An Exploratory Study of Address Geocoding Techniques with a Focus on Solutioning for the Indian Address Problem

[1]Harshit Gupta, [2]Vineeth Raj, [3]Dr. T. Manoranjitham
SRMIST

Abstract

Accurate geocoding of natural language addresses is essential to any industry dealing with logistics services, like the postal service, e-commerce, etc., and there are ongoing research efforts in address geocoding by on-demand service providers. However, currently, existing methods assume the existence of a standard addressing scheme which can be deconstructed and identified.To build automated, scalable and self-learning systems that are capable of analyzing information contained in natural language addresses the systems need, along with user input, verification, and updating of learning models. This does not work well in places which do not have a standard addressing scheme or where addresses can not be determined with the help of these addressing schemes due to the massive increase in population density and urbanization of countries. A number of techniques have been proposed by various authors to provide a solution for address geocoding of natural language addresses that have varying accuracy. There have also been proposed solutions as dynamic models that grow to accommodate changing landscapes.We also discuss relevant issues with these systems such as data collection, evaluation metrics and bench-marking to find accurate location. After going through these proposed solutions and current implementations, we conclude by studying the applicability of these solutions for Indian addresses, their limitations and provide possible directions for future work.

I. INTRODUCTION

As we all are going towards digitization we are experiencing a lot of changes around us with the help of the advancement of technology and infrastructure. All our necessary things are just click away from us, all thanks to on-demand services which are seamlessly helping the Indian customers and consumers to order get the services to their doorstep, the main challenge these on-demand providers facing in fulfilling the requirements and demand of there customers is providing their services on time which lags in India in comparison to other countries. The main reason of these issues is the lack of proper addressing scheme in India, however, these problems can partiality resolved with help of GPS but it does not guarantee exact location when the destination is in big buildings and not in a proper GPS coverage. Here comes the role of Addressing scheme which can truly solve this massive problem of the Supply Chain industry. The goal of this paper is to analyze the different addressing schemes used by other countries and several other solutions provided by several pieces of research and to provide suggestions and improvements in Indian Addressing System.

II. CHINESE-GEOCODING METHOD USING FUZZY TECHNOLOGY

Zhang X, et. al. [1], discuss the challenges faced when performing address geocoding for Chinese urban addresses and present a potential solution with means to deal with the challenges. The issues raised with Chinese addresses, as described, are as follows. *Firstly*, with the way in which the addresses are represented; the Chinese addresses do not have any standard delimiters between words making it difficult to tokenize the address into the constituent elements like Area, City, Province, State, etc. *Secondly*, the nature of Chinese addresses as “chaotic”, reasoning that there is no standard scheme for naming or orderly numbering which cause standard methods of address geocoding to fail. *Thirdly*, the constituent elements of an address, like the city, state, province, etc., have fixed naming authorities, and these are distributed amongst different government agencies; who have failed to collaborate and come up with means to standardize the addresses.

The paper reviews the existing address geocoding systems and their in-feasibility for usage with ex- isting model of Chinese addresses. GBF/DIME and TIGER/Line geocoding solutions were notfeasible since they were built with assumptions of American addressing scheme in mind where street interpolation can be easily done due to standardization. An ESRI geocoding solution would not work since the address model of ESRI assumes the format of address to be “house number + street name“ but many Chinese addresses, especially in urban areas are of the format “building number + residential complex”. The paper also studies a Japanese address geocoding solution described by Sagara T., et al.[2], where a combination of tree and trie structure to achieve higher flexibility and efficiency, a technique known as “inverted in- dex” or “inverted document”. The paper also evaluates a geocoding solution developed by a Chinese GIS company which uses regular expressions and keyword matching; which was not only significantly slower but also did not work well with non-standard addresses.

Zhang X, et. al., propose a geocoding solution building upon the strengths of studied solutions. The solution is similar to Sagara T., et al.,[2] while extending the trie structure to include commonly used aliases, and adopt- ing ESRI’s address model to make them both work in a single system consisting of a service layer containing the address matching module and a management layer containing the address management module.

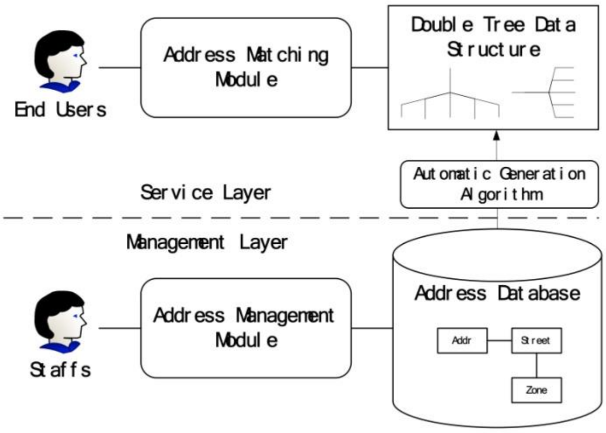


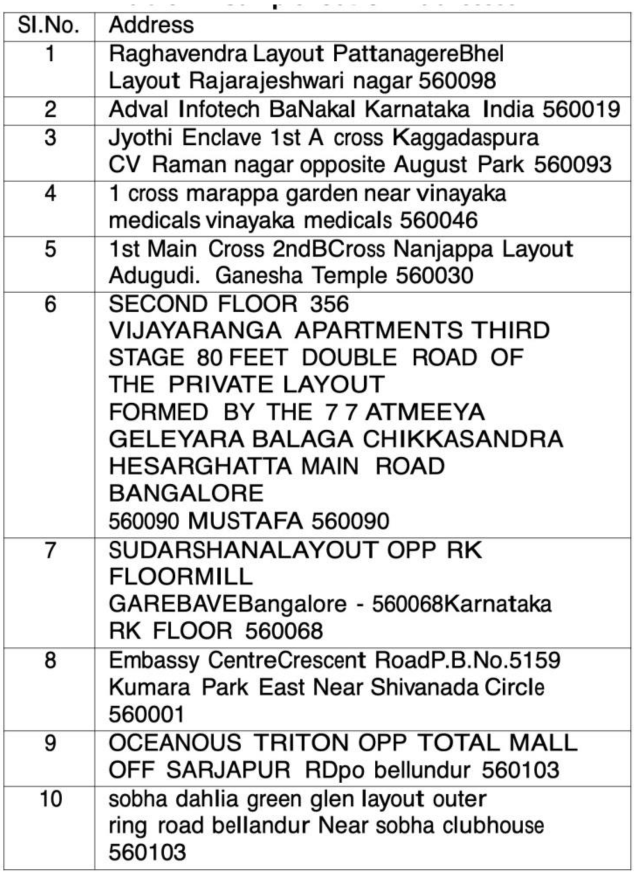
Fig. 1. Proposed solution for Chinese address Geocoding (Courtesy of Zhang X, et. al. [1])

III. GEOGRAPHICAL ADDRESS CLASSIFICATION WITHOUT USING GEOLOCATION COORDINATES

T. R. Babu, et. al. [3], propose a solution that was demonstrated in a major e-commerce organization, that can perform geographical address classification without the use of geolocation co-ordinates. The paper surveys related work and upon not finding any similar solution in literature, details an accurate method that can be used to classify addresses belonging to a set of predefined sub-regions. During online shopping, the input criterion for addresses is lax to provide flexibility, avoid any inconvenience that may be caused to the user, and to

provide an overall better experience to the customer. This however, results in non-standard addresses being entered in the system at places where there are already no prefixed addressing schemes which increase the cost of identification of the user, misrouting of shipments, etc.

There is a high variability in the number of terms that the user enters which could be attributed to the varying confidence level the user has with their own address. Some cases where the user enters a long ad- dress, include details like their availability and timings at the address, directions to reach the location, etc. Two solutions were proposed based on the current work where the first solution relies on the experience of the Field Executive (FE) of the logistics team in order for identification of addresses while the second solution, which is detailed by the paper is unsupervised classification based on text similarity.



Sl.No.	Address
1	Raghavendra Layout PattanagereBhel Layout Rajarajeshwari nagar 560098
2	Adval Infotech BaNakal Karnataka India 560019
3	Jyothi Enclave 1st A cross Kaggadaspura CV Raman nagar opposite August Park 560093
4	1 cross marappa garden near vinayaka medicals vinayaka medicals 560046
5	1st Main Cross 2ndBCross Nanjappa Layout Adugudi. Ganesha Temple 560030
6	SECOND FLOOR 356 VIJAYARANGA APARTMENTS THIRD STAGE 80 FEET DOUBLE ROAD OF THE PRIVATE LAYOUT FORMED BY THE 7 7 ATMEEYA GELEYARA BALAGA CHIKKASANDRA HESARGHATTA MAIN ROAD BANGALORE 560090 MUSTAFA 560090
7	SUDARSHANALAYOUT OPP RK FLOORMILL GAREBAVEBangalore - 560068Karnataka RK FLOOR 560068
8	Embassy CentreCrescent RoadP.B.No.5159 Kumara Park East Near Shivanada Circle 560001
9	OCEANOUS TRITON OPP TOTAL MALL OFF SARJAPUR RDpo bellundur 560103
10	sobha dahlia green glen layout outer ring road bellandur Near sobha clubhouse 560103

Fig. 2. A Sample set of addresses used to demonstrate variability and challenges (Courtesy of T.R. Babu, et. al. [3])

The paper goes on to provide a detail of the steps of the proposed solution, that is as follows:

1) Pre-Processing
2) Approaches to deal with large data sets
3) Solutions
4) Experimental Evaluation

Preprocessing is done in order to make the available data suitable for further processing. This involves steps like data cleanup, probabilistic separation, integration of domain knowledge, clustering, classification mod- els to tag fraudulent addresses (termed monkey-typed addresses, detailed by T.R. Babu, et. al., in [4]), and generating N-grams. Two approaches are suggested to deal with large dataset of addresses; Data reduction using N-grams and frequency patterns, and Dimension- ality reduction in order to reduce the cardinality of N- grams where the N-grams can be efficiently reduced to a unigram with no information loss (eg., (Vega, Mall) to (VegaMall)). In Solutions, and Experimental evaluation, T. R. Babu, et. al. [3], detail the difference in solutions obtained using supervised and unsupervised learning approaches, where similar results were obtained. A semi-supervised dataset generated with ensemble clas- sifiers and nearest neighbor approach are used to grow the data to 2.5 and 4.5 times the original dataset with improvement in accuracy.

IV. HISTORICAL COLLABORATIVE GEOCODING

Cura R and Dumenieu B describes how historical ad- dresses can be determined with the help of huge data set availability of addresses and placing these addresses on maps to identify and localize the address with doesn't follow the modern addressing scheme divisions like (house number, place, district, city, state, etc.)[5] thus provides a helpful measure to classify and determine the addresses in the old countries which colonized many centuries before and following the same pattern. With the help of old maps called gazetteers, they are trying to Geocode the address and matching can be done using customisable criteria which include several dimensions

- fuzzy semantic
- fuzzy temporal
- scale
- spatial precision

The research by Cura R, et. al [5], solves for a very common problem in India that is Indian Addressing scheme is not planned and using the ancient addressing schemes and name in the location which established decades ago, at the same time some areas are well planned with proper addressing schemes and location thus dividing the addresses into two different groups. Using this methodology can be beneficial for the areas which established decades ago but at the same time it does not provide the solution for unplanned cities and places.

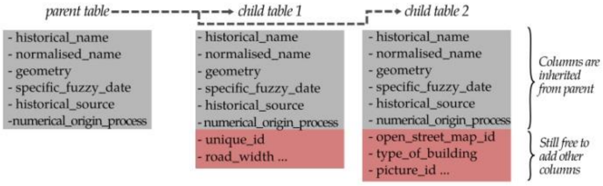


Fig. 3. The Process of Historical data Geocoding using the Table Inheritance Mechanism (Courtesy of Cura R, et. al. [5])

V. IMPACT OF IMPROPER ADDRESS SCHEME IN INDIA

Around the earth out of 7 billion people residing in households around 75 percent houses doesn't have a proper addressing scheme [7], making several facili- ties which requires the involvement of addresses got affected or delayed. These services include:

- Logistics and Transportation Services
- On-Demand Survives
- Address Verification
- Emergency Services

Logistics Services are in high use these days and deliveries are being made every single day in urban areas, increasing the requirement of a proper addressing scheme. Even after having a pincode system it is insufficient to reach or map the location due to the density of population in that particular area making it difficult to reach the location. Inability to locate address makes it difficult for the delivery executive to deliver the shipments and orders at the provided address. In Supply Chain industry is divided into mainly three parts which include pickup, transshipment and delivery. The delivery part of the supply chain is being done in the

Last Mile which is responsible for the delivery of the shipment. Around 30 percent of total supply chain cost is contributed by the Last Mile.

In the absence of geolocation, the addresses are sorted with the help of pincodes and since these pin- codes are not well structured and sorting is being done without the internal knowledge of these pincodes thus two delivery locations can have a huge distance making it much more expensive for the logistics vendor to deliver the parcel or order to the required location in specified time window[6].

This study helps in understanding the problems that are being faced in different services uses the addresses provided by the provider and how it is impacting the cost and economy to provide that service at the doorstep. Which can be concluded as there is a need for a better addressing scheme or a proper Geo Loca- tion method to lower the complexity of supply chain industry.

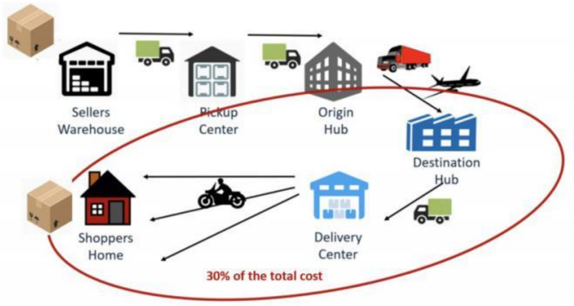


Fig. 4. Role of the Last Mile in the Supply Chain [6]

- Considering a given address by a user for spe- cific service then there should be some methods to classify the address as belonging to smaller sub regions accurately without using geolocation without the help of latitude and longitude, thus helping in providing the service with the specified time window.

VI. CONCLUSION AND FUTURE WORKS

- Various research has gone into the feasibility and implementation of geocoding solutions across var- ious parts of the world (eg. USA [8], China [1], Korea, Japan [2]).
- Without the existence of a well-defined addressing scheme (eg. East Asian Countries), the implemen- tation of geocoding solutions is difficult and spe- cific to the address implementation of the country.
- Any industry involving a supply chain needs to be capable of address geocoding. This can be man- ually done, but will be effort intensive, and may not be efficient, accurate, or scale-able; which is an essential requirement of a big logistics supplier.
- Existing geocoding solutions that have been im- plemented are not suitable for use with Indian addresses due to the lack of a standardized format.
  - This problem is further aggravated due to wrongly spelt words, wrong zipcodes caused due to a non-standard format, or missed spaces, wrong hierarchy, incomplete ad- dresses and a general lack of structure by virtue of a large diversity in culture and literacy.
- Even though there exist solutions which converts the given addresses into specified geo locations but have their own limitations while decoding the Indian addresses in terms of accuracy, cost, and vendor lock-in.

VII.REFERENCES

[1] X. Zhang, H. Ma, and Q. Li, "An address geocoding solution for Chinese cities", in Geoinformatics 2006: Geospatial Infor- mation Science, 2006.

[2] Sagara T., Arikawa M., Sakauchi M., "Spatial Document Man- agement System Using Spatial Data Fusion", International Con- ference on Information Integration and Web-based Applications
& Services (IIWAS2001), pp 399-409, Linz, Austria, 2001.

[3] T. R. Babu, A. Chatterjee, S. Khandeparker, A. V. Subhash, and S. Gupta, "Geographical address classification without using geolocation coordinates", in Proceedings of the 9th Workshop on Geographic Information Retrieval - GIR 15, 2015.

[4] T.R. Babu, V. Kakkar, "Address Fraud: Monkey Typed Address
Classification for e-Commerce Applications", SIGIR-Ecom,
2017

[5] Cura R, "Historical collaborative geocoding", SPRS Int. J. Geo- Inf. 2018, 7(7), p 262.

[6] Dr. Santanu Bhattacharya, Sai Sri Sathya, Dr. Kabir Rustogi and Dr. Ramesh Raskar, "Economic Impact of Discoverability of Localities and Addresses in India", arXiv:1802.04625v1 [cs.CY] 13 Feb 2018

[7] R. Feng., "Startup What3words Aims To Give Billions Of
People One Thing They Dont Have", Forbes, 2016.

[8] P. A. Zandbergen, "A comparison of address point, parcel and street geocoding techniques", Comput. Environ. Urban Syst.,
2008.