Cross-Country Analysis of User Profiles for Graph-Based Location Estimation

SHIORI HIRONAKA*, MITSUO YOSHIDA*, AND KYOJI UMEMURA*, (Member, IEEE)
Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi 441-8580, Japan
Corresponding author: Shiori Hironaka (s143369@edu.tut.ac.jp)
This work was supported by JST-Mirai Program Grant Number JPMJMI20B4, Japan.

ABSTRACT Social media is used all over the world. Although the home location attribute is available anywhere, few studies have been conducted on the differences in the performance of home location estimation between countries. Because social media is used differently in different countries, the difference between countries in terms of users’ relationships needs to be investigated. We consider the performance of graph-based home location estimation in capturing users’ relationships, as it is affected by users’ social media usage. We analyze user profiles in 10 countries by graph-based estimation and find that some profile attributes, in particular, contribute to the estimation performance for all the countries examined.

INDEX TERMS Cultural differences, home location estimation, online social networks, Twitter, user profile.

I. INTRODUCTION Social media is widely used everywhere in the world. People obtain and post news online. Social media data have been used for detecting phenomena such as influenza epidemics [1], earthquakes [2], and global mobility patterns [3]. In these cases, users’ home locations are utilized to combine real-world information with information obtained online. Aggregated tagged locations in users’ posts or self-declared locations in users’ profiles are ways of determining users’ home locations [4]. However, information regarding users’ home locations is sometimes missing; therefore, researchers have developed methods for estimating home locations from other information [5].

The social media outlet Twitter is used to post short texts called “tweets.” Because users post about themselves, a user’s home location can be estimated by the location-related words included in their posts [6]. This estimation method is known as a content-based approach. Users also have a tendency to interact and develop relationships with people similar to them on social media [7]. On the basis of this characteristic, a graph-based approach using relationships based on a closeness assumption, according to which users on social media connect with other users located geographically close by, is proposed for estimating users’ home locations [8]. Location estimation methods utilize these characteristics alone or in combination, which is called a hybrid approach.

Alonso-Lorenzo et al. [9] reported that estimation performance using a hybrid approach differs by country. The differences in estimation performance are affected by language and social media usage. A content-based approach depends on the language used in tweets, and different languages are used in different countries. Therefore, it is necessary to employ a suitable estimation method according to the users’ language, such as segmentation or tokenization. Unlike a content-based approach, the same graph-based method is expected to be applied even if the users use different languages. Because behaviors on social media differ by user demographics [10], the relationships among users should differ depending on the country. Therefore, the performance of graph-based estimation needs to be investigated.

To analyze the difference between countries in terms of user relationships, we compare user profiles with the performance of location estimation. We assume that a user’s profile represents the characteristics of the user and that the user’s relationships are explained by the user’s profile because these are created by or for the user. We then use graph-based location estimation to measure the characteristics of the user’s relationships.

Our research questions are as follows:
RQ1 Does the performance of graph-based location estimation vary by country?
RQ2 Is there a profile attribute that contributes to estimation performance? Is there a difference in the profile attributes contributing to estimation performance between countries?

RQ3 What characteristics of profile attributes contribute to estimation performance after considering the correlation between attributes?

RQ4 Can the language used explain the differences in the profile attributes contributing to estimation performance between countries?

We collect more than 1,000,000,000 geotagged tweets of more than 10,000,000 users and select more than 6,000,000 users for analysis. The number of unique places is more than 90,000. We select the 10 countries with the most users. We focus on 12 profile attributes, including the indirect attribute follow ratio, which is obtained from two other attributes that can be directly retrieved from the Twitter API. We roughly compare countries using filters based on profile attributes. We find that at least one attribute behaves similarly across all 10 countries.

II. RELATED WORK

Many home location estimation methods have already been proposed [4], [5]. These methods employ tweet contents or social graphs. In addition, other profile attributes such as location text [11], description [12], and time zone [11]–[13] can be employed to estimate a user’s home location. Profile attributes also contribute to estimation performance.

Some studies have mentioned users whose relationships are not related to geographic distance, many of whom include celebrity users who have many followers or are mentioned by many users. McGee et al. [14] found that users with many followers tend to be distant from connected users. Rahimi et al. [15] showed that celebrities, who are frequently mentioned users, interfere with graph-based estimation. Ebrahim et al. [16] categorized celebrities into global and local types and showed that their approach, excluding global celebrities who are mentioned worldwide, outperforms state-of-the-art estimation methods. Tian et al. [17] also proposed an algorithm to find global celebrities.

Some users who pose estimation difficulties have other characteristics. Hironaka et al. [18] compared estimation difficulty with users’ centrality and found that users with high hub and authority scores calculated by the HITS algorithm had difficult-to-estimate locations. Users with high authority scores are considered celebrities, whereas those users with high hub scores are the ones who follow many celebrities and constitute a different type of user. McGee et al. [14] also found that users who follow many users tend to be distant from connected users.

Location estimation depends on the behaviors of users on social media. Sloan and Morgan [19] investigated the use of geotagging and user demographics, such as the languages used. They showed that the rate of posting geotagged tweets varies depending on the language used by the user. However, the country in which the user lives has not been analyzed because this information cannot be directly obtained from the profile. Hedayatifar et al. [20] analyzed the geographical structure of the Twitter communication network. They found that languages and national borders appeared as breaks in communication. Tomina et al. [21] surveyed self-disclosure in profile, usage objectives, and anonymity consciousness on Twitter, targeting users in the United States, India, and Japan, and investigated how self-disclosure is influenced by usage objectives and anonymity consciousness. They suggested that cultural differences exist on social media and that social media is used in different ways depending on the culture.

In this paper, we analyze the differences in graph-based location estimation between countries. Previous studies have focused either on a single country or the whole world. In contrast, we compare estimation performance between countries.

III. DATA

We prepare datasets for each country. A dataset contains home location data used for training and testing, social graph data constructed from relationships between users, and profile data that are used to obtain user characteristics.

Data collection proceeds as follows: First, we collect geotagged tweets using Twitter Streaming API and assign a home location to each user. Next, we collect their follower-following relationships and construct a social graph. Finally, we extract profile data from the collected geotagged tweets.

A. HOME LOCATION

Home location is a user location attribute. In this paper, we assume that users are using social media around the area they actively post geotagged tweets. We consider the area where the user posted most of their tweets as the home location of the user.

Area data are provided by some governments, but collecting these for all countries is difficult. In this paper, we used a place object$^1$ that is assigned to a geotagged tweet as an area. A place object contains a place ID, name, country, place type, and bounding box. A place type likely indicates an area’s granularity.$^2$ We assign city-level home locations to users using place objects, where a place type is “city.”

We collected 1,495,799,237 geotagged tweets from January 1 to December 31, 2019, using Twitter Streaming API. The number of users who posted geotagged tweets with a city-level place type at least once was 14,917,399, and the number of users who posted geotagged tweets at least 10 times was 7,047,133 (47.2%). We then calculated the number of tweets per area for each user. We assigned a home location as the area where the user posted the most frequently

$^1$https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/place (viewed 2021-04-08)

$^2$The official documentation does not explain the possible values of place type. Five types (country, admin, city, neighborhood, and poi) appear in the data collected in this paper. We assumed that the size of the area would decrease in this order.

$^3$https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/api-reference/post-statuses-filter (viewed 2021-04-08)
to users who posted at least 10 times. Few tweets with place objects where country_code or bounding_box is null were excluded. There were no other exclusion rules for auto-posted users or celebrities. Finally, we obtained home location data for 6,685,718 users, which included 92,461 unique places.

The geographic distribution of the users with home locations is shown in Figure 1. Each user’s coordinate is calculated as the centroid of the bounding box of the user’s home location. Figure 2 shows the ranking of the number of users by country (Top 40). In this paper, we created datasets for the top 10 countries by the number of users: in order, the United States, Brazil, the United Kingdom, Japan, the Philippines, Turkey, Indonesia, India, Mexico, and Saudi Arabia. The number of users in each dataset is $|V|$, as shown in Table 1. The United States had the most users, followed by Brazil.

### B. SOCIAL GRAPH

We collect data of the relationships between users with home locations and construct a social graph based on the relationships. First, we collect the followees and followers of the users with assigned home locations using the Twitter API. Then, we create an edge if the users assigned the home locations are mutual followers.

We collected data and constructed a social graph for each country. Table 1 shows the basic statistics of the social graphs. Here, $|V|$ is the number of nodes (users), $|I|$ is the number of users who did not have any mutual following relationships.

4https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/api-reference/get-friends-ids (viewed 2021-04-08)
(isolated nodes), $|E|$ is the number of edges, $K_{\text{out}}$ is the average degree (the number of mutual friends), $S_{\text{out}}$ is the variance of the degree, and $M_{\text{out}}$ is the median of the degree. Users in the United States tend to have fewer mutual followings than Brazil. India has more isolated users than the other countries.

### C. PROFILE ATTRIBUTES

We use the given user profiles for analysis. Each tweet has a user profile (User object) attached at the posted time. We extract the newest profiles from the geotagged tweets collected in Section III-A.

We employ the following profile attributes as user characteristics: length of the username (screen_name), length of the name (name), length of the location (location), length of the description (description), number of followers (followers), number of friends (friends), number of likes (likes), number of statuses (statuses), and number of total tweets (tweets). Text length is counted as the number of UTF-8 characters, including spaces. In addition, we employ the following calculated attributes: the number of days since the account was created (spent_days), average number of tweets per day (avg_tweets), and ratio of followers to friends (follow_ratio). spent_days is defined as the number of days between January 1, 2020, and follow_ratio is defined as (friends + 1) / (followers + 1).

These profile attributes are categorized into three groups: degree-related attributes, text-related attributes, and activity-related attributes. The degree-related attributes are friends, followers, and follow_ratio. The text-related attributes are screen_name, name, location, and description. The activity-related attributes are favorites, listed, statuses, spent_days, and avg_tweets.

We consider degree-related attributes as directly affecting location estimation performance because these attributes are related to the shape of the social graph. It has been reported that the location of users with too many followers or followees is difficult to estimate [22]. The degree is also used to detect celebrities whose location is difficult to estimate [16]. In contrast, the locations of users with low activity are also difficult to estimate.

### IV. EXPERIMENTAL METHOD

Home location estimation is a task that estimates an unknown user’s home location from a social graph and known home location labels. User profiles of all users are given. We conduct two experiments to answer the research questions. In the first experiment, we evaluate the estimation performance for each country. In the second experiment, we investigate the estimation performances with profile attributes.

The experiments proceed in the following steps. First, we filter the users in the social graph using the user profile attributes. In the first experiment, the filter is not applied. Second, we construct a social graph with the filtered users. This social graph is a sub-graph of the social graph in the dataset. Then, we estimate the home locations of the users in the test data using the location labels in the training data and the sub-graph and evaluate the performance with the test data.

### A. FILTERING SOCIAL GRAPH

The filter described here filters given users based on a single profile attribute. Then, the filtered users are used to construct a filtered social graph.

In this paper, we use the following two filters to filter users based on their profile attributes: a high-cut filter, which filters users with higher values than the threshold value $\theta$, and a low-cut filter, which filters users with lower values than the threshold value $\theta$. Each filter has a threshold value $\theta$ as a parameter. The filters are expressed as follows:

$$\text{LowCut}(U) = \{ u | u \in U, a_u > \theta \}$$

(1)

$$\text{HighCut}(U) = \{ u | u \in U, a_u < \theta \},$$

(2)

where $U$ is a set of the target users of the filtering, and $a_u$ is a profile attribute of user $u$.

The filtered social graph is a sub-graph of the social graph in the dataset and is obtained with nodes as the filtered users. Excluded users are not in the graph. Edges that connect to excluded users are also excluded from the graph.

### TABLE 1. Basic statistics of the dataset.

| Country       | $|V|$ | $|E|$ | $|\theta|$ | $K_{\text{out}}$ | $S_{\text{out}}$ | $M_{\text{out}}$ |
|---------------|------|------|-----------|-----------------|-----------------|-----------------|
| United States | 1748492 | 206671 | 80973364 | 41.42 | 164.91 | 14 |
| Brazil        | 591526 | 93794 | 35659270 | 52.03 | 133.48 | 18 |
| United Kingdom| 395492 | 50172 | 14835714 | 33.29 | 86.88 | 11 |
| Japan         | 325846 | 70333 | 9305522 | 23.51 | 72.68 | 6 |
| Philippines   | 207758 | 71518 | 5756312 | 20.61 | 63.94 | 7 |
| Turkey        | 188928 | 47822 | 5857770 | 24.74 | 179.36 | 5 |
| Indonesia     | 197038 | 57446 | 2785522 | 10.96 | 32.56 | 4 |
| India         | 122646 | 59031 | 2318458 | 10.50 | 58.62 | 1 |
| Mexico        | 165315 | 35703 | 3933724 | 19.57 | 69.72 | 5 |
| Saudi Arabia  | 128700 | 46179 | 2590364 | 14.81 | 68.35 | 3 |
B. HOME LOCATION ESTIMATION METHOD

Home location estimation is a method that estimates a user’s home location from relationships between users while assuming that connected users are living geographically close by. In this experiment, we use the method proposed in the previous paper [22]. This method selects the home location that appears most frequently among the home locations of connected users as the estimated location. The method is simple and has been reported to be accurate [8].

C. EVALUATION METRICS

We use precision, recall, F1, and coverage to measure estimation performance. Precision is the ratio of the number of users whose locations are estimated to the number of users whose home location has been correctly estimated. Recall is the ratio of the number of users in the test data to the number of users whose home location has been correctly estimated. F1 is the harmonic mean of precision and recall. Coverage is the ratio of the number of users in the test data to the number of users whose locations are to be estimated.

The metrics are expressed as follows:

\[
\text{Precision} = \frac{\left| \{u \mid u \in E, \hat{l}_u = l_u\} \right|}{|E|} \quad (3)
\]

\[
\text{Recall} = \frac{\left| \{u \mid u \in E, \hat{l}_u = l_u\} \right|}{|T|} \quad (4)
\]

\[
\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)
\]

\[
\text{Coverage} = \frac{|E|}{|T|}, \quad (6)
\]

where \(T\) is the set of users in the test data, \(E\) is the set of users whose home locations are estimated \((E \subseteq T)\), \(l_u\) is the correct home location of user \(u\), and \(\hat{l}_u\) is the estimated home location of user \(u\) \((u \in E\) has \(\hat{l}_u\)).

V. RESULTS

We first compared the estimation performance between countries. Next, we analyzed the correlations between profile attributes. Then, we analyzed the relationship between user profiles and their relationships using home location estimation.

A. ESTIMATION PERFORMANCE BY COUNTRY

In this experiment, we evaluated the performance of home location estimation for each country. Estimation performance was evaluated by leave-one-out cross-validation using the home location data and social graph data in each country’s dataset. The results of each dataset are shown in Table 2. We also show the number of unique labels \(|A|\) and the average size of the areas in the table. Size is defined as an average of the diagonal length in kilometers of the area’s bounding box. It is assumed that the difficulty of estimation performance increases with increase in the number of unique labels or the size.

In Table 2, the highest F1 of 0.60 is achieved by Brazil, and the lowest F1 of 0.13 is achieved by Indonesia. The United States and the United Kingdom have more unique labels than other countries. However, the F1s of Japan and Indonesia are lower than those of the United States and the United Kingdom. The smallest number of unique labels is 172 in Saudi Arabia, which also has the second highest F1. The sizes of Brazil and Saudi Arabia are greater than those of the other countries and their F1s are relatively high, although Mexico’s F1 is not.

The granularity of place information from Twitter differs by country, even when only city-level places are chosen. Saudi Arabia is not a small country and has about 10% of the number of users as the United States, but the number of unique labels is only 1% of that of the United States, which is less than expected. Moreover, the total area of the United Kingdom is about 40 times smaller than that of the United States, but the number of unique labels is 2/3 that of the United States. Accordingly, the number of unique labels and the average size of areas differ by country. Although the number of unique labels and the average size of areas are expected to affect the estimation performance to a certain extent, the difference in estimation performance shown in Table 2 cannot completely be explained by this. Hence, we conclude that estimation performance differs by country.

If the relationships among users had similar characteristics in these countries, the difference in estimation could have been explained by the number of areas and the area size. However, this was not the case, suggesting that further analysis of users’ relationships is required. For this purpose, the absolute value of the performance of home location estimation may not be suitable.

B. CORRELATION BETWEEN PROFILE ATTRIBUTES

We focus on profile attributes to analyze users’ relationships. Some combinations of attributes are assumed to be strongly correlated, such as friends and followers, and statuses and avg_tweets. To understand these results, correlations between attributes need to be considered before further analysis.

We calculated Spearman’s rank correlation coefficient between the profile attributes in each dataset. Owing to space limitations, we only show the results of the United States, Japan, and India in Figure 3. In the graph-related attributes, friends and followers are moderately
FIGURE 3. Spearman’s rank correlation coefficient between profile attributes in the United States, Japan, and India. Correlation tends to be high within the same group. Similar results are obtained across 10 countries.

TABLE 3. Testing result of precision improvement by filtering using social graph related attributes. follow_ratio was effective in all countries.

| Country/Cut   | friends | followers | follow_ratio |
|--------------|---------|-----------|--------------|
| United States| +       | +         | +            |
| Brazil       | +       | +         | +            |
| United Kingdom| +     | +         | +            |
| Japan        | +       | +         | +            |
| Philippines  | +       | +         | +            |
| Turkey       | +       | +         | +            |
| Indonesia    | +       | +         | +            |
| India        | +       | +         | +            |
| Mexico       | +       | +         | +            |
| Saudi Arabia | +       | +         | +            |

correlated (0.687). In the activity-related attributes, avg_tweets and statuses are strongly correlated (0.925), and avg_tweets and favorites are moderately correlated (0.641). We categorized the profile attributes into three groups in Section III-C. Correlation tends to be high within the same group. There is little correlation between the text-related attributes and other attributes.

Analysis showed that correlations between profile attributes were similar among different countries. It is thus reasonable to conduct an analysis based on the increases or decreases in the value of profile attributes.

C. PROFILE ATTRIBUTES CONTRIBUTING TO ESTIMATION

To analyze the profile attributes that contribute to estimation for each country, we filtered the social graphs using the two filters described in Section IV-A and evaluated precision using the filtered graph by leave-one-out cross-validation. The two filters require a threshold of \( \theta \) to filter users. We changed the threshold \( \theta \) from the minimum value to the maximum value of each attribute. In the experiment, all possible thresholds of screen_name, name, location, and description were selected. The threshold range of spent_days was divided into 100 equal sections. The threshold ranges of friends, followers, favorites, statuses, listed, avg_tweets, and follow_ratio were divided into 100 equal sections after they had been logarithmically transformed. We combined two filters, 100 thresholds, and all attributes.

The results are shown as follows. First, we find the threshold when the maximum precision is achieved and coverage is greater than 0.05 for each filter and attribute. Then, we test whether the precision is significantly improved by the filtering. The result is marked as “+” if the precision is significantly better than the precision without filter for a given condition. If we find that the precision is significantly improved, we consider the attribute as contributing to the estimation performance.

Significance testing is conducted as follows. Let \( n \) be the number of users whose location is estimated, and \( x \) be the number of users who have correctly estimated their home location. Precision \( p \) is calculated as \( p = x/n \). When \( n \) users are randomly sampled, the distribution of the precision
TABLE 4. Testing result of precision improvement by filtering using text-related attributes. ”+” means that precision with filtering using the attribute was improved over precision without filtering. Long name and description contributed to difficulty in estimation.

| Country/Cut    | screen_name | name | location | description |
|---------------|-------------|------|----------|-------------|
|               | Low         | High | Low      | High        | Low | High | Low | High |
| United States | +           | +    | +        | +           | +   | +    | +   | +    |
| Brazil        | +           | +    | +        | +           | +   | +    | +   | +    |
| United Kingdom| +           | +    | +        | +           | +   | +    | +   | +    |
| Japan         | +           | +    | +        | +           | +   | +    | +   | +    |
| Philippines   | +           | +    | +        | +           | +   | +    | +   | +    |
| Turkey        | +           | +    | +        | +           | +   | +    | +   | +    |
| Indonesia     | +           | +    | +        | +           | +   | +    | +   | +    |
| India         | +           | +    | +        | +           | +   | +    | +   | +    |
| Mexico        | +           | +    | +        | +           | +   | +    | +   | +    |
| Saudi Arabia  | +           | +    | +        | +           | +   | +    | +   | +    |

TABLE 5. Testing result of precision improvement by filtering using activity-related attributes. Activity-related attributes contributed to estimation performance.

| Country/Cut    | favorites | listed | statuses | spent_days | avg_tweets |
|---------------|-----------|---------|----------|------------|------------|
|               | Low | High | Low | High | Low | High | Low | High |
| United States | +         | +     | +    | +    | +   | +    |
| Brazil        | +         | +     | +    | +    | +   | +    |
| United Kingdom| +         | +     | +    | +    | +   | +    |
| Japan         | +         | +     | +    | +    | +   | +    |
| Philippines   | +         | +     | +    | +    | +   | +    |
| Turkey        | +         | +     | +    | +    | +   | +    |
| Indonesia     | +         | +     | +    | +    | +   | +    |
| India         | +         | +     | +    | +    | +   | +    |
| Mexico        | +         | +     | +    | +    | +   | +    |
| Saudi Arabia  | +         | +     | +    | +    | +   | +    |

follows a normal distribution with \( p \) mean and \( p(1-p)/n \) variance: \( p \sim N(p, p(1-p)/n) \). Then, we calculate the confidence interval of \( p \). We judge that the precision has improved if the precision is better than that without filtering and the 95% confidence interval does not overlap with that without filtering.

The results for degree-related, text-related, and activity-related attributes are shown in Table 3, 4, and 5, respectively.

In Table 3, the precisions are improved using the graph-related attributes friends, followers, and follow_ratio. The users with high friends and low followers are difficult to estimate in many countries. The results indicate that the home locations of users with high follow_ratio are difficult to estimate in all countries. All attributes related to the shape of the social graph are effective in the estimation. In particular, follow_ratio is the only attribute that is effective in all countries.

As shown in Table 4, precision is improved using the text-related attributes name, location, and description. The results show that the locations of users with long names, location texts, and description texts are difficult to estimate in certain countries. This result indicates that text length affects the estimation performance regardless of language. The results also show that the locations of users with short location texts and descriptions are difficult to estimate in some countries. In addition, screen_name contributes to estimation performance in few countries. The results suggest that older users are more likely to have shorter usernames and new users tend to have longer usernames. However, new users can choose a short username, and there is not much difference between countries.

In Table 5, precision is improved using the activity-related attributes favorites, listed, statuses, spent_days, and avg_tweets. The locations of users with high favorites, low listed, low statuses, high spent_days, and low avg_tweets tend to be difficult to estimate. The results indicate that either too low or too high activity on social media can make estimation difficult.

VI. DISCUSSION

We compared the estimation performances between countries to evaluate each country's dataset. We then analyzed the relationships between user profiles and their relationships using home location estimation. We found that estimation performance varies by country, and graph-related attributes, text-related attributes, and activity-related attributes affect the difficulty of estimation.

In this section, we discuss the background of the attributes contributing to graph-based location estimation. First, we discuss why each profile attribute contributed to the estimation performance. In addition, we discuss the fact that different attributes were obtained for different countries, focusing on the language used in each country.

A. FACTOR OF ESTIMATION DIFFICULTY

We can break down the difficulty of estimation into two factors. One is that there are few clues for estimation, and the other is that clues for estimation do not satisfy the closeness assumption.

The graph-based home location estimation method cannot estimate any home location if there is no edge, which
TABLE 6. Language usage by country aggregated from geotagged tweets.

| Country            | First Language | #tweets | %   | Second Language | #tweets | %   | Total |
|--------------------|----------------|---------|-----|----------------|---------|-----|-------|
| United States      | English        | 352,035,944 | 0.84 | undefined       | 42,846,417 | 0.10 | 417,880,361 |
| Brazil             | Portuguese     | 189,045,554 | 0.85 | undefined       | 17,620,236  | 0.08 | 222,665,790 |
| United Kingdom     | English        | 62,046,023  | 0.85 | undefined       | 6,566,395   | 0.09 | 72,612,418  |
| Japan              | Japanese       | 84,094,749  | 0.94 | undefined       | 2,197,056   | 0.02 | 89,951,805  |
| Philippines        | Tagalog        | 35,545,508  | 0.58 | English         | 16,989,051  | 0.28 | 61,545,259  |
| Turkey             | Turkish        | 38,855,627  | 0.84 | undefined       | 4,092,457   | 0.09 | 46,978,084  |
| Indonesia          | Indonesian     | 31,388,056  | 0.75 | English         | 9,528,756   | 0.28 | 40,916,812  |
| India              | English        | 15,086,743  | 0.44 | Hindi           | 3,946,467   | 0.09 | 19,033,210  |
| Mexico             | Spanish        | 23,536,026  | 0.78 | undefined       | 3,199,624   | 0.11 | 30,038,648  |
| Saudi Arabia       | Arabic         | 26,190,880  | 0.81 | undefined       | 3,747,522   | 0.12 | 32,938,402  |

is the estimation clue [15]. The results in Table 3 indicate that the home locations of users with low degrees (low friends or low followers) are difficult to estimate. Low activity such as statuses and avg_tweets is associated with low followers. Therefore, users with low activity would have few estimation clues. Although our datasets contain only users who posted tweets at least 10 times, the results showed a tendency that the locations of users with lower activity than others are more difficult to estimate. The effects of these characteristics on difficulty in estimation hold in almost all countries.

Even if there are edges, the estimation will fail if the relationship does not meet the assumption. For example, celebrities or influencers may have relationships that do not relate to geographic distance because they have many followers who are not friends. We found that high follow_ratio is effective in all countries. A previous study [18] reported that location estimation is difficult in the cases of two types of users, hub users and authority users, as described in Section II. Authority users are considered to be celebrities or influencers. Hub users are users with high friends. Thus, our results are consistent with the previous study.

Locations of other users besides celebrities and influencers were also difficult to estimate. The home locations of users with long self-declared text (i.e., high name or description) were difficult to estimate in some countries. We suppose that users easily make online relationships if they reveal more information about themselves. Twitter can be viewed as an interest network: users with the same interests connect online. However, online relationships do not reflect geographic distance, so their locations are difficult to estimate. Our results suggest that the amount of self-declared information is related to the tendency to develop online relationships, which affects the difficulty of location estimation in some countries.

We found two factors affecting estimation difficulty. Moreover, the characteristics of users whose locations cannot be estimated owing to lack of estimation clues were clarified. Further research is needed on the characteristics of users whose locations are difficult to estimate because they do not meet closeness assumptions, other than celebrities or influencers.

B. LANGUAGE USAGE

There are no borders on the Internet, but language can be a barrier to the distribution of information. We investigated languages used in each country and counted the number of tweets and their language information for each country from the geotagged tweets in 2019 collected in Section III-A. Language information was collected from the lang field of the Tweet object, and country information was collected from the country field of the Place object.

Table 6 shows the top two languages used in each country. We observe that 80% of the tweets are posted in the official language of the country, except for India and the Philippines. In India, 44% of tweets are in English and 28% are in Hindi. In the Philippines, 58% are in Tagalog and 28% are in English. Both India and the Philippines are unique in that they have more than one official language, including English.

The countries with majority of tweets in English are the United States, the United Kingdom, India, and the Philippines. If the same language is used in these countries, the profile attributes contributing to the estimation, especially text-related attributes, should be similar. Thus, not all the results are the same. The agreement between the results of profile attributes contributing to estimation difficulty in the United States and the United Kingdom is two-thirds. This suggests that there are cultural differences among countries in which the same languages are used.

VII. CONCLUSION

In this paper, we collected tweet data from 10 countries to investigate user profiles with characteristics of users’ relationships through graph-based home location estimation. First, (RQ1) we found that the absolute value of location estimation accuracy differs by country. The number of areas in the country or the average size of areas could not completely explain the difference. Second, we focused on the profile attributes contributing to the estimation. (RQ2) We found that profile attributes contributed to estimation difficulty and that the attributes varied between countries. To understand the results, we analyzed the rank correlation between attributes and found that the correlation between the attribute groups tends to be small. Then, we interpreted the result. (RQ3) We found two factors of estimation difficulty: lack of estimation clues and unmet
estimation assumption. The home location of users with low degrees or low activity tended to be difficult to estimate owing to a lack of clues. This held true in the cases of almost all countries. Nevertheless, we found that follow_ratio uniformly affects estimation difficulty. We also found that the amount of self-declared information is correlated with the difficulty of estimation because users who share more information tend to make more online relationships. Because these text-related attributes did not correlate with follow_ratio, some cultural aspects of the corresponding countries are captured. We then investigated the languages used in each country. (RQ4) The text-related profile attributes contributing to estimation difficulty are different even between countries where English is used. This suggests that profile attributes can differ owing to cultural differences among countries in which the same language is used.

REFERENCES

[1] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: Detecting influenza epidemics using Twitter,” in Proc. 2011 Conf. Empirical Methods Natural Lang. Process., 2011, pp. 1568–1576.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: Real-time event detection by social sensors,” in Proc. 19th Int. Conf. World Wide Web (WWW), 2010, pp. 851–860.

[3] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, “Geo-located Twitter as proxy for global mobility patterns,” Cartogr. Geograph. Inf. Sci., vol. 41, no. 3, pp. 260–271, May 2014.

[4] X. Zheng, J. Han, and A. Sun, “A survey of location prediction on Twitter,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 9, pp. 1652–1671, Sep. 2018.

[5] X. Luo, Y. Qiao, C. Li, J. Ma, and Y. Liu, “An overview of microblog user geolocation methods,” Inf. Process. Manage., vol. 57, no. 6, Nov. 2020, Art. no. 102375.

[6] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: A content-based approach to geo-locating Twitter users,” in Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), 2010, pp. 759–768.

[7] L. Backstrom, E. Sun, and C. Marlow, “Find me if you can: Improving geographical prediction with social and spatial proximity,” in Proc. 19th Int. Conf. World Wide Web (WWW), 2010, pp. 61–70.

[8] D. Jurgens, T. Finethy, J. Mccorriston, Y. T. Xu, and D. Ruths, “Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice,” in Proc. 9th Int. AAAI Conf. Web Social Media, 2015, pp. 188–197.

[9] J. Alonso-Lorenzo, E. Costa-Montenegro, and M. Fernandez-Gavilanes, “Language independent big-data system for the prediction of user location on Twitter,” in Proc. IEEE Int. Conf. Big Data (Big Data), Dec. 2016, pp. 2437–2446.

[10] Z. Wood-Doughty, M. Smith, D. Broniatowski, and M. Dreidze, “How does Twitter user behavior vary across demographic groups?” in Proc. 2nd Workshop NLP Comput. Social Sci., 2017, pp. 83–89.

[11] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser, “A multi-indicator approach for geolocalization of Tweets,” in Proc. 7th Int. AAAI Conf. Weblogs Social Media, 2013, pp. 573–582.

[12] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, “Unifying text, metadata, and user network representations with a neural network for geolocation prediction,” in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, 2017, pp. 1260–1272.

[13] J. Mahmud, I. Nichols, and C. Drews, “Home location identification of Twitter users,” ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, p. 47, 2014.

[14] J. McGee, J. Caverlee, and Z. Cheng, “Location prediction in social media based on tie strength,” in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM), 2013, pp. 459–468.

[15] A. Rahimi, T. Cohn, and T. Baldwin, “Twitter user geolocation using a unified text and network prediction model,” in Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process., 2015, pp. 630–636.

[16] M. Ebrahimi, E. ShafieiBavani, R. Wong, and F. Chen, “Twitter user geolocation by filtering of highly mentioned users,” J. Assoc. Inf. Sci. Technol., vol. 69, no. 7, pp. 879–889, Jul. 2018.

[17] H. Tian, M. Zhang, X. Luo, F. Liu, and Y. Qiao, “Twitter user location inference based on representation learning and label propagation,” in Proc. Web Conf., Apr. 2020, pp. 2648–2654.

[18] S. Hironaka, M. Yoshida, and K. Umemura, “User’s centrality analysis for home location estimation,” in Proc. IEEE/WIC/ACM Int. Conf. Web Intell., Oct. 2019, pp. 59–63.

[19] L. Sloan and J. Morgan, “Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter,” PLoS ONE, vol. 10, no. 11, Nov. 2015, Art. no. e0142209.

[20] L. Hedayatifar, A. J. Morales, and Y. Bar-Yam, “Geographical fragmentation of the global network of Twitter communications,” Chaos, Interdiscip. J. Nonlinear Sci., vol. 30, no. 7, Jul. 2020, Art. no. 073133.

[21] T. Tominaga, Y. Hiijkata, and J. A. Konstan, “How self-disclosure in Twitter profiles relate to anonymity consciousness and usage objectives: A cross-cultural study,” J. Comput. Social Sci., vol. 1, no. 2, pp. 391–435, Sep. 2018.

[22] C. A. Davis, Jr., G. L. Pappa, D. R. R. de Oliveira, and F. de L. Arcanjo, “Inferring the location of Twitter messages based on user relationships,” Trans. GIS, vol. 15, no. 6, pp. 735–751, Dec. 2011.

[23] R. Wesslen, S. Nandu, O. Eltayeby, T. Gallicano, S. Levens, M. Jiang, and S. Shaikh, “Bumper stickers on the Twitter highway: Analyzing the speed and substance of profile changes,” in Proc. 12th Int. AAAI Conf. Web Social Media, Jun. 2018, pp. 696–699.

[24] J. Shima, M. Yoshida, and K. Umemura, “When do users change their profile information on Twitter?” in Proc. IEEE Int. Conf. Big Data (Big Data), Dec. 2017, pp. 3119–3122.