Assessing the Accuracy of Crowdsourced POI Names

By: Abdulelah A. Abuabat, Mohammed A. Aldosari, and Hassan A. Karimi

Outline

1. Motivation
2. Dataset & Data Sources
3. Methodologies
    a. Similarity Algorithm
    b. Thresholds
4. Results
5. Limitations
6. Takeaways

Motivation

- The debut of Web 2.0 has led to the emergence of many new applications, such as crowdsourcing applications. Such applications has become popular since users of different interests can voluntarily add, edit, and review data. They collaborate together to accumulate such a data and review it over time.

- Considering that contributors may not follow specific standards to contribute new data, or they may not be aware of them if existed, it is imperative to pay attention to quality of VGI data. Assessing the quality of such a data is beneficial to other users, such as those who use the service for navigation.

- The contribution of this paper is to check the reliability of point of interest (POI) names in *OpenStreetMap* (OSM) to see to what extent the OSM POI names are similar to the corresponding names in the reference dataset.

Dataset & Data Sources

- Our data was from *OpenStreetMap* (OSM), which is a popular map-based Volunteered Geographic Information (VGI) where anyone can contribute geospatial data, and it is intended to be widely available and used by others without any restrictions.

- The reference dataset is provided by *Placesdatabases*, a commercial vendor of spatial data.

- Our focus was only the Pennsylvania state, USA. The number of POIs, which have names in the OSM dataset is 89207. The OSM dataset contains the editing histories of these POIs.

Methods

In our project, we have used the following three methods to measure the similarity between the POI names in the OSM dataset with their corresponding POI names in the reference dataset.

1. The overall similarity between the POI names in the last version of the OSM dataset and their corresponding POI names in the reference dataset.
2. The overall similarity between the POI names in the last version of OSM dataset and its earlier version and consider only those OSM POI names that perfectly match (100%) their corresponding POI names in the reference dataset.
3. The average percentage of edits needed for an OSM POI name to match perfectly (100%) its corresponding name in the reference dataset.

Methods

In our project, we have used the following three methods to measure the similarity between the POI names in the *OpenStreetMap* dataset with their corresponding POI names in the reference dataset.

1. The overall similarity between the POI names in the last version of the OSM dataset and their corresponding POI names in the reference dataset.

We assumed that the latest version contains the most accurate POI names. Contributors may update POI names as they recognize errors, and POI names may evolve over time to be accurate and reflect the real names. However, POI names may not be correct if contributors have different views as to which is the correct name of a POI.

Methods

In our project, we have used the following three methods to measure the similarity between the POI names in the *OpenStreetMap* dataset with their corresponding POI names in the reference dataset.

2. The overall similarity between the POI names in the last version of OSM dataset and its earlier version and consider only those OSM POI names that perfectly match (100%) their corresponding POI names in the reference dataset.

The objective is to analyze whether or not the OSM POI names have been edited and revised frequently.

Methods

In our project, we have used the following three methods to measure the similarity between the POI names in the *OpenStreetMap* dataset with their corresponding POI names in the reference dataset.

3. The average percentage of edits needed for an OSM POI name to match perfectly (100%) its corresponding name in the reference dataset.

The objective is to realize how many edits on average are needed for POI names in OSM to be accurate and perfectly match their corresponding POI names in the reference dataset.

Similarity Algorithm

- Two strings can be similar semantically or lexically. String similarity measure can be divided into two main categories: term-based and character-based.
- Our work was focused on measuring the similarity between pairs of POI names. Therefore, we have compared string pairs, lexically, by taking the character-based approach.
- We have used the *Levenshtein Distance Strings Metric algorithm* to conduct this work.

Levenshtein Distance Strings Metric Algorithm

As documented in many discussions, including Wikipedia; for two strings x and y with length of i and j, respectively, the Levenshtein algorithm defined as:

$$LevDist\,(i,j) = \begin{cases} if\ \min(i,j) = 0: \\ \quad \max(i,j) \\ else: \\ \quad min \begin{cases} LevDist\,(i-1,j)+1 \\ LevDist\,(i,j-1)+1 \\ LevDist\,(i-1,j-1)+k \end{cases} \end{cases}$$

Table : An Example Showing the Results of the Levenshtein Distance Strings Metric Algorithm

1st String	2nd String	Similarity
University of Pittsburgh	University of Pittsburgh	100%
	University Pittsburgh	96%
	University of Pitt	86%
	Pittsburgh	59%

Thresholds

Our approach of matching the POIs in the OSM dataset and the reference dataset may produce inaccurate results because of two main issues.

1. The nearest POI in the reference dataset may not be the correct corresponding POI. This issue might occur due to location accuracy.

2. Multiple POI locations may overlap, in other words, POIs inside a POI. For example, two POIs might overlap within the same boundary like McDonald's as a restaurant and Walmart as a supermarket.

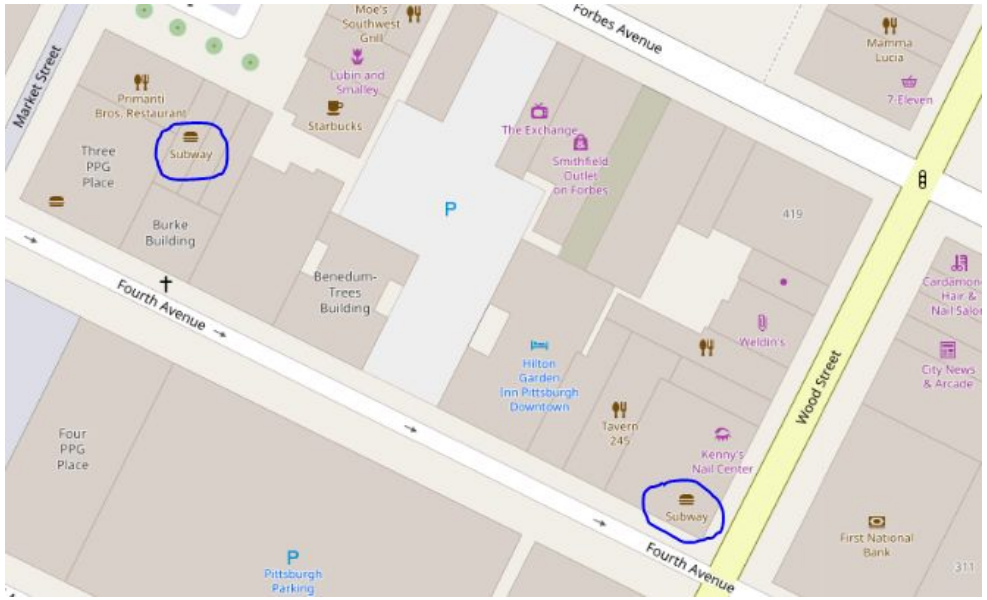
Threshold #1

Our approach of matching the POIs in the OSM dataset and the reference dataset may produce inaccurate results because of two main issues.

1. The nearest POI in the reference dataset may not be the correct corresponding POI. This issue might occur due to location accuracy, see the figure below.

Cut-off Threshold:

We determined a distance threshold to reduce matching errors through an analysis where we checked the locations of the two nearest Starbucks branches. We found that they are approximately 400 meters away from each other. We used 400 meters as a threshold for the maximum distance between these two POIs. Thus, an OSM POI will not be incorrectly matched with a similar but not corresponding POI in the referenced dataset.



Example of same POIs located close to each other within a distance below the threshold

Threshold #2

Our approach of matching the POIs in the OSM dataset and the reference dataset may produce inaccurate results because of two main issues.

2. Multiple POI locations may overlap, in other words, POIs inside a POI. For example, two POIs might overlap within the same boundary like McDonald's as a restaurant and Walmart as a supermarket, see the figure below.

Cut-off Threshold:

To address this issue, we examine several names for the same POI, especially names with abbreviations, such as "Saint → St.", "Fifth → 5th", "Avenue → Ave.", and state abbreviation "New York → NY", to find a minimum similarity percentage. We found that 40% is reasonable as the minimum similarity percentage. Table II shows an example of this test.

Example of POIs located inside a POI. Walmart Pharmacy and McDonald's inside Walmart supermarket

Results

- We would like to mention that about 17136 POIs, which is 19.2% of the dataset, have 100% similarity with the reference dataset.

- Method 1, we found 80.62% overall similarity between the POI names in the OSM dataset and the POI names in the POI names in the reference dataset. Thus, we may say that POI names in OSM have potential to be an accurate and reliable source.

- Method 2, we found 98.74% match between the POI names in the latest version of OSM dataset and the earlier version of OSM dataset. This means that if the POI name in OSM is entered accurately.

- Method 3, we found that after 3.9% of the number of edits, OSM POI names will match the corresponding names in the reference dataset correctly. For instance, if a POI name is edited 100 times, it is likely that the accurate name remains the same after the fourth edit.

Limitations

- Contributors may tend to follow different approaches to identify and specify the location (latitude, longitude) of a POI on a map where each approach may result in a different location. For instance, a user may specify the center point as the designated location, while another may specify one of the edge as the designated location.
  - To address such issue, the similarity threshold discussed earlier is used to preclude those comparisons where names are significantly different.

- Contributors may tend to use different approaches while naming POI. They may use different words or symbols interchangeably while they mean the same thing.
  - To address such an issue, users should be reminded with the common naming conventions used during the process of naming POIs. Also, users should follow the naming used by local governments while naming POI to avoid naming conflicts.

Takeaways

- To assess the OSM POI naming, two approaches can be followed: assessing the POI naming lexically or semantically. The scope of this project is to assess the POI naming lexically and analyze how similar OSM POI naming is similar to the names in the reference dataset.

- Contributors should be provided with guidelines and conventions, such as where to pick the location of the POI and how to name it according to the governmental naming conventions.

- In terms of the 40% as the minimum similarity percentage, it is true that this threshold may lead to false positive results. However, having a high threshold may also lead to another issue which is false negative cases that should be included in the analysis to assess their similarity. Thus, we address in our work that we examine the different perspective of writing POI names, and we estimate that 40% percent can be a considerable choice.

References

[1] W. H. Gomaa and A. A. Fahmy. "A survey of text similarity approaches," International Journal of Computer Applications, no. 13, pp.68, 2013.

[2] C. Barron, P. Neis, and A. Zipf, "A comprehensive framework for intrinsic OpenStreetMap quality analysis," Transactions in GIS, vol. 18, no. 6, pp. 877-895, 2014.

[3] G. Touya, V. Antoniou, A. Olteanu-Raimond, and M. Van Damme. "Assessing Crowdsourced POI Quality: Combining Methods Based on Reference Data, History, and Spatial Relations," ISPRS International Journal of Geo-Information 6, no. 3. pp. 80, 2017.

[4] D. Jonietz, and A. Zipf. "Defining fitness-for-use for crowdsourced points of interest (POI)," ISPRS International Journal of Geo-Information 5, no. 9, pp. 149, 2016.

Thank you!