiles are used to index each street into a SolR index with the following relevant information: street name, city, province, a field (*completeName*) including the street name, city name, and province name, used to query the index using similarity measures. We also index a list of latitude/longitude coordinates associated with odd and even civic numbers. The order of the civic numbers is based on creating a linear geographical representation of the street, as opposed to a list ordered on civic number values. This is due to the fact that real world data does not necessarily guarantee civic numbers to appear in order on a given street. That is, the geographical locations of civic numbers "1, 3, 5" could in fact form a straight line when ordered as "1, 5, 3".

When a user performs a search, we parse the geographical portion of the query, and look for a form such as : "<civic number> <street name> [city] [province]". We remove the civic number from the query, and use the remainder to query the SolR core to identify the appropriate street from the index described previously. We use the *completeName* field in order to find the best possible match indexed in our core.

Given a positive result from the SolR core, we use the street profile information to geocode the position of the civic number entered by the user. The list of civic numbers in the street profile are mapped to latitude/longitude pairs.

We then find the civic number in the list that has the closest value to the one found in the user's query, say q. If q is in the list of civic numbers in the street profile, we return its corresponding geocoordinates. In the case of q not being in the street profile, we select the civic number with the closest value, noted p. We then select the next neighbor to p, noted n. To select the appropriate n, we take either the civic number listed prior or after p, so as to ensure that $c > q > n$, or $c < q < n$. That is, we want to find two civic numbers surrounding q. Then, we use the civic numbers of p and n, and the value of q, to determine the relative distance of q along the line between p and n (e.g. if $p = 100$, $n = 200$, and $q = 40$, the distance d is 40% along the line between p and n). Finally, we determine the approximate latitude and longitude by interpolating the geocoordinates of q based on the latitude and longitudes of p and n and the computed distance d. Note that since we are working on a very limited scale in terms of distances, we use linear interpolation between the two geocoordinates of p and n.

3. CONCLUSION

We present an approach to geocode addresses in the context of a search engine using open data. The open data from the National Road Network is processed and aggregated to build street profiles. These street profiles contain information about the streets such as the name, city, province. They also include a list of civic numbers points with their corresponding geocoordinates. The street profiles are indexed in SolR and used to resolve the geocoordinates of address user queries by interpolation based on the existing points.

4. REFERENCES

[1] National road network. `http://geogratis.gc.ca/api/en/nrcan-rncan/ess-sst/c0d1f299-179c-47b2-bcd8-da1ba68a8032.html`.

[2] P. A. Zandbergen. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, 32(3):214 – 232, 2008. Discrete Global Grids.