# Unveiling Locations in Geo-spatial Documents

Gyan Ranjan*, Juong-Sik Lee†, Deepti Chafekar† and Umesh Chandra†
*University of Minnesota, Twin Cities, MN and †Nokia Research Center, Palo Alto, CA
*E-mail: granjan@cs.umn.edu, {juong-sik.lee, deepti.chafekar, umesh.1.chandra}@nokia.com*

## Abstract

Resolving geo-identities of addresses in emerging economies[1] where users rely primarily on short messaging as the means of querying, poses several daunting challenges: lack of proper addressing schemes, non-availability of cartographic information and non-standardized nomenclature of geo-spatial entities such as streets and avenues, to name a few.

In this work, we propose a simple and elegant approach to solve this problem for emerging economies. By treating address texts as short documents and exploiting latent proximity information contained in them — for example, landmark like references, similarity of address texts etc — we transform the problem of resolving geo-identity to a search problem on short annotated *geo-spatial* documents, collected through extensive survey of six cities in India. Our solution spans all the phases of building a geo-identity resolution system, even though our emphasis is on the collection and organization of the corpus to facilitate a search engine backend for the task. Through experimentation based on a representative test set collected from the real world, we demonstrate how this approach achieves over 94% accuracy in resolution and an order of magnitude reduction in system state (memory) with nearly zero false-negatives - a significant improvement over the state of the art in emerging markets.

## 1. INTRODUCTION

Addresses provide a universal way of referencing locations in a human understandable format. For example, we are used to querying for locations and routes on online search engines by providing address texts as queries. Moreover, a person is likely able to provide, and perhaps guess, approximate address texts, either of his/her current location or of a region/point of interest (POI), such as a restaurant or a hospital, and thus query for information in the vicinity with ease. This is especially pertinent to emerging economies (such India and other parts of Asia-Pacific and Africa), where advanced map-based user interfaces and GPS chipsets have relatively lower penetration on end user devices and short messaging services are the dominant means of communication. Understanding addresses is, therefore, key to building effective and scalable location aware services similar to *locate-me*, point-to-point navigation, location-aware recommendation systems, in emerging markets.

We refer to the general problem of mapping an address text to a pair of latitude-longitude coordinates as *geo-coding* of address texts. We now illustrate the challenges for this task in the emerging markets with the help of a real world example:

EXAMPLE 1. *Barista[2], 3 C's Cinema Road, Near Alankar's Theater, Lajpat Nagar III, New Delhi, India.*

When entered as a query to three different search engines viz. Google, OVI and Yahoo, this address results in a *no match*. Whereas OVI and Yahoo search engines fare no better with altered queries, Google (`maps.google.com`), provides critical insight into the possible causes of failure as described below.

Location A in figure 1(b) refers to the original query address that resulted in a failure. On reducing the query text to simply *Barista, New Delhi*, to find all possible locations for this cafe chain within the city of New Delhi, we obtain a list of addresses suggested by Google, one of which is: *Barista-Alankar 3 Cs, Lajpat Nagar III, New Delhi, Delhi, India*. This location is shown on the map as location B. Moreover, proximous to location B, Google resolves another address text: *3 Cs Cinema, Feroze Gandhi Marg[3], Lajpat Nagar III, Lajpat Nagar, New Delhi, Delhi, India* (shown in the figure as location C). To the human eye it is obvious that the n fact, the query is a combination of the two address. Therefore, the *no match* result can, in some sense, be thought of as a *false negative*.

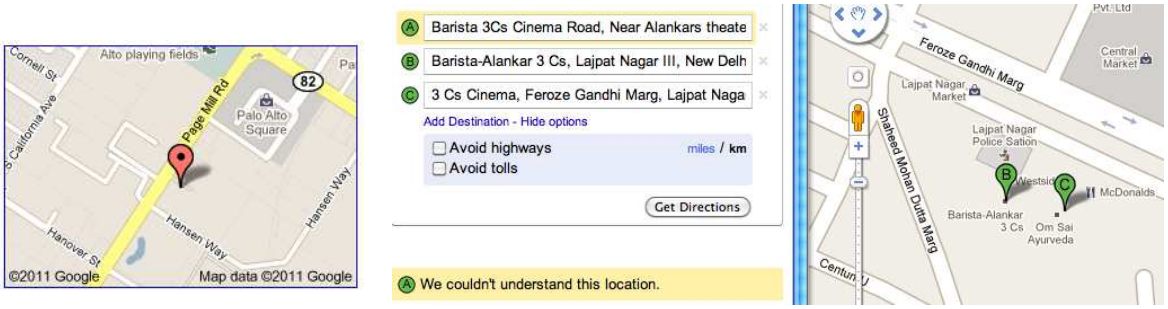
We infer, on the basis of the evidence at hand, that Google, and probably the other search engines, match address queries against each address in the corpus individually to resolve geo-spatial identity. This strategy might work well in the context of United States and Europe, where cities are often divided into relatively smaller blocs identified by the intersections of street and avenues. This clearly does not apply to India and other countries in general. In contrast we find loosely defined street and *area* name [4] combinations

[1]Emerging markets/economies: An umbrella term used for countries outside of North America and Europe

[2]A popular chain of cafes in New Delhi.

[3]The word *Marg* means road. Therefore, the street name here is Feroze Gandhi Road

[4]Loosely defined sub-divisions of a city.




(a) A successful query, (b) And a failed one.

**Figure 1: An example: Resolving geo-identities in developed vs. emerging markets.**

in most address texts. A dictionary of standardized street and/or area names is hard to come by. Moreover, whereas a street may be officially called *Feroze Gandhi Marg* (say), the colloquially understood name is often associated with a locally recognized landmark (in this case *3 C's cinema* theater which lends its name to the street as *3 C's Cinema Road*). Therefore, even when official cartographic data is available, geo-identity resolution is likely to fail if the system in question does not take into account such colloquial references.

In view of these observations, we provide a first principles based approach — from collecting, annotating and organizing address information in emerging markets — to designing a geo-identity resolution system that overcomes the apparent handicaps of current systems. Our key insight is that unstructured address texts, can be organized in a simple and elegant way to avoid the *false negatives*. Our solution involves collating geo-spatially annotated address texts (collected in a planned manner as explained in detail later) to form a corpus of geo-spatial documents. Through this we generalize the geo-identity resolution problem to a problem in the document search space which, in turn, allows us to exploit the versatility and power of commercial search engines. The proposed solution achieves over 94% accuracy with nearly zero *false negatives*. We demonstrate that the obvious tradeoff for this transformation i.e. of getting *false positives* is within manageable limits. This is indeed our principal contribution.

## 2. PROBLEM FORMALIZATION AND DATASETS

We state the problem of geo-coding of addresses afresh from the perspective of a geo-spatially annotated corpus of surveyed addresses, each with a latitude-longitude pair associated with it.

**Geo-Id Resolution:** Given a corpus of address texts $\mathcal{A}$, such that $\forall A \in \mathcal{A}$, the geo-spatial expanse of $A$ is known ($lat_A$, $lon_A$), and a query address $Q$, find the address/es that *best* determine the geo-spatial expanse of $Q$. As we shall see in subsequent sections (see §3), there is an elegant way to use geo-spatial hashing of annotated address texts in the corpus to achieve precisely this functionality. But first, we describe and explore the datasets used for the study.

**Datasets:** Our primary data source is collected through an intensive *war-driving* effort conducted in six major cities in India. We call this the *active survey* phase in our system design. Trained personnel (surveyors)[5] visit popular regions (localities/areas) in each of the six cities to collect addresses for POIs manually. The address text for each points-of-interest (POI) is entered through an application running on a handheld device by the surveyor. The application provides a template form distinguishing different components in an address (such as POI name, building number, building name, proximal landmarks, street name, area name, city etc.), thereby automatically labeling strings into different geo-spatial entities

Additionally, we collected a set of addresses ($\approx 2,500$) spread across six major cities in India by crawling several local search websites (such as `www.justdial.com`), and used OpenStreetMaps (OSM) (`www.openstreetmaps.org`), an open cartographic data project, to infer latitude-longitude coordinates (at least for a few popular POIs in each of the six cities) both on OSM and Google maps portal. For popular POIs in well known localities in these cities (in particular restaurants, stores and cinema theaters), the agreement between OSM and Google maps was quite high. We treat the set of such POI addresses and their locations as test queries with ground-truth (by consensus) in our study.

## 3. FROM ADDRESSES TO GEO-SPATIAL DOCUMENTS

In this section we describe a simple and elegant methodology of organizing surveyed POI addresses, that helps us overcome the challenges discussed in preceding sections. Address texts, when seen in isolation, can show significant lexical dissimilarity even for POIs which are in geo-physical vicinity of each other. For example, two POIs located in close proximity may have different street names in them (one official and another colloquial). Others, and this is also not uncommon in our secondary dataset, might only have the name of a locally popular landmark in them (for e.g. near Alankar's Cinema Theater, Lajpat Nagar III, New Delhi.). An address, in the Indian context, is what it is perceived to be instead of what it ought to be.

In order to accommodate such variations of percept we devise a simple methodology of representing geo-physical regions of interest in a city. We overlay a logical square grid over the city thereby dividing it into unit cells of a desired dimension (say $\approx 250m \times 250m$). We treat each of these grid-cells as geo-spatial documents, the contents of which is

[5]With growing popularity GPS-enabled devices in emerging markets, in near future this can be crowd-sourced.

a collection of labeled strings contributed by each of the surveyed POIs that map to it. The city, is thus transformed to a collection of discrete documents that contain all possible strings (names of buildings, streets and areas) that appear in the addresses belonging to that city. Clearly, this assimilative process reduces the number the documents against which a queried address needs to be evaluated for lexical similarity. Indeed, we see that the 12, 000 surveyed POIs for city 1, map to a mere 1,150 geo-spatial documents each of size $250m \times 250m$ and similar reductions are observed for other cities too. Moreover, and perhaps more importantly, as a geo-spatial document combines street, area and landmark names together for all the POIs within its expanse, the question of uniqueness becomes important — how different is a given geo-spatial document from others in terms of lexical content.

In order to ascertain this we extract all non-dictionary words (proper nouns) for each of the geo-spatial documents including landmark, street and city names. For a document $X$ we denote the term vector as $T^X$. Next we compute the lexical similarity of these term vectors for all pairs of geo-spatial documents using the Jaccard distance: $J(X,Y) = |T^X \cap T^Y|/|T^X \cup T^Y|$. We observe, that as the geo-physical distance increases, so does the Jaccard distance i.e. lexical similarity between geo-spatial documents decreases consistently with distance. This implies that documents have significant lexical variation despite being aggregates of multiple POI addresses, which bodes well for a search like solution that we present in the next subsection.

We can now re-pose the problem of geo-identity resolution of address texts in terms of $\mathcal{X}_C = \{X : X$ is a geo-spatial document in city C $\}$. The geo-identity of the query $Q$ is then resolved to within the grid cell corresponding to the geo-spatial document $X^*$, which attains highest score for Q. The scoring algorithm of a search engine provides a ranking of the documents in the corpus for a given query. In practice, however, we may find a subset $\hat{\mathcal{X}} \subset \mathcal{X}$ of geo-spatial documents that attain competitive scores for a given query. In such cases, depending upon the distribution of the scores, we compute the smallest contiguous sub-region formed by adjacent grid cells [3].

## 4. EXPERIMENTS AND RESULTS

We implement the query serving engine of our geo-identity resolution system on Solr/Lucene [1], an open source, enterprise scale search engine from Apache Inc. Solr/Lucene that provides highly configurable interfaces. The search engine accepts a free-text query address as its input and returns a rank-order of geo-spatial documents along with the ranking scores obtained by each geo-spatial document for the given query. We configure the scoring algorithm to reward matches not only on the basis of number of common terms between the query and the geo-spatial documents but also on the basis of contiguous terms.

**Creating Test Cases:** We use the corpus and the secondary data sources to create three different kinds of test queries, viz. *in-corpus, partial in-corpus and induced false positives* queries. We now discuss these in detail below with respect to the number of false positives and false negatives obtained in each case:

a. **In-corpus queries:** Given an address $A \in \mathcal{A}$, the corpus of addresses surveyed, let $X_A \in \mathcal{X}$ be the geo-spatial document to which $A$ maps based on its co-ordinates $(lat_A, lon_A)$. We randomly select $2,000$ such addresses in City 1 and search them against the geo-spatial documents for City 1, to obtain $X_A^*$, the geo-spatial document that attains the highest score and observe that $X_A^* = X_A$ in all the cases. Also, when more than one geo-spatial documents attain the maximum score, we find that these documents neighbors in the city-grid. Secondly, the second-highest ranked geo-spatial document, attains at most 65% the score attained by by $X_A^*$. Fig. 2(a) shows the distribution of top-10 scores for 5 such queries, for which the second highest scores are the highest over the the set of $2,000$ test queries. Clearly, search engine scores distinguish well between true matches and partial matches when the corpus contains the query address. Thus false negatives is not a concern in our system.

b. **Partial in-corpus queries:** We now select a subset of 500 addresses from the secondary dataset that contain popular landmarks in them. Based on these unique landmarks, we obtain the geo-spatial coordinates of each of these addresses from the OpenStreetMaps interface by locating them on the city map. Note that each such query address is only required to have landmark and either street/area name or both in it, therefore these addresses represent the case of partial information in the corpus. Once again, we obtain score distributions similar to those for in-corpus address queries (see Fig. 2(b)), with relatively smaller number of false positives ($\approx 5\%$ on average) for the highest ranked document (see Fig. 2(c)). On inspection, we observe that false positives are often obtained when landmarks are of commercial chain types (e.g. Mac Donald's etc.) present in similarly named areas in the same city (e.g. Shanti Nagar and Shanthal Nagar). Moreover, in all such cases with a false positive as the highest scored document, we do obtain a tie with the true match if it is present in the corpus. Such tied results can therefore be thought of as competing suggestions as is commonly observed in the case of web-based search engines (e.g. Google/Yahoo/Ovi). Also, the second highest scores still have a maximum of 72% of the highest and 54% on an average for the 500 queries.

c. **Induced false positives:** We now create a set of modified addresses from a sample subset of in-corpus queries as follows: for an address $A \in \mathcal{A}$, the corpus of geo-annotated addresses, we select named entities POI name, landmark, street name and area name. We then permute the proper nouns from these fields to obtain a new address $\widetilde{A}$ (e.g. $A$: Feroze Gandhi Marg, near Alankar Theater $\rightarrow \widetilde{A}$: Feroze Gandhi Theater, near Alankar Marg). We therefore deliberately create false positives in our search indexes. Now, when $A$ is used as a query, the induced false positive $\widetilde{A}$'s score provides an estimate of the expected false positive scores. Similarly, when $\widetilde{A}$ is used as a query, $A$ is the false positive. We see that the max normalized scores attained by the false positives in either direction are much lower than those obtained by the respective $X^*$ (see Fig. 2(d)).

**Document Size and Performance:** The size of the unit cell reflects the granularity of geo-identity resolution achievable, the reduction in the system memory map and the rate

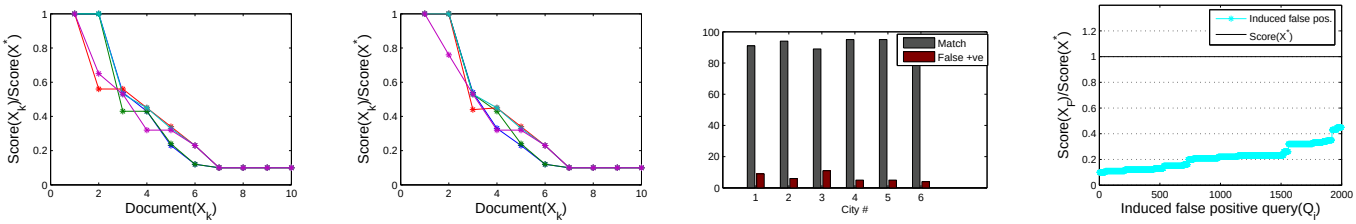



**Figure 2: (a) Top 10 documents by score for 5 in-corpus queries (b) partial in-corpus queries (c) True match (%) vs. false positives (d) True match scores vs. false positives.**

of false positives and negatives. As the unit cell size decreases, the number of geo-spatial documents increases and the density of terms per geo-spatial document decreases. However, the number of false negatives does not rise significantly even for $100 \times 100m^2$ cells, even though the number of competing documents per partial in corpus query rises. This is because our solution uses density based clustering of documents with competing scores and therefore the region assigned to a query usually falls within the unit cell of a higher dimension. On the higher end, the number of false positives does go up with increasing document sizes. In fact, with the unit cell length of $1km$, we observe a sudden 25% rise in the average number of false positives per query. This effect is largely due to higher term similarity in geo-spatial documents at this limit. We find that the discriminating terms per document thin out at such scales. We found the range $200m$ to $500m$ as the seemingly optimal for all six cities.

## 5. RELATED WORK

Most of the work on geo-coding of address texts is implemened in commercial search platforms like Google, OVI and Yahoo. Recently, other applications such as inferring context of users' search queries [4] and also in assessing the geographic cues in news queries [5]. Such studies rely heavily on identifying geo-spatial texts as differentiating factors amongst a set of short text documents. Similarity measures for short text documents, particularly queries, have been studied in [7, 8] and applied to indexing in [2]. Typically, such solutions rely on one or more variants of the probabilistic latent symantic indexing [6] method. Sahami et al first used the search engine score for document classification [9]. Their solution is clearly the inspiration for our work where we select the search engine score as an indicator of the level of similarity between a query address and the geo-spatial documents in the corpus. Last but not the least, density based clustering schemes help accommodate documents with competing scores. We use one of the more famous solutions for this task [3].

## 6. CONCLUSION, DISCUSSION AND FUTURE WORK

We studied the problem of resolving geo-identity of address texts in emerging markets. Despite the lack of proper addressing schemes, cartographic information and standardized nomenclature of geo-spatial entities, such as streets and areas, we show that address texts still have some latent structural information. When viewed as short documents address texts convey rich associative geo-spatial clues, such as proximity information and landmark like references to locally popular named entities that help differentiate a geo-spatial locality from others. Based on these observations, we posed the problem of resolving geo-identities of address texts as a problem in the document search space where geo-physical locations become geo-spatial documents containing text information. Deployed on a search engine backend, our solution achieves over 94% accuracy in resolution and an order of magnitude reduction in system state with nearly zero false-negatives. Given the potential for location based services in emerging markets, there is tremendous scope for future research. Our work is only a first step in this direction.

## 7. ACKNOWLEDGEMENT

We are thankful to Mohammad Rahimi, Luis Sarmenta and Alison Lee who participated in preliminary discussions on this work.

## 8. REFERENCES

[1] *http://lucene.apache.org/solr/.*

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, *Indexing by latent semantic analysis*, vol. 41, 1990, pp. 391–407.

[3] Martin Ester, Hans peter Kriegel, Jorg S, and Xiaowei Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, AAAI Press, 1996, pp. 226–231.

[4] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, *Analysis of geographic queries in a search engine log*, Proc. of LocWeb (2008).

[5] Ahmed Hassan, Rosie Jones, and Fernando Diaz, *A case study of using geographic cues to predict query news intent*, Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (New York, NY, USA), GIS '09, ACM, 2009, pp. 33–41.

[6] T. Hoffman, *Probabilistic latent semantic analysis*, Proc. of UAI '99, 1999.

[7] Rosie Jones, Benjamin Rey, and Omid Madani, *Generating query substitutions*, In WWW, 2006, pp. 387–396.

[8] Donald Metzler, Susan Dumais, and Christopher Meek, *Similarity measures for short segments of text*, In Proc. of ECIR-07, 2007.

[9] M. Sahami and T. D. Heilman, *A web-based kernel function for measuring the similarity of short text snippets*, Proc. of WWW, 2006.