Assessing the Accuracy of Crowdsourced POI Names

Abdulelah A. Abuabat, Mohammed A. Aldosari, Hassan A. Karimi
School of Computing and Information
University of Pittsburgh, Pittsburgh, USA
e-mail:{aaa185, maa303, hkarimi} @ pitt.edu

***Abstract*—The purpose of this paper is to assess the accuracy of the Volunteered Geographic Information (VGI), specifically naming point of interests (POIs), in OpenStreetMap (OSM). For this, we compared, from a lexical perspective, the similarity of POI names in the OSM dataset with their corresponding names in a reference dataset. The overall similarity is 80.62% suggesting that POI names in OSM have potential to be an accurate and reliable source.**

Keywords- Crowdsourcing; Volunteered Geographic Information (VGI); text similarity analysis; point of interest (POI); OpenStreetMap (OSM).

I. INTRODUCTION

The debut of Web 2.0 has led to the emergence of many new applications, models, tools, and projects, among other things. One of these projects is crowdsourcing. Although the crowdsourced data occasionally is redundant and noisy [5], crowdsourced data often shows decent quality [4], flexibility [4], reliability [6], and economy [6]. One of the well-known crowdsourcing projects is OpenStreetMap (OSM), a popular map-based Volunteered Geographic Information (VGI) [16] project designed to crowdsource geospatial data to build a freely accessible world map. Contributors can easily create a new POI, modify an existing POI, or extract data from a region of their interest. They collaboratively contribute to OSM through the OSM official website, desktop or mobile based applications.

Generally, VGI-based projects are known to be effective since contributors are not restricted to add or edit data/information. Contributors can report a change that might occur faster than commercial geographical information providers. For instance, contributors could report a road closure that occurs due to a natural disaster, and the change would be reflected immediately [9]. Considering that contributors do not follow specific standards to contribute new data, it is imperative to pay attention to quality of VGI data [7] [8]. Of possible VGI data errors, those related to naming POIs are focused on this paper. The contribution of this paper is to check the reliability of POI names in OSM, as a representative collaborative mapping project. The remainder of this article is structured as follows. Section II discusses related work. Section III discusses the methods used to measure POI naming accuracy. The analysis performed and its results are discussed in Section IV. Section V concludes the work.

II. RELATED WORK

OSM is a collaborative mapping project where anyone can contribute geospatial data and it is intended to be widely available and used by others without any restrictions [8] [13]. As OSM has become popular and widely used, attentions have been paid to its data quality [3] [8] [14]. Assessing the quality of data/information collected by the crowd in OSM is of great importance to the products and services that are based on the OSM data and maps. The two approaches in assessing VGI data quality are quality measure and quality indicator. In the quality measure approach, VGI data are compared with a reference dataset. In the quality indicator approach, VGI data are evaluated through intrinsic methods. In these methods, VGI data quality is evaluated by means other than a reference dataset [2]. For instance, contributors' behaviors are analyzed to estimate the overall data quality. VGI data quality, in terms of basic data quality measures, such as completeness, attribute accuracy, and semantic accuracy, have been extensively studied. It is argued in [15] that in OSM and similar projects, attribute names may be highly inaccurate due to lack of standards and clear naming conventions. In [12], answering the question regarding the number of contributions needed to map an area accurately was focused. It was found out that five contributions would result in an acceptable level of positional accuracy. In a recent work in [3], the authors assessed a subset of the OSM dataset with a reference dataset by analyzing the POIs that have changed frequently in terms of their names and positions. In their work, they focused on only one POI type, subway station, since it has a frequent number of changes regarding its name and position. They proposed an approach for identifying whether two POIs are homologous based on three measures: position, name, and amenity type. They manually evaluated these points and found that the majority of them are similar; 328 out of 329 POIs correctly matched their corresponding OSM POIs. Different from this work, we consider all amenity types in an OSM dataset and evaluate their similarity in terms of POI names. Furthermore, our work also evaluates names as they are edited and reviewed by other contributors.

III. METHODOLOGY

In this section, we will give a general overview of OSM, the methods we followed in our analysis to assess the quality of OSM POI names, and the measure we used to calculate the similarity between pairs of corresponding names in the OSM dataset and the reference dataset.

A. Overview

In OSM, the data model is classified into three types: nodes, ways, or relations. Nodes represent POIs which are objects or entities, e.g., a school or a restaurant. Ways represent groups of interrelated nodes, such as a group of

 ISBN: 978-1-61208-636-1

POIs in a building. Relations represent the relationships between nodes, ways or other relations. Contributors provide information, to the best of their knowledge, about POIs. They create new POIs and add some information about each. Other contributors may update a POI content by correcting errors and/or adding further information. While creating or editing a POI, contributors may choose among agreed-upon information from which they can make a selection, such as amenity types. They may also include some other information about a POI, for example, name and address. Our work in this paper is focused on assessing the overall quality of POI names in OSM. Currently, there is a void in the literature about text verification methods to check the correctness of the written words. For instance, if a contributor writes "Universty of Pittsbrg" instead of "University of Pittsburgh", OSM allows the contributor to save the incorrect name. Moreover, contributors may follow different naming conventions while creating or editing a POI. For instance, a contributor may write a street name as "Fifth Ave.", while another contributor may edit it to "Fifth Avenue". As it is stated by [10], most OSM contributors are amateur and have diverse backgrounds, education, and cartographic knowledge. In addition to the OSM dataset that was extracted from the OSM world history file, we have obtained a reference dataset as a ground truth data. The reference dataset is provided by Placesdatabases.com, a commercial vendor of spatial data. As it is mentioned in [11], the ground truth data is also susceptible to errors, and the assumption that the ground truth data is fully reliable is not valid. For instance, ground truth data may be outdated or may not be updated regularly as new data is added to the map compared to the VGI data which may be updated as soon as new data comes in. Placesdatabases.com claims that their data is completely refreshed every three months.

B. Methods

In this work, we use 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.

Method 1. In this method, the overall similarity between the POI names in the last version of the OSM dataset and their corresponding POI names in the reference dataset is measured. Since POIs in OSM are usually updated frequently through a set of revisions, we assume 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.

Method 2. In this method, we measure 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.

Method 3. In this method, we measure 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.

C. Similarity Measure

String similarity analysis is considered a significant tool in different applications, such as text mining, text classification, document analysis and clustering, and information retrieval. Two strings can be similar semantically or lexically. String similarity measure can be divided into two main categories: term-based and character-based. Since our work is focused on similarity measure between pairs of POI names, we compare string pairs lexically by taking the character-based approach. We use the Levenshtein Distance Strings Metric algorithm, which is character-based and calculates the minimum number of single character edits, i.e., deletion, substitution, and insertion, for the comparison. Table I shows an example of this algorithm that is used to compare two strings.

To compare two POI names, we consider the location, represented as latitude and longitude, of each POI in the OSM dataset. Next, we search the selected OSM POI with the nearest two POIs in the reference dataset, using the Euclidean distance. After finding the nearest two POIs, we check the POI names, by using the Levenshtein Distance Strings Metric algorithm, to see which one has the highest names similarity.

Our approach of matching the POIs in the OSM dataset and the reference dataset may produce inaccurate results because of two main issues. First, the nearest POI in the reference dataset may not be the correct corresponding POI. This issue might occur due to location accuracy [1] [2]. Second, 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, as in Figure 1. To address these issues, we set specific conditions to improve the matching quality.

TABLE I. 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%
	School of Computing and Information	20%
	NA	0%

 ISBN: 978-1-61208-636-1

These conditions were derived by conducting an analysis in the city of Pittsburgh, as described below. 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. For instance, one of the Subway branches, in Figure 2, may be incorrectly matched with the other branches in the reference dataset although their POI names may be similar. The reason why we did not consider a larger or smaller distance is because some places are very large, like a university campus or a shopping center, and some are very small like Starbucks. Therefore, by using a threshold like the one here, we can ensure that a large POI, which may contain other small POIs within its boundary, will be included in the process, see Figure 1. Additionally, in the matching process, we include both POIs so that the most similar POIs, in terms of names, are considered [3]. To address the second issue, which is POIs overlapping, 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.

IV. DISCUSSION AND RESULTS

The number of POIs, which have names in the OSM dataset in Pennsylvania is 89207. Of these, 17136 POIs (19.2%) have 100% similarity with the reference dataset. In the next two sub-sections, we will discuss the results of each method and the obstacles we faced.

A. Results

By applying 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. By applying 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 the first time, there is a high probability that it will remain to be accurate and unchanged in subsequent versions. By applying 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. As we can see, the percentage of edits needed to ensure accurate POI names is relatively low. This means that if a POI name is accurate the first time it is entered into the OSM dataset, chances are low that it will be edited in subsequent versions. In other words, most often the contributors tend to enter the correct names of POIs in the first place.

TABLE II. SIMILARITY RESULTS FOR NAMES AND THEIR POTENTIAL EQUIVALENT NAMES.

1st String	2nd String	Similarity
Saint Louis	St. Louis	80%
Fifth Avenue Station	5th Ave.	43%
New York	NY	40%
Starbucks	Subway	27%
Walmart	McDonald's	24%

B. Limitations

As the goal of this work is to assess the quality of OSM POI names by comparing them against the names in the reference dataset, there are several considerations, related to quality standards which are mentioned in [12], that are worth mentioning. For instance, contributors may 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. This issue might also be found in the reference dataset. For example, matching McDonald's in Figure 1 would find a closest POI in the reference dataset

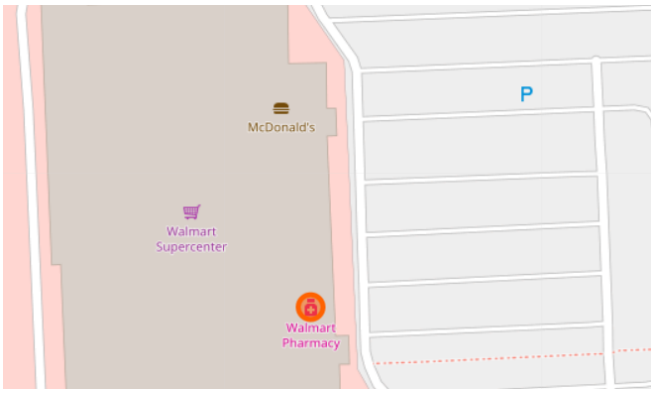

Figure 1. Example of POIs located inside a POI. Walmart Pharmacy and McDonald's inside Walmart supermarket [17].

Figure 2. Example of same POIs located close to each other within a distance below the threshold [17].

where the distance between the locations in OSM and in the reference dataset is very small, thus considered overlapped; one scenario is that the McDonalds's in OSM is matched with the Walmart Supercenter in the reference dataset. In situations like this, the similarity threshold, discussed above, is used to preclude those comparisons where names are significantly different.

In addition to the issue of matching the OSM POI names with their corresponding names in the reference dataset, there is an issue of semantic similarity. Contributors may use different words or symbols interchangeably while they mean the same thing. For instance, a contributor may write a POI's name as "School of Computing & Information" instead of "School of Computing and Information". In such situations, our proposed approach of similarity measure may not produce 100% match, despite the fact that both names are semantically the same. One way to address this issue is by reminding the contributor of the common naming conventions used during the process of naming POIs. Also, in OSM we observed that contributors interchangeably write names in short forms, e.g., "5th Ave." instead of "Fifth Avenue". In such situations, while both names are semantically the same, the similarity percentage will be low. However, adhering to a naming convention is one way to address the semantic similarity issue, but there still remains the problem of different naming standards in different countries.

V. CONCLUSION

In this paper, we focused on assessing the accuracy of VGI in naming POIs. We implemented three methods to measure accuracy of POI names: the overall similarity between the OSM dataset and the reference dataset, the similarity between the last version and an earlier version of the OSM dataset, and the average number of edits needed to have OSM POI names to be 100% similar to their equivalent in the reference dataset. We focused on the lexical perspective of the names, rather than the semantic view of the names, and found that most POI names in OSM are accurate. This work introduces new research questions: How can the accuracy of POI names be improved? Can there be a unified style for naming POIs? Can there be an algorithm that helps contributors by suggesting names?

ACKNOWLEDGMENTS

While each author contributed equally, we would like to thank the School of Computing and Information at the University of Pittsburgh for the encouragements and the support.

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] M. Van Exel, E. Dias, and S. Fruijtier. "The impact of crowdsourcing on spatial data quality indicators," In Proceedings of the 6th GIScience international conference on geographic information science, pp. 213, 2010.

[5] G. Barbier, R. Zafarani, H. Gao, G. Fung, and H. Liu. "Maximizing benefits from crowdsourced data," Computational and Mathematical Organization Theory 18, no. 3, pp. 257-279, 2012.

[6] O. Alonso, and S. Mizzaro. "Using crowdsourcing for TREC relevance assessment," Information Processing & Management 48, no. 6. pp. 1053-1066, 2012.

[7] A. J. Flanagin, and M. J. Metzger. "The credibility of volunteered geographic information," GeoJournal 72, no. 3-4, pp. 137-148, 2008.

[8] L. A. Ali, F. Schmid, R. Al-Salman, and T. Kauppinen. "Ambiguity and plausibility: managing classification quality in volunteered geographic information," In Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 143-152. ACM, 2014.

[9] M. Zook, M. Graham, T. Shelton, and S. Gorman. "Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake," World Medical & Health Policy 2, no. 2, pp. 7-33, 2010.

[10] H. Senaratne, A. Mobasheri, A. L. Ali, C. Capineri, and M. Haklay. "A review of volunteered geographic information quality assessment methods," International Journal of Geographical Information Science 31, no. 1, pp. 139-167, 2017.

[11] 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.

[12] M. Haklay, S. Basiouka, V. Antoniou, and A. Ather. "How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information," The Cartographic Journal 47, no. 4, pp. 315-322, 2010.

[13] S. S. Sehra, J. Singh, and H. S. Rai. "Assessing OpenStreetMap Data Using Intrinsic Quality Indicators: An Extension to the QGIS Processing Toolbox," Future Internet 9, no. 2, pp. 15, 2017.

[14] R. Karam and M. Melchiori. "A crowdsourcing-based framework for improving geo-spatial open data," In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, pp. 468-473 , 2013.

[15] J. Girres and G. Touya. "Quality assessment of the French OpenStreetMap dataset," Transactions in GIS 14, no. 4, pp. 435-459, 2010.

[16] M. F. Goodchild. "Citizens as sensors: the world of volunteered geography," GeoJournal 69, no. 4, pp. 211-221, 2007.

[17] OpenStreetMap. www.openstreetmap.org/. Accessed 20th November 2017.

 ISBN: 978-1-61208-636-1