EaserGeocoder: Integrative Geocoding with Machine Learning (Demo Paper)

Sina Rashidian
Stony Brook University
Stony Brook, NY, USA
sina.rashidian@stonybrook.edu

Shubham Kumar Jain
Stony Brook University
Stony Brook, NY, USA
jain.shubham@stonybrook.edu

Xinyu Dong
Stony Brook University
Stony Brook, NY, USA
xinyu.dong.1@stonybrook.edu

Fusheng Wang
Stony Brook University
Stony Brook, NY, USA
fusheng.wang@stonybrook.edu

ABSTRACT

Increased availability of large amounts of address data provides opportunities for data driven studies to improve decision making in business applications and support precision public health with high resolution geolocations. Geocoding large number of addresses is challenging due to high cost and often disclosure of sensitive data to vendors over the Web. Most geocoders take advantage of Web APIs which require sending private addresses over the Internet, which may not be an option for many applications with sensitive data including public health and geo-medicine. Meanwhile, the cost for geocoding massive number of addresses could be high and becomes a major hurdle for many users. To overcome these challenges, we developed an open source on-premise geocoding software *EaserGeocoder*, which uses a novel integrative geocoding model to achieve high accuracy through integrating multiple open data sources. *EaserGeocoder* takes advantage of machine learning based approaches to determine best answers from multiple data sources. *EaserGeocoder* can also be easily parallelized to achieve high scalability through parallelized search and distributed computing. *EaserGeocoder* is on a par with commercial geocoding systems, outperforms open source systems, and is available for free.

CCS CONCEPTS

• **Information systems** → **Geographic information systems**;

KEYWORDS

Geographic Information System, Geocoding, Text Searching

ACM Reference Format:

Sina Rashidian, Xinyu Dong, Shubham Kumar Jain, and Fusheng Wang. 2018. EaserGeocoder: Integrative Geocoding with Machine Learning (Demo Paper). In *26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '18)*, November 6–9, 2018, Seattle, WA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3274895.3274929>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGSPATIAL '18, November 6–9, 2018, Seattle, WA, USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5889-7/18/11.

<https://doi.org/10.1145/3274895.3274929>

1 INTRODUCTION

Spatial big data is exploding due to widely accessible geospatial data with government open data initiatives, geo-crowdsourcing, and social media. For example, Electronic Health Records (EHR) become essential resources for precision medicine and healthcare, which come with addresses of patients. Meanwhile, the increasing spatial resolution of spatial data provides unique opportunities for providing insights and adding new values to both industry and healthcare applications, such as precision public health [10]. This ultimately relies on geocoding massive amounts of address data, which could be sensitive in nature due to privacy.

For instance, addresses of patients are available in Statewide Planning and Research Cooperative System (SPARCS) [2] dataset, which provides an opportunity for large scale high resolution public health studies based on the addresses. The first step in order to perform high resolution spatial data analytics is to geocode patients' addresses to geolocations. There are also business applications with user addresses, which can provide location based analysis such as location intelligence, but the addresses can be confidential. Such applications face major challenges on geocoding, including:

- **Privacy.** For instance, in population or public health studies, patients' addresses are part of the Protected Health Information (PHI) data, which must comply to Health Insurance Portability and Accountability Act (HIPAA) [14]. Sharing patients' addresses over the Internet could pose major risk and lead to violations. Similarly, preserving the privacy of customer or user addresses is a major goal for business applications. This makes it difficult or infeasible to use Web API based geocoding systems as they require sharing data over the Internet.
- **High Cost.** Commercial geocoding is volume based in most cases. Geocoding 10 years' SPARCS data (~200M records) could cost around 100K USD. This is a major obstacle for many users, in particular, research projects or small business owners.
- **Scalability.** Commercial geocoders often have upper limits for daily and monthly usage and open source ones are not designed for high volume and have limited scalability. This is a major challenge for large number of addresses.
- **Noisy Data.** In reality, collected addressed data provided by users or patients often come with noise due to typos or miscommunication. Effective techniques to handle such noisy addresses are critical to achieve high accuracy for geocoding.

In summary, the current geocoding space lacks a cost-effective (or free) geocoder that can achieve high accuracy and scalability, while preserving data privacy. There is a major demand for an on-premise open scalable geocoding system that can support geocoding massive private addresses with high accuracy and low cost.

In this project, we intend to build a geocoding system based on multiple open and free reference address datasets, see Section 3. The main question would be: “Are these reference sources would be sufficient to support geocoding as they are designed for other purposes?” We believe that an integrative approach on different reference sources are capable to complement each other and achieve high accuracy. For this demonstration, we focused on geocoding for New York State, but all methods are expandable to other states.

This motivates us to develop an effective and scalable integrative geocoder for massive address datasets (EaserGeocoder). The EaserGeocoder takes advantage of multiple open data sources by having an integrative approach. It would search each dataset for the closest match, to generate a candidate list. We propose a novel machine learning based approach to automatically rank candidates to generate the final answer by learning from a public address data with a comprehensive set of address features. Our contributions are summarized as follows.

- We develop an integrative geocoding model capable of using multiple reference sources complement each other to achieve high accuracy and minimize error.
- We propose a novel machine learning based method for solving the problem of choosing the best answer by learning from a ground truth dataset to produce a prediction model based on Gradient Tree Boosting.
- We take advantage of parallel processing and in-memory caching to achieve high scalability.
- The system is open source and can be conveniently deployed on-premise without sharing of address data.

2 BACKGROUND AND OVERVIEW

2.1 Geocoding Reference Types

One of the most fundamental parts of any geocoding system is its references. These are data containing addresses with associated geolocations. There are three types of reference sources in general: linear-based, polygon-based and point-based. In summary, point-based ones are most accurate, as they contain one geolocation for each address. Polygon-based data are next. They consist of geolocations of points on the polygon of the building, city or town. For a single address, the centroid of polygon is used as the geolocation. At last, in linear-based references each line is represented as a set of geolocation points. For extracting geolocation of one specific address, interpolation methods are needed to estimate the location of the desired building, having higher spatial error than others [16, 19, 20].

2.2 Classic Geocoding Model

In classic geocoding model, one data source is imported into a database (usually TIGER [7]). For each address, address elements such as street name, city name, zip code, and etc after parsing and tokenizing are sent as a query to database to find closest match. The answer with highest text similarity is chosen in this case, and

geolocation regarding to that in database is returned. Limitations of this model are 1. Supporting only one *type* of reference sources 2. Searching based on only text similarity 3. Using only one data source, in most cases 4. Not considering spatial differences between data sources in case multiple are imported [19].

To overcome these limitations, we are proposing a new model called Integrative Geocoding Model [15]. First, we go through different references which are available for this project in section 3 and then elaborate our model in section 4 in details.

3 OPEN REFERENCE DATA

Open data initiatives and crowdsourcing have produced highly valuable geospatial data, including open government data, tax parcel data, community contributed data, and state specific data. We will focus on following data sources, but our approach can be generalized to incorporate more data sources if available. Note that we only support geocoding for US addresses at the moment, and for this demo we targeted New York state.

- **TIGER:** Topologically Integrated Geographic Encoding and Referencing provides a US national data source consists of spatial extracts from the Census Bureau’s MAF/TIGER database [7]. This street network map widely used in geocoding systems, since it has a vast coverage and data is reliable [17].
- **Tax Parcels:** It is a legal description and division of land with the main purpose of providing a precise unit of taxation [11]. It contains exact addresses with one point representing for each address. Major drawbacks of this data source are missing many buildings addresses and partial information for records [13, 20].
- **Community Contributed Datasets:** We considered Open Street Map (OSM)[6], and OpenAddresses (OA)[5]. Data cleaning is a critical step for using such data sources as they come with much noise, formatting issues and inconsistency.
- **Street and Address Maintenance(SAM) Program (NY):** As a result of NY state efforts for collecting street and address point data for Next-Generation 911 emergency systems this dataset is created. Currently, it supports 51 out of 62 counties in NY. [4]

As these data sources are not totally disjoint and could have a lot of common addresses, we analyzed their contribution in our geocoding system. Turns out, OSM data addresses are covered by others. Therefore, we eliminated its usage in our system.

4 INTEGRATIVE GEOCODING

Our approach consist of 5 steps summarized in below:

- (1) Parsing: Input address is broken into a set of features by tokenizing, cleaning and standardization.
- (2) Customized Query: For each individual reference data source a customized query will be generated.
- (3) Database search: Queries are submitted to the database to search across all data sources.
- (4) Relaxed Search: In case previous step does not return sufficient answers, search is repeated with less restrictions.
- (5) Intelligent Answer Selection: We choose the best answer from the candidates, using machine learning techniques.

Figure 1 illustrates all steps above. The training phase is the prerequisite for step 5 and it is needed to be done in advance, as it is shown at the top in figure with dashed lines. Furthermore, we are

Figure 1: Integrative Geocoding Model

employing parallel processing by taking advantage of multi-thread programming to enhance performance. The algorithm is described thoroughly at [15], in this paper we mention the machine learning section briefly and focus on how we want to demonstrate this work.

4.1 Customized Query Generation & Relaxation

Based on characteristics of each data source, we generate specific query for that source. There are many cases that a feature in one data source does not always exist (e.g. zip code in Tax Parcels). In that case, adding that feature as a mandatory filter to the search process, could even impair searching quality. EaserGeocoder generates customized queries based on features of each reference, therefore it reaches answers faster and reduces number of reattempts to get a proper response. For each data source, there are a couple of relaxed queries custom designed, in case there were not any answers with mandatory conditions they are used for reattempts.

4.2 Intelligent Answer Selection - A Machine Learning Approach

Selecting best answer is challenging as different sources do not share same set of features and free sources tend to be noisy and incomplete even in features they supposed to have. Common traditional approach for answer selection is based on assigning scores with pre-defined weights to same address elements between input query and candidates [12, 18]. Although, this method makes common sense, it would not be optimized where there are a variety of data sources with different set of features.

Training Dataset We collected 19,095 random samples of physician addresses in New York State provided by The Centers for Medicare and Medicaid Services (CMS)[1]. The “ground truth” of geolocations for this dataset is generated from Google geocoding API [3]. Given this dataset to our system, we generate four geocoding candidates for each address using four references and calculate the distance between each candidate and the “ground truth” geolocations. We take 400 meters as an error threshold to classify each candidate as “correct” or “wrong”. Then we defined a feature vector estimated by comparing input address and matching record for each candidate, to form the features for the classifier [15].

Features The matching record in each reference is compared with the original address from the ground truth dataset, and generates a set of features used for the classifier. Distance function uses *Absolute Distance of Building Numbers*, *Normalized Edit Distance for Street, City and County Names*, *Equality and Availability of Zip Code*, and *Equality Tests for Base Number, Prefix and Postfix*. In addition, the following features are considered too: the reference type (TIGER, OA, Parcels and SAM), input zip code, and response class.

Gradient Tree Boosting Based Classifier We use Gradient Tree Boosting based Classifier to train a prediction model. Gradient Boosting is a machine learning technique which generally uses weighted sum of multiple weak prediction models (decision trees) to predict on new datasets. Since relationships between parameters in this problem is nonlinear, decision trees as fundamental prediction models are a perfect fit. After training with the ground truth dataset, the classifier learned about one hundred and fifty small decision trees. We took advantage of XGBoost library dedicated to gradient boosting algorithms designed to be highly efficient, flexible and portable [9] for this purpose. After training, given an input address and the set of candidates extracted from multiple references, a feature vector (similar to training part) will be sent to the classifier. A probability is assigned to each candidate based on underlying decision trees. The final result would be candidate with the highest probability. The classification step is the last step of the geocoding pipeline, Figure 1. Accuracy of this model is measured and discussed in [15] in details.

5 DEMONSTRATION

We will demonstrate the EaserGeocoder software (Section 5.1), example integrative geocoding process (Section 5.2), and comparison with other geocoding systems (Section 5.3).

5.1 Software

EaserGeocoder comes with both a Web based application [8] and an offline standalone application. We will demonstrate both of them.

Web Based Application. This Web based application supports single address geocoding and batch geocoding. For single address geocoding, given address components, EaserGeocoder will return the geolocation of the address and centroids of city, zip code and county, if applicable, as shown in Figure 2. For batch processing, a file containing addresses will be uploaded, and geocoded geolocations will be returned as a file [8].

Standalone Application. The core functionalities of EaserGeocoder are implemented in Java and released as a jar file, which can be used for batch processing. It is OS independent and tested on Windows, Linux and MacOS. EaserGeocoder uses master-slave based parallelization for achieving scalability. The software can be downloaded at its website [8]. The program takes an input CSV file containing addresses to be geocoded. In this demo, we will use addresses in New York state only. The program can be customized to run only a small sample for testing, and a number of threads can be specified for parallelization. For example, a thread number set to two times of number of CPU cores provides good scalability based on our experiments.

(a) Geocoding with zip code and city name

(b) Geocoding with minimum amount of data

Figure 2: Single Address Geocoding - EaserGeocoder

Software Resources Running the program requires some resources including geocoding references (Section 3). These references are parsed, cleaned and imported to Apache Solr and are ready to use. The estimated size of combined reference data sources is about 7GB for New York State. We also take advantage of TIGER gazetteer data [7], decision trees learned and information about building number formats in different counties, zip codes and cities which were preprocessed in advance. All are accessible through the website [8].

5.2 Anatomy of Integrative Geocoding

We will demonstrate the integrative geocoding process through step-by-step illustration, including intermediate results. For example, we will show how decision trees help to choose the best answer from candidates. We will prepare following information for the demonstration: 1) Results from each data source; 2) distance vectors of candidate results and input address; 3) customized queries for each data source; and 3) decision trees scoring. Such information will help to differentiate EaserGeocoder from traditional geocoding systems.

5.3 Comparison

A comparison functionality is integrated into the software to evaluate EaserGeocoder's performance. The demonstration will show a list of addresses coming with ground truths. It reports how many addresses are geocoded successfully, with information on match rate accuracy, running time, average distance to given ground truth (spatial accuracy), number of perfect answers (answers exactly match a record in one of our data sources) the accuracy and spatial distance. Moreover, we report which address belongs to which data source, and how many total addresses come from each data source.

In the demonstration, we will also compare EaserGeocoder with other geocoding systems such as Google, MapQuest and Here.

6 ACKNOWLEDGEMENTS

This research is supported in part by grants from National Science Foundation ACI 1443054 and IIS 1350885.

REFERENCES

- [1] 2016. Physician and Other Supplier Data CY 2013. Retrieved August 11, 2016 from <https://www.cms.gov/>
- [2] 2016. SPARCS. <https://www.health.ny.gov/statistics/sparcs/>
- [3] 2017. Google Maps Geocoding API. <https://developers.google.com/maps/>
- [4] 2017. NYS GIS Clearinghouse - NYS Address Points. Retrieved May 17, 2017 from <http://gis.ny.gov/gisdata/inventories/details.cfm?DSID=921>
- [5] 2017. OpenAddresses. Retrieved Feb, 2017 from <http://results.openaddresses.io/>
- [6] 2017. OpenStreetMap Nominatim. Retrieved May, 2017 from <http://nominatim.openstreetmap.org>
- [7] 2017. TIGER Products - Geography - U.S. Census Bureau. Retrieved May 26, 2017 from <https://www.census.gov/geo/maps-data/tiger.html>
- [8] 2018. EaserGeocoder. <http://bmdb.cs.stonybrook.edu/easergeocoder/>
- [9] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.
- [10] Xin Chen and Fusheng Wang. 2016. Integrative Spatial Data Analytics for Public Health Studies of New York State. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 391.
- [11] New York State Geographic Information Systems Clearinghouse. 2014. NYS GIS Clearinghouse - NYS Tax Parcels. Retrieved Feb 20, 2017 from <http://gis.ny.gov/gisdata/inventories/details.cfm?DSID=1300>
- [12] Daniel W Goldberg and Myles G Cockburn. 2010. Improving geocode accuracy with candidate selection criteria. *Transactions in GIS* 14, s1 (2010), 149–176.
- [13] Geoffrey M Jacques. 2012. A research agenda: does geocoding positional error matter in health GIS studies? *Spatial and spatio-temporal epidemiology* 3, 1 (2012), 7–16.
- [14] U.S. Department of Health & Human Services. 2015. Health Information Privacy | HHS.gov. <https://www.hhs.gov/hipaa/>
- [15] Sina Rashidian, Xinyu Dong, Amogh Avadhani, Prachi Poddar, and Fusheng Wang. 2017. Effective Scalable and Integrative Geocoding for Massive Address Datasets. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 26.
- [16] Gerard Rushton, Marc P Armstrong, Josephine Gittler, Barry R Greene, Claire E Pavlik, Michele M West, and Dale L Zimmerman. 2006. Geocoding in cancer research: a review. *American journal of preventive medicine* 30, 2 (2006), S16–S24.
- [17] Xuan Shi, Bowei Xue, and Imam M Xierali. 2016. Identifying the Uncertainty in Physician Practice Location through Spatial Analytics and Text Mining. *International Journal of Environmental Research and Public Health* 13, 9 (2016), 930.
- [18] Duck-Hye Yang, Lucy Mackey Bilaver, Oscar Hayes, and Robert Goerge. 2004. Improving geocoding practices: evaluation of geocoding tools. *Journal of medical systems* 28, 4 (2004), 361–370.
- [19] Paul A Zandbergen. 2008. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32, 3 (2008), 214–232.
- [20] Paul A Zandbergen. 2009. Geocoding quality and implications for spatial analysis. *Geography Compass* 3, 2 (2009), 647–680.