On the Accuracy of Online Geocoders

Dirk Ahlers
OFFIS Institute for Information Technology
Oldenburg, Germany
ahlers@offis.de

Susanne Boll
University of Oldenburg
Germany
susanne.boll@uni-oldenburg.de

ABSTRACT

Geocoding is the conversion of a textual description of a location to geographic coordinates. With online geocoders being freely available to researchers and practitioners alike, their influence on data quality needs to be estimated. To this end, we first describe the basics of address-based geocoding and accuracy issues. We then look at two of the most widely-used geocoders and provide analysis of their automatic geocoder results, discuss their quality, and present a correction methodology to increase the accuracy of geocoding, using two independent sources, heuristics on accuracy metadata and conflation techniques.

1. INTRODUCTION

Most geospatial Web applications demand the transformation of a textual description of a place into a geographic coordinate by a geocoding process, to support the mapping of data or user queries.

The major mapping providers also offer online geocoders to prepare textual data for mapping. They are provided free of charge and are readily available and are therefore a preferred source for researchers and practitioners alike. These geocoder services are used in a multitude of applications such as rapid prototyping, research tools, and the vast amount of geographic mashup services on the Web.

We look into the two most widely-used online geocoders, namely those of Google and Yahoo!, which are both freely available and have coverage for Germany. The aim is to identify and quantify inaccuracies as well as differences and similarities within the heterogeneous results to derive a correction method to reduce the overall error in address-level geocoding.

Our application scenario of geographic Web Information retrieval aims to identify and extract location references in unstructured Web pages and to provide spatial search capabilities on these documents [1], aiming at the high granularity of individual addresses [2]. To actually enable spatial search and analysis capabilities such as vicinity searches, the textual address has to be converted to a geographical coordinate by a geocoder. The retrieved data then is suitable for geospatial search applications and supports assistance to mobile users.

2. RELATED WORK

In the field of geographic information retrieval, the issue of uncertainty is usually discussed with a focus on term disambiguation for placenames in unstructured documents [7], [3]. In the context of IR, the challenge lies in correct identification and assignment of geospatial properties.

A complementary topic is the accuracy of the geographical coordinates that are assigned to the extracted information. [9] discusses the topic of geographic uncertainty, ranging from theoretical aspects over modelling, handling, and mapping up to data acquisition and positioning. [6] gives an overview of the state of the art in geocoding. An analysis of geocoder error levels and an initial correction method can be found in [8]. [5] examines the positional errors of geocoders by the distances between entities. [4] describe a correction methodology based on direct access to interpolating line-based reference data and additional external sources for parcel sizes and distributions.

3. GEOCODER ACCURACY

An address in itself is a hierarchical textual description of a certain place and can be geocoded to a geographical point within a small radius. We have discussed some of the issues of address recognition, identification, extraction and verification in our previous work [2]. The location granularity of full addresses is rather high at a building level and is then well useable by pedestrians or other mobile users. Especially at such high granularity, small errors can easily accumulate and become significant. For a consumer of geocoding results, data quality is not always easily to be assessed and needs to be considered.

3.1 Requirements

The application scenario of pedestrian assistance has accuracy requirements at the granularity of individual addresses. This leads to the interesting question of what exactly makes a 'correct' and 'accurate' coordinate for a geocoded address. Apart from the semantics of the point approximation of an extended geographic entity, we discuss two main requirements. One is the absolute positional accuracy as the congruence to the physical world. The other is the relational accuracy which should maintain spatial relations such as distance and direction between entities and buildings, keeping them clearly distinguishable.

3.2 Analysis

The area of study for the address-level geocoding is the city of Oldenburg in Northern Germany. Preprocessing was done by our validating address parser [2]. We had addresses geocoded by both the Google and Yahoo geocoder APIs. For initial analyses, we have mapped whole streets by iterating through all possible house numbers.

We compared the results by various statistical measures and by visualization and mapping for analysis of the disagreement of both geocoders. Furthermore, we used manual comparison with official cadastral data to estimate the absolute positional error. We additionally examine the accuracy annotations of the geocoders as a granularity and reliability measure. However, this does not capture uncertainty due to the geocoder's reference database, number of street segments, interpolation etc. Still, even this rough measure provides interesting insights into coverage and completeness of the respective reference databases. The two geocoders prove to have differing coverage, with no single source always preferable to the other. Similarly, the analysis of distances, positional errors and error distribution shows non-uniform character between the two sources.

We have found a variety of mismatches which in part can be traced or attributed to a range of issues. We find geocoding inaccuracies due to line simplification or interpolation errors as well as suspected mismatches between the reference data and reality, such as out-of-date street directories, overlapping streets, missing street segments, or non-existing house numbers, to name just a few. The full paper will explore and discuss them in more detail.

For some cases, correct and wrong results can have the same properties which are hardly detectable without grounding to the real world. However, in many other cases, outliers or inaccuracies can be detected within street data and can then be used to reduce the reliability measure of that data. Therefore, the challenge is to distinguish and rank results based on their derived accuracy and reliability to support a usage decision.

While some overall measures and assumptions can be made, differences within individual streets do not allow for a preference of one source over another. One important finding, therefore, is that the geocoders cannot be evaluated in whole, but should be evaluated on a per-street basis.

4. CORRECTION METHODOLOGY

The analysis in the previous section was done mainly at the street level. However, for our application, we only want to map the address data we find and are often not interested in whole streets.

Correction of inaccuracies in the data sources proves very demanding, as no obvious feature could be identified which would give an estimate of the correctness or completeness of the geocoder for a specific street without resorting to a ground truth. For the limited area we examined above, we can conclude that while one source has a more thorough coverage, the other in part has better accuracy. By combining both online geocoders, strengths of the individual sources can be exploited while their limitations can be alleviated. We can then realise a method to reduce the error and enhance the accuracy.

A first – already efficient – error reduction strategy is to simply select the source with the better accuracy indicator. In cases with matching accuracy indicators, a comparison of their respective geographic positions is made. If the results lie within a small threshold from each other, an averaged position can be sufficient. For larger differences, it is not directly clear which geocoder provides the better results. However, due to some properties of the underlying datasets, we can use a conflation approach that results in better accuracy and coverage than each of the individual sources alone.

The approach in this case is to use a small part of the immediate environment of an address by additionally examining the directly surrounding house numbers and subjecting them to spatial analysis. Surrounding of an address or simply the adjacent buildings cannot work by simple number distances. Due to the way numbers can be assigned, this needs to take numbering schemes and spatial distances into account. Using the results of the evaluation, certain types of errors can be detected. Exploiting this information allows us to arrive at a more reliable answer.

The conflation criteria then are the stated granularity of the geocoders, the distance of points, outlier detection, variations and fluctuations in neighbouring house numbers etc. The complete algorithm will be presented in the full paper. Using our algorithm for the test area, we find a strong increase in accuracy and are able to reliably assign Still, some issues persist which cannot be rectified by our approach and currently remain undecidable. We are continuously working on these in our ongoing work.

5. CONCLUSION

We showed that free geocoding services already support a high level of granularity but that the accuracy at highest granularity levels still introduces some errors. We demonstrated that by combination of multiple sources, we can deliver geocoded locations from full addresses at a better accuracy than from individual sources alone.

The full paper will go into further detail on related work and techniques and offer more detailed analyses and classification of different error types encountered during the evaluation of the geocoders. Based on this, we will give an in-depth discussion of our correction methodology and provide an evaluation of our approach.

6. REFERENCES

[1] D. Ahlers and S. Boll. Location-based Web search. In A. Scharl and K. Tochtermann, editors, *The Geospatial Web*. Springer, 2007.

[2] D. Ahlers and S. Boll. Retrieving Address-based Locations from the Web. In C. Jones and R. Purves, editors, *GIR'08*. ACM, 2008.

[3] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web Content. In *SIGIR'04*. ACM, 2004.

[4] R. Bakshi, C. A. Knoblock, and S. Thakkar. Exploiting Online Sources to Accurately Geocode Addresses. In *GIS'04*. ACM, 2004.

[5] M. R. Cayo and T. O. Talbot. Positional error in automated geocoding of residential addresses. *Int. Journal of Health Geographics*, 2(1):10, 2003.

[6] D. W. Goldberg, J. P. Wilson, and C. A. Knoblock. From Text to Geographic Coordinates: The Current State of Geocoding. *Journal of the Urban and Regional Information Systems Association*, 19(1):33–46, 2007.

[7] K. S. McCurley. Geospatial mapping and navigation of the web. In *WWW'01*. ACM, 2001.

[8] J. H. Ratcliffe. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science*, 15(5), 2001.

[9] J. Zhang and M. Goodchild. *Uncertainty in Geographical Information*. New York, 2002.