

Article

Constructing Geographic Dictionary from Streaming Geotagged Tweets

Jeongwoo Lim *, Naoko Nitta, Kazuaki Nakamura and Noboru Babaguchi

Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan;
naoko@comm.eng.osaka-u.ac.jp (N.N.); k-nakamura@comm.eng.osaka-u.ac.jp (K.N.);
babaguchi@comm.eng.osaka-u.ac.jp (N.B.)

* Correspondence: jeongwoolim@nanase.comm.eng.osaka-u.ac.jp; Tel.: +81-6-6879-7746

Received: 28 March 2019; Accepted: 2 May 2019; Published: 8 May 2019



Abstract: Geographic information, such as place names with their latitude and longitude (lat/long), is useful to understand what belongs where. Traditionally, Gazetteers, which are constructed manually by experts, are used as dictionaries containing such geographic information. Recently, since people often post about their current experiences in a short text format to microblogs, their geotagged (tagged with lat/long information) posts are aggregated to automatically construct geographic dictionaries containing more diverse types of information, such as local products and events. Generally, the geotagged posts are collected within a certain time interval. Then, the spatial locality of every word used in the collected geotagged posts is examined to obtain the *local words*, representing places, events, etc., which are observed at specific locations by the users. However, focusing on a specific time interval limits the diversity and accuracy of the extracted local words. Further, bot accounts in microblogs can largely affect the spatial locality of the words used in their posts. In order to handle such problems, we propose an online method for continuously update the geographic dictionary by adaptively determining suitable time intervals for examining the spatial locality of each word. The proposed method further filters out the geotagged posts from bot accounts based on the content similarity among their posts to improve the quality of extracted local words. The constructed geographic dictionary is compared with different geographic dictionaries constructed by experts, crowdsourcing, and automatically by focusing on a specific time interval to evaluate its quality.

Keywords: geographic information; social network services; streaming information; online information extraction

1. Introduction

Geographic information, such as place names with their latitude and longitude (lat/long), is useful to understand what belongs where in the real world. The traditional geographic dictionaries called Gazetteers [1] are typically constructed by experts based on the information collected from official government reports. While the gazetteers contain the information about the popular administrative regions, such as states, counties, cities, and towns, much larger-scale geographic dictionaries containing the information about smaller-scale areas, such as beaches, parks, restaurants, and stadiums, have been constructed manually by crowdsourcing (e.g., GeoNames [2] and OpenStreetMap [3]). These dictionaries are often used for geoparsing [4–6], which is to extract place names in texts, so that the geographic coordinates can be assigned to the texts. Various types of place names can be extracted by looking up the dictionaries; however, the information provided by these geographic dictionaries is still limited to place names.

Since people often post about their experiences to photo-sharing services and microblog services, such as Flickr [7] and Twitter [8], their geotagged (tagged with lat/long information) posts can be aggregated to obtain more diverse types of geographic information, including local food, products, dialectal words, etc. One approach is to cluster a large number of geotagged posts based on their geotag, textual, and visual similarity [9–17]. This approach is generally useful for finding places attracting many people, where sufficient geotagged posts to form a cluster are posted. Another approach is to examine the spatial distribution of each word in the collected geotagged posts to extract *local words*, which indicate specific locations [18–27]. Since all words can be ranked based on certain types of scores representing their spatial locality, this approach is more suitable for discovering more diverse types of geographic information, including minor places. Further, additionally examining the temporal locality enables us to collect words representing events, which are observed at specific locations only at certain periods of time [28,29].

Since people often do not upload photos immediately after taking the photos, it is difficult to obtain real-time geographic information from photo-sharing services. On the other hand, microblog services often contain more real-time information due to the simple nature of their posting functions. Thus, researchers often apply the same techniques to microblog services to extract more up-to-date geographic information. They often use sliding time windows to check the temporal burstiness of local words or update the spatial distribution of each word [30–36]. However, one of the problems with the existing work is that the spatial locality of words is examined within a predetermined time window or a time window of fixed length. As a result, only the local words whose spatial distributions are localized within the given time window can be extracted. However, since the frequency of local words would depend on the popularity of places, events, etc., represented by the words, the suitable time window to examine the spatial locality should vary for each word. Another problem is that microblog services contain much more bot accounts compared to photo-sharing services. They are often used for providing information to a mass audience for specific purposes, such as advertisement, job recruiting, and weather forecasts. Since their posts are very similar, the spatial distributions of the words used in their geotagged posts can be largely distorted from their true distributions.

In order to solve these problems, this paper proposes an online method for constructing an up-to-date geographic dictionary by continuously collecting local words and their locations from streaming geotagged posts to Twitter (hereafter referred to as geotagged *tweets*) [37]. Our first idea is to record the usage history separately for each word. The usage history of each word is updated every time the word is used in the streaming geotagged tweet. It is accumulated until there is a sufficient number of tweets to examine the spatial locality of the word, which is equivalent to adaptively determining the suitable time window. When the spatial locality is either high or low enough, the word is determined as either a local word or general word, and its old usage history is deleted to be reinitialized. This enables us to repeatedly examine the recent usage history to accurately handle the temporal changes of its spatial distribution. Secondly, we validate the extracted words after removing similar tweets from the usage history so that the spatial locality of each word can be accurately examined by avoiding the influence of the tweets from bot accounts. The validated local words are then stored in a dictionary along with the posted texts containing the words, their posted time, geotags, and any accompanying images as the descriptions of the places, events, etc., represented by the words. Applying our proposed method to the streaming geotagged tweets posted from the United States in a month enabled us to continuously collect approximately 2,000 local words per day which represent many minor places, such as streets and shops; local specialties, such as food, plants, and animals; and current events, while forgetting the information about old events. The usefulness of our constructed geographic dictionary was shown by comparing with different geographic dictionaries constructed by experts, crowdsourcing, and automatically from the streaming tweets posted during a specific period of time and by visualizing the geographic information stored in our dictionary.

2. Related Works

Many methods have been proposed for extracting the geographic information from textual geotagged posts to Flickr and Twitter. For example, many researchers have collected a large number of geotagged posts from a specific area and tried to find the points of interest (POI) [11–14], areas of interest (AOI) [15,16], or regions of interest (ROI) [17] in the area by clustering the collected posts based on their geotag and textual similarity. Then, the words frequently used only in each cluster can be determined to describe the area represented by the cluster, for example, based on term frequency and inverse document frequency (TFIDF), which is widely used in the information retrieval and text mining field to find important words for each document in a corpus. The focus of this approach is often on the first step for accurately extracting regions attracting many people, which can be named afterwards.

Another direction is to extract *local words*, which indicate specific locations, by examining the spatial distribution of each word in a set of collected geotagged posts [18–27]. The focus of this approach is generally on the first step for accurately extract local words, whose corresponding regions are determined afterwards. The whole area where the geotagged posts are collected is often divided into sub-areas, such as cities [20,21,25] and grids of equal or varying sizes, to get the discretized spatial distributions of words [23]. The spatial distribution can also be estimated as the continuous probability density distribution, for example, by non-parametric models, such as Kernel Density Estimation [19], and parametric models, such as Gaussian Mixture Model [22,24], and the models represented with a focus and a dispersion [26,27]. Different types of score are then calculated from each distribution to represent its spatial locality, such as the entropy [19,20,23,24], CALGARI [25], geometric localness [22], TFIDF-based scores [20,21,23], dispersion [26], Ripley's K Statistic, and geographic spread [19,20]. Some scores are calculated considering the difference from global distribution which can be obtained from the distributions of users or stopwords. Such scores include χ^2 statistics, log-likelihood, information gain [19,20], Kullback–Leibler divergence [19,22], and total variation [22]. While the words are often ranked in the order of their scores to obtain the top-k words as the local words, Cheng et al. [27] used supervised methods to determine the local words. By using manually prepared local and non-local words as the training data, the local word classifier is trained based on the two estimated parameters for the spatial distribution, which represent the spatial focus and dispersion.

While these approaches have tried to extract stationary geographic information, such as landmarks, local products, and dialectal words, there is also a lot of research for extracting local events, which are temporary geographic information, from geotagged posts to Flickr and Twitter. Specifically, *local events*, which are defined as real-world happenings restricted to a certain time interval and location, are often detected as clusters of words describing the events. For example, Watanabe et al. [28] firstly find current popular places by clustering spatially close geotags posted within a recent specific time interval, and extract words from the geotagged posts in each cluster to describe the local event happening at the place represented by the cluster. Chen et al. [29] firstly extract local words representing local events based on the word usage distribution. They obtain the discretized 3-dimensional spatial and temporal distribution for each word, and after applying the Discrete Wavelet Transform to each dimension, find dense regions in the distribution based on the Wavelet coefficients. The words for which any dense region is found are determined as local words. Then, the local words are grouped based on their co-occurrence in the geotagged posts and spatial and temporal similarity. Although the information about local events can be extracted by additionally considering the temporal dimension, these methods are designed to be applied to the collected geotagged posts in a batch manner to extract past local events.

Since up-to-date information is constantly posted to Twitter, many online methods have been proposed for extracting the information about current local events in real time from geotagged tweets [30–35]. They often use overlapping or non-overlapping sliding time windows. The local events within a current time window are detected in similar ways as described above, and their reliability can be checked based on their temporal burstiness, which is examined by comparing the usage frequency of their descriptive words between the current and previous time windows.

The clusters are then updated by using the geotagged tweets in the next time window. While their goal is to detect clusters of geotagged tweets or words to sufficiently describe local events, Yamaguchi et al. have focused on accurately extracting local words by examining the difference between the spatial distributions of each word and users by considering their temporal changes [36]. In order to realize the real-time processing, the Kullback–Leibler divergence between the spatial distributions is updated by only considering the oldest geotagged tweet in the previous time window and the new geotagged tweet; thus the spatial distribution of each word is always examined for the fixed number of most recent geotagged tweets containing the word.

As discussed in this section, both the stationary and temporary geographic information is extracted either by finding dense clusters of geotagged posts based on textual, spatial, and temporal similarity or by extracting local words whose spatial and temporal distribution is highly localized. The latter approach is more suitable for the automatic construction of a geographic dictionary, since the extracted local words are often used for geoparsing or the location estimation of users or non-geotagged posts [18–22,26,27,38–42] in a similar way to manually constructed geographic dictionaries. Especially, in order to collect diverse types of geographic information, including minor places, events, etc., which are unlikely to form dense clusters, we propose a unified framework based on the latter approach for extracting local words representing both stationary and temporary geographic information. The proposed method contributes to efficiently and effectively construct an up-to-date geographic dictionary by:

- continuously extracting both popular and minor stationary and temporary local words by adaptively determining the time window for each word so that its spatial locality can be examined at the suitable timing.
- examining more accurate spatial distribution of each word by removing the geotagged tweets from bot accounts.

3. Proposed Method

The goal of this work is to construct a geographic dictionary using streaming tweets, which are geotagged with coordinates $x = (lat, lon)$. The geographic dictionary consists of local words l_k ($k \in \mathbb{N}$, where \mathbb{N} is a set of natural numbers), their associate sets of N_k geotags $G_k = \{x_{k,n} | n = 1, \dots, N_k\}$, the types of the local words: stationary or temporary, a set of tweets $S_k = \{w_{k,n} | n = 1, \dots, N_k\}$, and a set of images $V_k = \{I_{k,i} | 1 \leq i \leq N_k, i \in \mathbb{N}, \}$ representing the words. Additionally, for the temporary local words, the time of its first and last geotagged posts T_k^F and T_k^L is recorded to describe its observed time duration.

Basically, the proposed method extracts local words based on their spatial locality in the geotagged tweets collected during a certain time interval. However, the suitable time interval for examining the spatial locality depends on the popularity of the places, products, events, etc., represented by the words. Further, the same words can be used to represent events happening at different times and locations. Such different characteristics of the local words need to be considered to accurately extract diverse types of local words [37].

In order to handle the differences in when the spatial distribution gets localized among local words, the proposed method separately records the usage history of each word. Every time a geotagged tweet is received, the usage histories of the words from the received tweet are updated and the locality of each updated usage history is checked to determine if the corresponding word is a local word. As a result, the time interval to check the spatial locality is adaptively determined according to the usage pattern of each word and the local words can be added to the dictionary at the timing when their spatial locality gets high enough. For example, since the spatial locality of frequently used local words representing popular places, products, events, etc., would get high very quickly, these local words can be added to the dictionary soon after their usage histories are initialized. Additionally, even for the infrequently used local words representing less popular places, products, events, etc., their

usage histories would be kept until their spatial localities get high enough. As a result, if waited long enough, these infrequently used local words can also be added to the dictionary.

Further, in order to handle the temporal changes of the spatial distributions of temporary local words, only the recent usage history should be checked for each word. Thus, the usage history of a word is removed either when the spatial locality gets high enough for the word to be a local word or when the spatial locality gets low enough for the word to be determined as a general word which can be used anywhere. Clearing out the past usage history enables us to examine only the recent spatial locality for each word. This also enables us to check the temporal location changes for the same words, since the spatial locality of the same word is examined over and over with different timing. We generally consider local words as temporary ones when they are firstly added to the dictionary. Then, when the same word is determined as a local word again afterwards, its location consistency can be checked. If its location has changed, their geotags in the dictionary need to be updated. Only those whose spatial localities are consistently high at the same locations over a certain time duration are determined as stationary local words. If the local word in the dictionary is determined as a general word afterwards, its records need to be deleted as old temporary information.

Finally, Twitter has many bot accounts who often post tweets using similar formats. The spatial distributions of the words contained in the geotagged tweets from these bot accounts can be distorted from their true distribution. In order to examine the spatial distributions constructed only from the tweets posted by real users about their real-world observations, when a word is determined as a local word or general word, its spatial locality is reverified after removing similar tweets, which are likely from bot accounts, from its usage history.

To summarize, our proposed method consists of the following three steps, as shown in Figure 1:

(1) Location-based local/general word extraction

Every time a geotagged tweet is received, the usage histories of the words from the received tweet are either initialized or updated. Then, the spatial locality of the updated usage history is checked to determine if the corresponding word is a local or a general word.

(2) Content-based local/general word re-verification

For the word determined as a local or a general word, similar tweets, which are likely posted from bot accounts, are deleted from its usage history, and its spatial locality is verified again.

(3) Location consistency check

For the word determined as a local word, its location consistency over time is checked to determine if it is a stationary word.

The details of each step are explained in the following subsections.

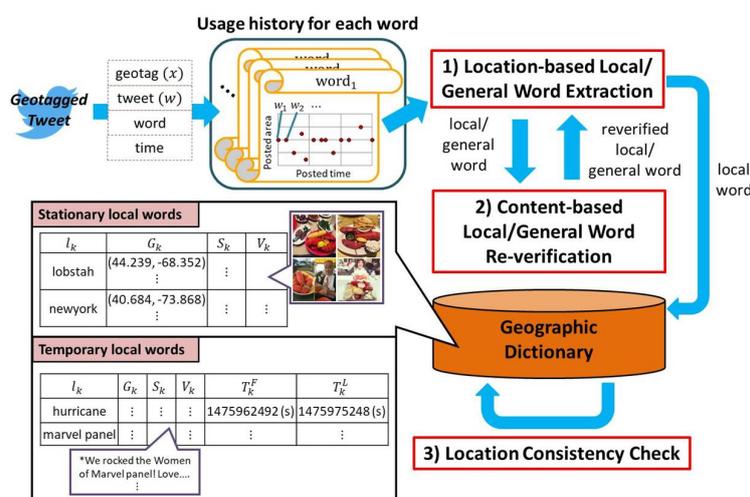


Figure 1. Overview of proposed method.

3.1. Preprocesses

In order to efficiently update the spatial distribution for each word and examine its spatial locality, the usage frequency histogram is used as the discretized spatial distribution. As a preprocess, given a set of all geotagged tweets posted during a certain period of time, the world is recursively divided into J areas so that each area $a_j (j = 1, \dots, J)$ has the same number of tweets. At each iteration, an area is divided into two subareas at the median point alternately for each axis (latitude and longitude). When a streaming tweet is received afterwards, which area the tweet is posted from is determined according to the latitude and longitude ranges of each area. In order to improve the reliability of the extracted local words, the usage frequency of each word is only updated at most once per user.

Further, we consider that the local words are mainly nouns, such as the names of places. We firstly remove URLs from each tweet, so that the links often posted in tweets would not be handled as the candidates for local words. Then, a part-of-speech tagging is applied to each tweet and only nouns are extracted [43]. Especially, compound nouns, which are the combinations of two or more words, often represent more restricted areas than their component words. For example, *Huntington beach* represents more restricted area than *Huntington* or *beach*. In order to extract meaningful local words, such as the names of places, the proposed method extracts compound nouns from each tweet as nouns [44]. Additionally, tweets often contain hashtags, which are the tags with # placed in front of a word or unspaced phrase. Since hashtags are often used to represent tweets with the same theme or content, we consider them as descriptive as the compound nouns. Thus, after extracting compound nouns, any alphanumeric nouns and hashtags are handled as the candidates for the local words in the following processes.

3.2. Location-Based Local/General Word Extraction

The proposed method separately records the usage history of each word. When a tweet attached with the geotag $x = (lat, lon)$ is received at the time t , for each of Z words $u_z (z = 1, \dots, Z)$ contained in the tweet, the tweet is added to its usage history along with its geotag and time. Additionally, a histogram of usage frequency $f_{z,j}$ in each area a_j is represented as $F_z = \{f_{z,j} | j = 1, \dots, J\}$, and when the received tweet was posted from the area $a_{\hat{j}} (x \in a_{\hat{j}})$, $f_{z,\hat{j}}$ is incremented by 1. When there is no usage history for u_z , the usage history is initialized with the tweet, the geotag x , the time t , and F_z set as $f_{z,\hat{j}} = 1$ and $f_{z,j} = 0$ for $\forall j \neq \hat{j}$.

Then, the TFIDF-based score SL_z , which reflects the spatial locality of u_z , is calculated based on F_z as follows:

$$SL_z = f_z^{mode} \cdot IDF_z \quad (1)$$

$$IDF_z = \log \frac{J}{|A_z|} \quad (2)$$

$$A_z = \{a_j | f_{z,j} \neq 0\}, \quad (3)$$

where f_z^{mode} is the mode of the area-based usage frequency of u_z and $|A_z|$ is the number of areas where u_z is used. SL_z gets higher when u_z is frequently used only in specific areas and gets lower when u_z is used in more areas. Thus, u_z which satisfies $SL_z \geq R$ is determined as the local word l_k , while u_z which satisfies $SL_z \leq r$ is determined as a general word, where R and r are the thresholds explained later. The usage history of u_z is deleted when u_z is determined as a local or general word, which corresponds to determining the end of the current time window for u_z , as shown in Figure 2. The time when u_z is used the next time would be the start of its new time window.

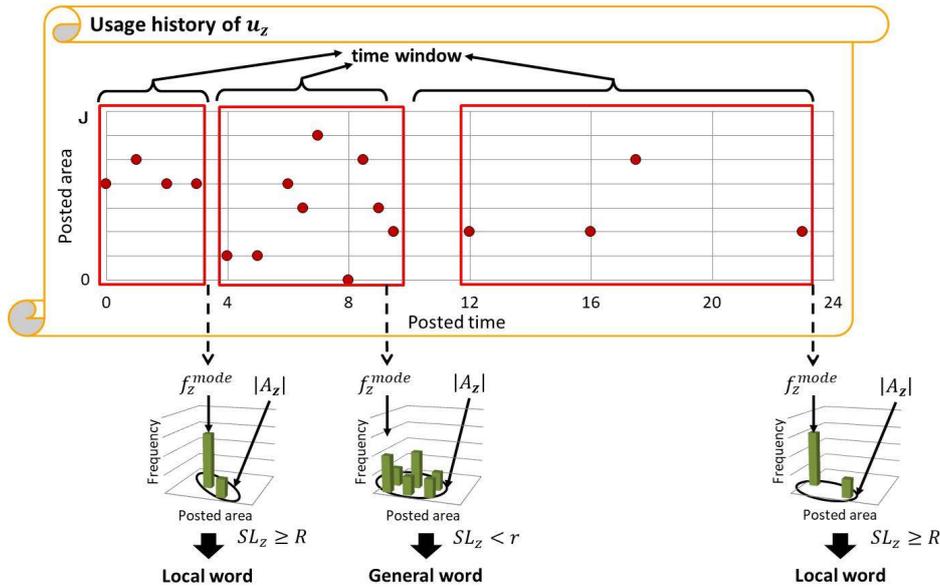


Figure 2. Usage history of u_z is accumulated for time windows of different duration so that its spatial locality is properly examined.

The proposed method basically waits until a word is used at least θ times in one of the areas ($f_z^{mode} = \theta$), where θ is the parameter which needs to be set as the lowest peak to determine a local word. Then, it waits until the peak of the spatial distribution gets high enough to be a local word or the spread of the spatial distribution gets wide enough to be a general word. The thresholds R and r are related to the maximum and minimum number of areas the local and general words can be used according to f_z^{mode} , respectively. As shown in Figure 3, assuming that a word can only be used in at most λ_{max}^θ areas to be a local word when $f_z^{mode} = \theta$, R can be determined as:

$$R = \theta \cdot \log \frac{J}{\lambda_{max}^\theta}. \tag{4}$$

Then, when $f_z^{mode} > \theta$, the word can be only used in at most $\lambda_{max}^{f_z^{mode}}$ areas ($|A_z| \leq \lambda_{max}^{f_z^{mode}}$) to be a local word. $\lambda_{max}^{f_z^{mode}}$ becomes larger for higher f_z^{mode} as follows:

$$\lambda_{max}^{f_z^{mode}} = \frac{J}{\exp \frac{R}{f_z^{mode}}}. \tag{5}$$

Similarly, assuming that a word needs to be used in more than λ_{min}^θ areas to be a general word when $f_z^{mode} = \theta$, r can be determined as:

$$r = \theta \cdot \log \frac{J}{\lambda_{min}^\theta}. \tag{6}$$

When $f_z^{mode} > \theta$, the word needs to be used in more than $\lambda_{min}^{f_z^{mode}}$ areas ($|A_z| \geq \lambda_{min}^{f_z^{mode}}$) to be a general word. $\lambda_{min}^{f_z^{mode}}$ becomes larger for higher f_z^{mode} , also determined by Equation (5) by replacing R with r . Setting θ low would help to extract the minor local words, such as the names of small places. Further, λ_{max}^θ and λ_{min}^θ are the parameters to be set to determine the thresholds R and r , which automatically determine $\lambda_{max}^{f_z^{mode}}$ and $\lambda_{min}^{f_z^{mode}}$, respectively. $\lambda_{max}^{f_z^{mode}}$ should be low to accurately extract local words, but not too low to extract local words which can be used in multiple areas. $\lambda_{min}^{f_z^{mode}}$ should not be neither too

low nor too high to extract the local words which can be used in multiple areas but to properly remove irrelevant words used in many areas as general words.

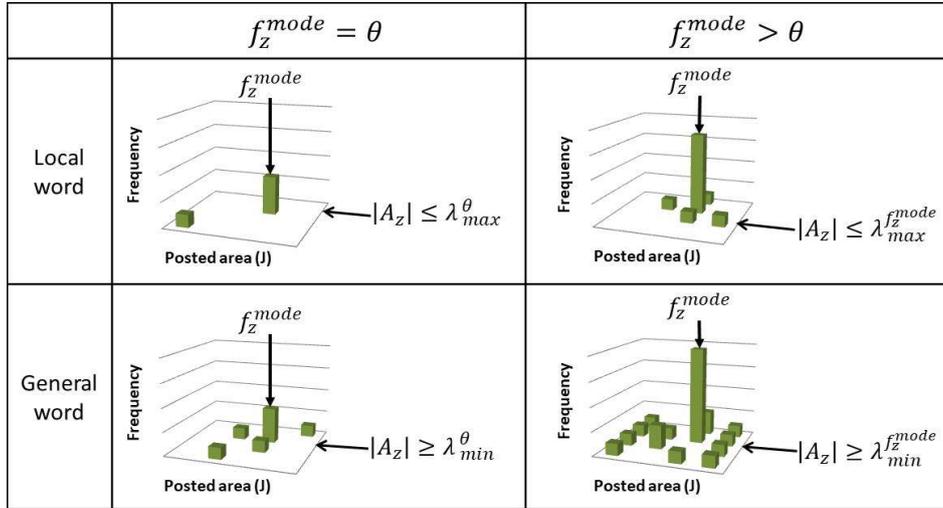


Figure 3. How the area-based frequency histogram of the word u_z is used to determine if u_z is a local or a general word. Intuitively, by using the same thresholds R and r for SL_z , the thresholds representing the maximum and minimum number of areas $\lambda_{max}^{f_z^{mode}}$ and $\lambda_{min}^{f_z^{mode}}$ for determining local/general words are changed according to the peak f_z^{mode} of the frequency histogram.

3.3. Content-Based Local/General Word Re-Verification

The geotagged tweets are sometimes posted from bot accounts which automatically post local advertisements, news, etc. Since their tweets are often written in similar formats, they can largely affect the spatial distributions of the words used in these tweets. Thus, when a word is determined as a local or general word, the similarity among the tweets in its usage history is checked to remove tweets which are likely to be from bot accounts.

Figure 4 shows how the tweets from bot accounts are removed from the usage history. When the word u_z is determined as a local or general word in the previous step, a set of tweets $S_z = \{w_{z,n} | n \in \mathcal{N}\}$ have been recorded in its usage history. Firstly, for each tweet $w_{z,n} \in S_z$, its maximum similarity to other tweets in S_z is calculated as the bot score $BS_{w_{z,n}}$ as follows. After removing URLs and mentions or replies (words starting with @, such as @username), each tweet $w_{z,n}$ is represented as a set of words $W_{z,n}$. The similarity between a pair of tweets $Sim^{con}(w_{z,n}, w_{z,m})$ is calculated using Jaccard Similarity [45] between the two sets of words. Thus, $BS_{w_{z,n}}$ is calculated as:

$$BS_{w_{z,n}} = \max_{w_{z,m} \in S_z, m \neq n} Sim^{con}(w_{z,n}, w_{z,m}) \tag{7}$$

$$Sim^{con}(w_{z,n}, w_{z,m}) = \frac{|W_{z,n} \cap W_{z,m}|}{|W_{z,n} \cup W_{z,m}|} \tag{8}$$

where $W_{z,n}$ represents a set of words composing the tweet $w_{z,n}$.

After calculating the bot score $BS_{w_{z,n}}$ for all $w_{z,n} \in S_z$, the tweet $w_{z,n}$ which satisfies $BS_{w_{z,n}} \geq Sim_{th}^{con}$ is removed from S_z , where Sim_{th}^{con} is a threshold to remove the similar tweets from the usage history. Let us note here that any similar tweets posted from real users, such as retweeted or quoted tweets can also be filtered out in this process.

After removing similar tweets, F_z is updated to calculate SL_z . If $SL_z \geq R$, u_z is determined as a local word l_k , and its usage history is reinitialized after updating the geographic dictionary. When l_k is added to the dictionary for the first time, the tweets and geotags recorded in its usage history are copied to the dictionary as $S_k = \{w_{k,n} | n = 1, \dots, N_k\}$ and $G_k = \{x_{k,n} | n = 1, \dots, N_k\}$, and the

time of the first and last tweets are recorded as T_k^F and T_k^L , respectively. When the tweets contain images, they are also added to the dictionary as $V_k = \{I_{k,i} | 1 \leq i \leq N_k, i \in \mathbb{N}\}$. If l_k is already in the geographic dictionary, its records are updated. If $SL_z \leq r$, its usage history is reinitialized; and, if u_z is in the geographic dictionary, it is removed as an old temporary local word together with its associated information.

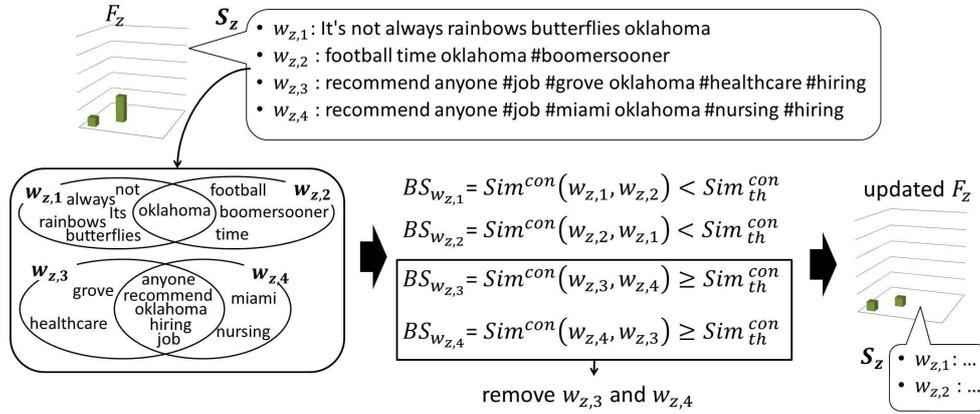


Figure 4. Area-based frequency histogram F_z of the word u_z is updated by removing the similar tweets from the set of tweets S_z in its usage history based on the bot score $BS_{w_{z,m}}$.

3.4. Location Consistency Check

The local words whose spatial distributions are localized at the same locations over a long period of time E are expected to represent stationary geographic information, which is consistently observed at the same locations. Let t_k represent the time duration of the usage history of the word l_k when it is determined as the local word for the first time. When $t_k \geq E$, where E is the threshold of the time duration, l_k is automatically determined as a stationary local word.

Otherwise, l_k is determined as a temporary local word, and its location consistency is examined when the same word is determined as the local word again. As shown in Figure 5, the location consistency is checked by comparing F_k^{old} and F_k^{new} which represent the histograms of the local word l_k in the dictionary and in its current usage history, respectively. If F_k^{old} and F_k^{new} are similar enough, F_k^{old} is updated by combining F_k^{new} . Otherwise, F_k^{old} is overwritten with F_k^{new} and T_k^F is reset to the time of the first tweet in the usage history. The tweets, geotags, and images in the usage history are also combined/overwritten to the dictionary accordingly, and T_k^L is updated as the time of the last tweet in the usage history. l_k is determined as a stationary local word if F_k^{old} and F_k^{new} are similar enough and the time duration of E has passed since T_k^F .

The histogram intersection is used as the similarity Sim^{loc} between F_k^{old} and F_k^{new} and they are considered similar enough when $Sim^{loc}(F_k^{old}, F_k^{new}) \geq Sim_{th}^{loc}$, where Sim_{th}^{loc} is a similarity threshold to determine the location consistency of the same word. F_k^{old} is updated as follows.

$$F_k^{old(\tau+1)} = \begin{cases} F_k^{old(\tau)} + F_k^{new} & (if\ Sim^{loc}(F_k^{old(\tau)}, F_k^{new}) \geq Sim_{th}^{loc}) \\ F_k^{new} & (otherwise), \end{cases} \tag{9}$$

where $F_k^{old(\tau+1)}$ and $F_k^{old(\tau)}$ represent the area-based frequencies in the dictionary after and before the update.

Finally, in order to forget the old temporary local words, any temporary local word l_k , for which the time duration of E has passed since T_k^L , is removed with its associated information.

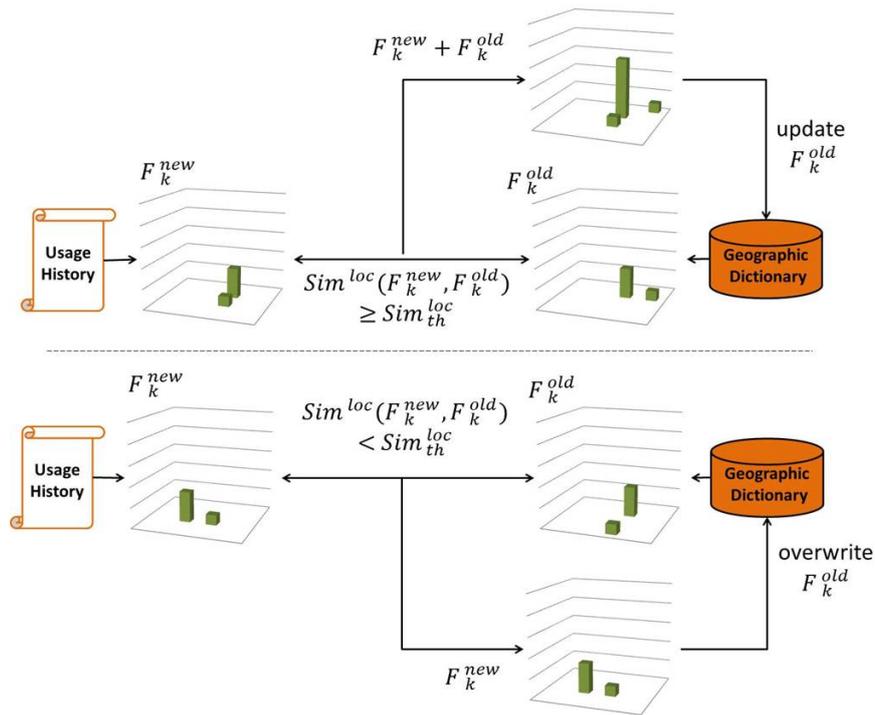


Figure 5. When the local word l_k in the geographic dictionary is determined as a local word again, its past area-based frequency histogram F_k^{old} recorded in the geographic dictionary is updated according to the similarity to F_k^{new} , which is its area-based frequency histogram in the current usage history.

4. Experiments

We evaluate our method by using 6,655,763 geotagged tweets posted during 30 days from September 2016 to October 2016 from the United States defined with the latitude and longitude ranges of [24, 49] and [−125, −66], respectively. Firstly, the effects of the parameters are evaluated by using the geotagged tweets on the first day in order to determine suitable parameter values. Then, the geographic dictionary is constructed iteratively over 30 days by using the determined parameters to evaluate the correctness of the extracted information.

4.1. Evaluations of Parameter Influence

Our proposed method has several parameters: the number of areas J to construct the area-based frequency histogram, R to extract local words, r to remove general words, Sim_{th}^{con} to remove tweets from bot accounts, and Sim_{th}^{loc} to determine the stationary local words. Here, we examined how changing the parameter values could affect the performance of our proposed method by using 215,885 geotagged tweets posted during the first day as the test set.

Firstly, as the parameter which is independent on other parameters, we examine how Sim_{th}^{con} affects the bot removal accuracy. Based on the assumption that bot accounts can post much more tweets during a day, we collected the tweets from accounts which posted more than 150 tweets as the tweets from bot accounts. The examples of the collected tweets are shown in Table 1. Further, the tweets from accounts which posted only a single tweet were also collected as the tweets from real users. For each tweet from the bot accounts and real users, we obtained its similarity to the most similar tweet from the users of the same category. Figure 6 shows the histograms of the obtained similarity for bot accounts and for real users. Figure 7 further shows the ratio of correctly removed tweets from bot accounts and the ratio of falsely removed tweets from real users when changing the similarity threshold Sim_{th}^{con} . Naturally, setting Sim_{th}^{con} low would remove tweets both from bot accounts and real users, while setting Sim_{th}^{con} high would keep tweets both from real users and bot accounts. Since we

want to remove as many tweets from the bot accounts without falsely removing the tweets from real users, $Sim_{th}^{con} = 0.4$, which gave the best results, is used in the following experiments. As shown in Figure 7, 99% of the tweets from bot accounts, which are not exactly the same, but similar to each other as shown in Table 1, were removed without falsely removing many tweets from real users.

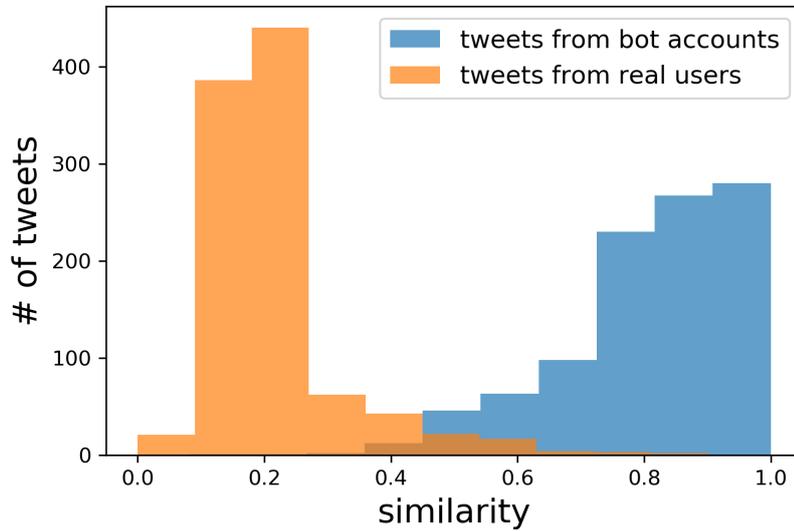


Figure 6. Maximum similarity for tweets from bot accounts and real users. Bot accounts tend to post similar tweets among themselves as shown in Table 1 (with the maximum similarity over 0.4), while real users tend to post unique tweets (with the maximum similarity under 0.4).

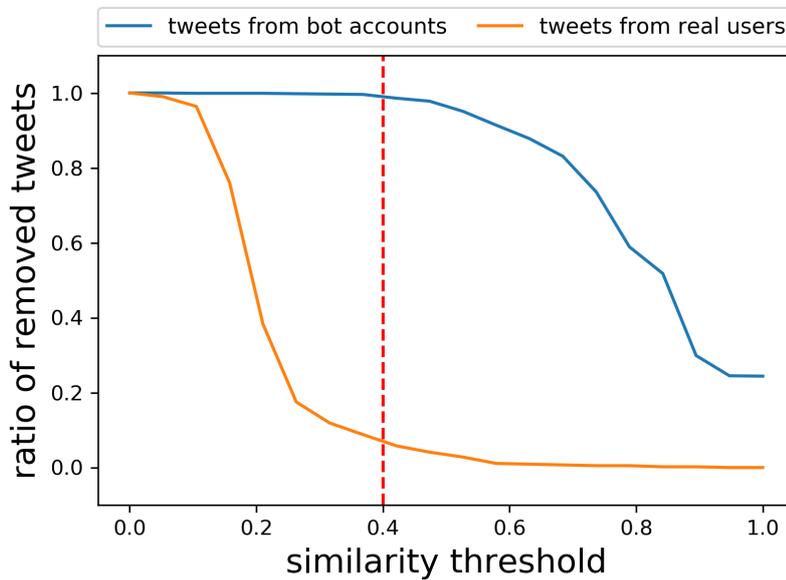


Figure 7. Ratio of correctly removed tweets from bot accounts and falsely removed tweets from real users. Setting $Sim_{th}^{con} = 0.4$ gave the best results, removing 99% of the tweets from bot accounts without falsely removing many tweets (less than 10%) from real users. Setting Sim_{th}^{con} higher would miss more tweets from bot accounts, while setting Sim_{th}^{con} lower would falsely remove tweets from real users.

Table 1. Examples of tweets from accounts which posted more than 150 tweets in a day. They are used as examples of tweets from bot accounts. ‘URL’ represents a link contained in the tweet.

Closed Homeless Concerns request at 240 Shotwell St URL. Case closed. case resolved. done.
Closed Street or Sidewalk Cleaning request at 640 Polk St URL. Case closed. case resolved. completed.
Closed Graffiti request at 106 Noe St URL. Case closed. case resolved. URL
Opened Parking and Traffic Sign Repair request via iphone at 222 Mason St URL. URL.
Opened Street or Sidewalk Cleaning request via iphone at 340 Stockton St URL. Wooden pallet. URL
Opened Graffiti request via iphone at 487 Church St URL. Planter graffitied. URL

Can you recommend anyone for this #job in #Tucson, AZ? URL #Nursing #Veterans #Hiring URL
Interested in a #job in #Providence, RI? This could be a great fit: URL #Physician #Veterans URL
Want to work in #Canton, OH? View our latest opening: URL #Job #Nursing #Veterans #Jobs #Hiring URL
Can you recommend anyone for this #job? Nurse Manager (OR) - URL #Nursing #Miami, FL #Veterans URL
Interested in a #job in #Charleston, SC? This could be a great fit: URL #Nursing #Veterans URL
If you’re looking for work in #SanDiego, CA, check out this #job: URL #Nursing #Veterans URL

‘Colin Powell’ just started trending with 15808 tweets. More trends at URL #trndnl
Trend Alert: ‘GOOD Music’. More trends at URL #trndnl URL
Washington was the city of United States with more Trends on Wednesday 7: URL #trndnl
54% of the United States’s Trends for Wednesday 7 were hashtags: URL #trndnl
On Wednesday 7, #NFLaFilm was Trending Topic in United States for 6 h: URL #trndnl
Trend Alert: ‘What’s the Value of Exercise’. More trends at URL #trndnl URL

Secondly, we collected place names from GeoNames [2] as the examples of local words and stop words [46] as the examples of general words. They were used as the test data to evaluate the effects of other parameters. As discussed in Section 3.2, R can be determined by setting the maximum number of areas λ_{max}^{θ} for a word to be a local word when $f_z^{mode} = \theta$. Since our goal is to obtain as many local words as possible, including those used by only a few users, we have set $\theta = 3$ and $\lambda_{max}^{\theta} = 2$, which means that when a word is used by at most three different users in one area ($f_z^{mode} = 3$), the word can be determined as a local word as long as the word is used in fewer than two out of J areas.

Table 2 shows how changing J affects the numbers of candidate place names and stop words which were used at least 3 times in one of the J areas ($f_z^{mode} \geq 3$), and the numbers of correctly/falsely extracted place names and stop words. More place names were extracted with smaller J ; however, more stop words were falsely extracted when J was too small. Based on the results, $J = 64$ was the best value to extract more local words without falsely extracting general words. Figure 8 shows how the United States was divided when $J = 64$.

Further, Figure 9 shows the histogram based on the number of areas for the place names and stop words when $f_z^{mode} = 3$ and $J = 64$. It can be seen that the place names tend to be used in much fewer areas; thus place names are much more localized than stop words. This histogram further verifies that $\lambda_{max}^{\theta} = 2$ is the appropriate threshold to collect the place names without falsely extracting stop words. Figure 9 also shows that most stop words can be used in more than 24 areas when $f_z^{mode} = 3$, which means that $\lambda_{min}^{\theta} = 24$ would be the appropriate threshold to determine general words. The thresholds corresponding to R and r when $f_z^{mode} = 3$ are shown with the red dashed lines in Figure 9. Words between the thresholds R and r need to wait for more usage history to be collected.

Table 2. Number of candidate local words and general words and correctly/falsely extracted place names and stop words when $\theta = 3$.

	J				
	16	32	64	128	256
# of candidate place names	3353	3115	2919	2766	2649
# of correctly extracted place names	2556	2495	2419	2366	2316
# of candidate stop words	181	148	124	97	75
# of falsely extracted stop words	15	4	1	2	1

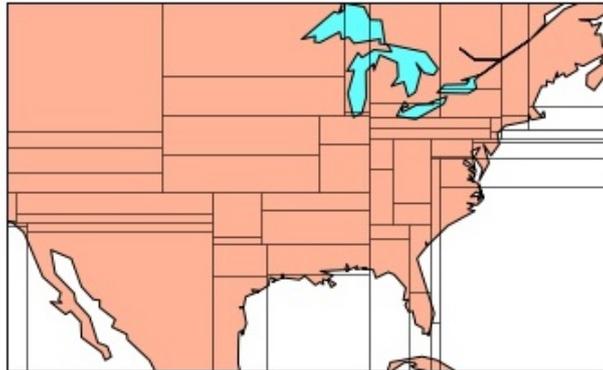


Figure 8. How the United State was divided into $J = 64$ areas.

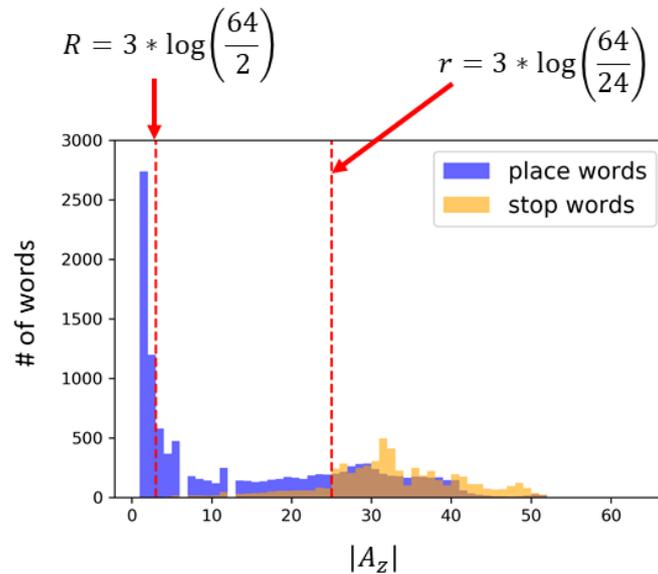


Figure 9. Histogram of the number of areas in which place names and stop words were used when $f_z^{mode} = \theta (= 3)$. Approximately 80% of the place names and less than 1% of stop words were used in fewer than two areas when $f_z^{mode} = 3$. On the other hand, approximately 70% of the stop words and 2% of the place names were used in more than 24 areas when $f_z^{mode} = 3$. The dashed red lines show these thresholds corresponding to R and r .

Figure 10 shows the relations of f_z^{mode} and $|A_z|$ in the usage history of the place names and stop words when $J = 64$. The curves are plotted by using the functions defined by Equation (5) and show the thresholds $\lambda_{z_{max}}^{f_z^{mode}}$ and $\lambda_{z_{min}}^{f_z^{mode}}$ when $R = 3 \cdot \log \frac{J}{2}$ and $r = 3 \cdot \log \frac{J}{24}$ which are determined by setting $\theta = 3$, $\lambda_{max}^\theta = 2$, and $\lambda_{min}^\theta = 24$. The words over $\lambda_{z_{min}}^{f_z^{mode}}$ are determined as general words and the words under $\lambda_{z_{max}}^{f_z^{mode}}$ are determined as local words. Words between the curves $\lambda_{z_{min}}^{f_z^{mode}}$ and $\lambda_{z_{max}}^{f_z^{mode}}$ need to wait for more usage history to be collected. As discussed above, over 80% of place names were correctly extracted as local words while only 1 stop word was falsely extracted. Further, 80% of stop words were correctly removed as general words while only 2% of place names were falsely removed. The removed place names were ‘accident’, ‘ball’, ‘blue’, ‘box’, ‘bright’, ‘campus’, ‘canon’, ‘center’, ‘chance’, ‘college’, ‘diamond’, ‘earth’, ‘energy’, ‘faith’, ‘freedom’, ‘garden’, ‘golf’, ‘grace’, ‘green’, ‘grill’, ‘honor’, ‘hope’, ‘joy’, ‘king’, ‘lake’, ‘lane’, ‘lucky’, ‘media’, ‘park’, ‘post’, ‘power’, ‘price’, ‘progress’, ‘short’, ‘star’, ‘start’, ‘story’, ‘strong’, ‘success’, ‘sunrise’, ‘sunshine’, ‘trail’, ‘university’, ‘veteran’, ‘wall’, ‘west’, ‘white’, ‘wing’, ‘winner’, ‘wood’, and ‘worth’. Although these words are in GeoNames, they can often be used in any locations. Thus, they can actually be considered as the correct removal.

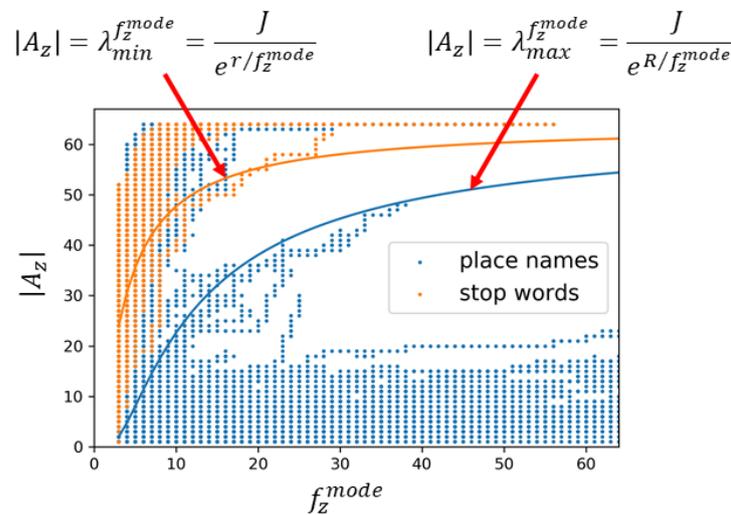


Figure 10. Relations between the area-based maximum frequency f_z^{mode} ($f_z^{mode} \geq \theta (= 3)$) and the number of areas $|A_z|$ for place names and stop words.

Finally, we extracted local words by setting $Sim_{th}^{con} = 0.4$, $J = 64$, $R = 3 \cdot \log \frac{J}{2}$, and $r = 3 \cdot \log \frac{J}{24}$. Since we do not have the ground truth for events which happened on the first day, we examined the similarity between F_k^{old} , the area-based frequency histogram in the dictionary, and F_k^{new} , the area-based frequency histogram in the recent usage history, to see the location consistency of the actual local words over time. Place names in GeoNames were used as the actual local words l_k . Figure 11 shows the histogram of the average similarity between F_k^{old} and F_k^{new} for the place names in GeoNames. The place names tend to be consistently posted from similar areas, and the similarities of the area-based histograms of the same place names during different periods of time were over 0.7 for 85% of the place names. Accordingly, we set $Sim_{th}^{loc} = 0.7$ for determining stationary local words in the following experiments.

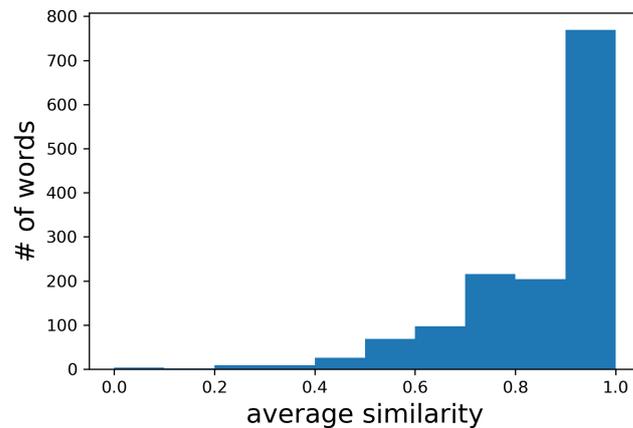


Figure 11. Average similarity between F_k^{old} and F_k^{new} for place names in GeoNames.

4.2. Comparisons with Other Dictionaries

Based on the results of the previous experiments, we extracted local words by setting $Sim_{th}^{con} = 0.4$, $J = 64$, $R = 3 \cdot \log \frac{J}{2}$, $r = 3 \cdot \log \frac{J}{24}$, $Sim_{th}^{loc} = 0.7$, and $E = 24$ h. Since the time duration for distinguishing between the stationary and temporary local words can vary largely and there is no ground truth (even a name of a country can be considered as temporary since it can be changed), we set $E = 24$ h by considering that most events do not last for more than one day. Figure 12 shows the number of local words extracted by the end of each day.

Most of the popular place names in GeoNames seem to have been determined as stationary local words in the first two weeks, while less popular place names in GeoNames were slowly added to the dictionary as stationary words afterwards. As the local words which are not in GeoNames, more and more stationary local words and temporary local words were consistently collected over time, as shown in Figure 12a and in the tail of Figure 12b. Figure 13 shows the histogram of the number of users for these local words and 95% of the extracted local words were used by fewer than 100 users. This verifies that the proposed method successfully extracted large number of minor local words.

In order to evaluate the correctness of the extracted local words, we compared our constructed dictionary with other geographic dictionaries, each of which was created differently. As manually created geographic dictionaries, we used Census 2017 U.S. Gazetteer [1] created by experts and GeoNames [2] created by crowdsourcing. Further, as the dictionary created from geotagged tweets, we constructed two dictionaries by applying Cheng's batch method [27] to a set of geotagged tweets posted during the first 10 days and 30 days. Cheng's method uses a classifier to determine if the spatial distribution of a word is localized or not. The classifier was trained by using the place names collected from Gazetteer and stop words as positive and negative training samples, respectively. In order to accurately estimate the spatial distribution of a word, words used more than 50 users during the first 10 days and 30 days were selected (referred to as Cheng_10days and Cheng_30days). Then, the spatial distributions were estimated [47] for these words to classify them into local/general words. We also added the place names in Gazetteer which were classified as the local words by the trained classifier. For each local word, the estimated center was determined as its location. Table 3 shows the number of local words in each dictionary.

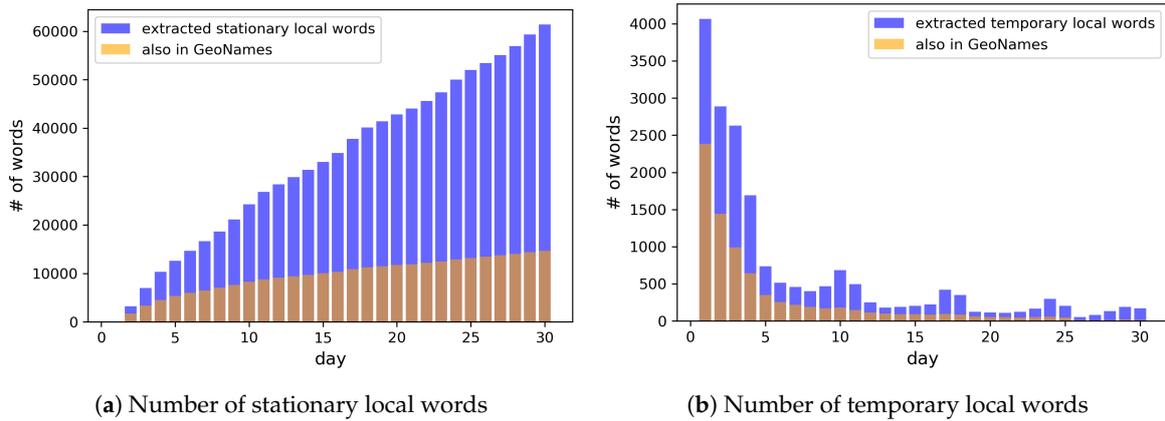


Figure 12. Number of extracted local words. A part of them are also in GeoNames.

Table 3. Numbers of local words in geographic dictionaries.

Proposed	GeoNames	Gazetter	Cheng_10days	Cheng_30days
0–61,546	2,189,612	29,072	2004	4732

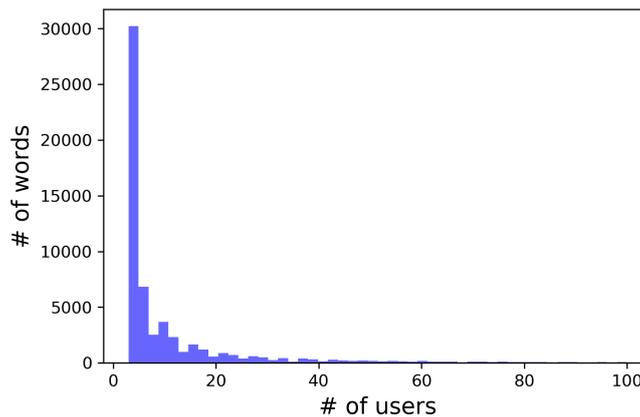


Figure 13. Number of users for the extracted local words. Those used by 5, 10, and 20 users accounted for 50%, 70%, and 80% of the extracted local words.

Firstly, we used GeoNames as the ground truth of local words and examined the correctness of their locations in the dictionaries constructed from geotagged tweets. Since manually constructed geographic dictionaries sometimes provide different locations for the same local words, we double-checked the locations in GeoNames with those provided by Google’s Geocoding API, and used the ones closest to the estimated locations as the ground truth. For the proposed method, the location for each local word was estimated as the center of its collected geotags, where the word is used most frequently. In the same way as Cheng’s method, the center was searched by dividing the whole area into grids of 1/10 of latitude and 1/10 of longitude, and then was estimated as the mean of the geotags within the most frequent grid. Figure 14 shows the cumulative distribution of the location errors for the local words extracted by the proposed method and Cheng’s method. Compared to Cheng’s method, our proposed method collected approximately 6 times more local words in GeoNames by examining the spatial locality in their suitable time windows. Although the ratio of the words with large errors increased slightly, the errors were within 10 km for 92% of the extracted local words.

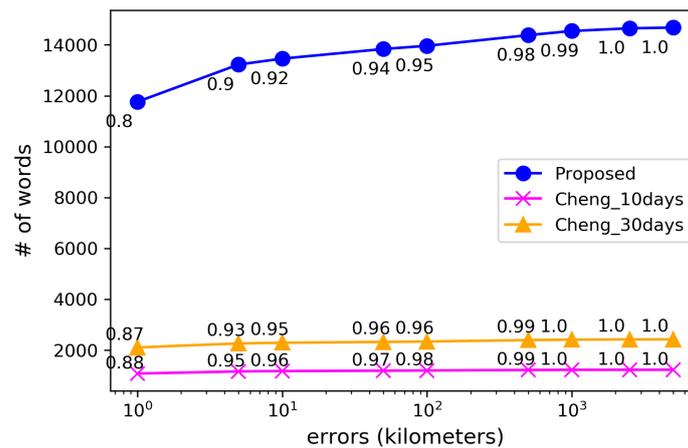


Figure 14. Location errors of the local words extracted by the proposed method and Cheng’s method, which are also in GeoNames. The numbers of the lines represent the ratio of the words with the corresponding errors.

In order to further evaluate the correctness of the extracted local words, we used the extracted local words and their locations for content-based tweet location estimation. We uniformly sampled 1% of geotagged tweets from each day, and after removing near-duplicates, obtained 36,626 geotagged tweets as the test tweets for the location estimation. The geotags of these test tweets are considered as the ground truth of their locations. The remaining 6,655,763 geotagged tweets were used for constructing the geographic dictionary. When given a test tweet, its location can be estimated if the tweet contains any word listed in the geographic dictionary. The location of the test tweet is estimated as the location of the word in the dictionary. When the test tweet contains several local words or there are multiple candidate locations for the same local word in the dictionary, the location closest to the ground truth is selected from the candidates when using GeoNames, Gazetteer, and the dictionary constructed by Cheng’s method. When using our dictionary, the local word which has been used in smallest number of areas is selected since such word is considered to indicate the most restricted area. Then, the location of the test tweet is estimated as the center of the geotags of the selected local word. Let us note that, since only the formal names of cities, such as *chicago city*, are listed in the Gazetteer, while more simple words, such as *chicago*, are often used in tweets, we removed words like ‘city’, ‘town’, ‘village’, and ‘cdp’, from the local words in the Gazetteer as a preprocess. When using our constructed dictionary, the location of each tweet was estimated by using the local words updated until the test tweet was posted.

Figures 15 and 16 show the estimation errors from days 1–10 and 11–30 when using each dictionary, respectively. The locations were estimated for the largest number of tweets by using GeoNames, since it has a largest number of local words obtained by crowdsourcing. However, the errors in the estimation using manually created dictionaries, such as GeoNames and Gazetteer, tend to be much larger than when using the dictionaries created from geotagged tweets, since the spatial locality of the words is not considered in constructing the dictionaries. With much fewer local words than Gazetteer, the dictionary constructed by Cheng’s method can estimate the location for much more tweets with small errors. With more diverse types of local words, the dictionary constructed by our method during the first 10 days already performed slightly better than the dictionary constructed by Cheng’s method using the geotagged tweets posted during the 30 days. In the last 20 days, the performance of our dictionary further improved, while that of the dictionary constructed by Cheng’s method using the geotagged tweets during the first 10 days degraded. Additionally, we have also provided the estimation results when using only stationary local words. In the first 10 days, the performance was much worse than when using all words since sufficient number of stationary words had not been collected yet. However, in the last 20 days, the performance got even better than when using all local words since the stationary

local words consistently indicate the same locations. The results have verified that setting E to 24 h was reasonable, and our proposed method can automatically extract the most diverse and accurate local words from the geotagged tweets.

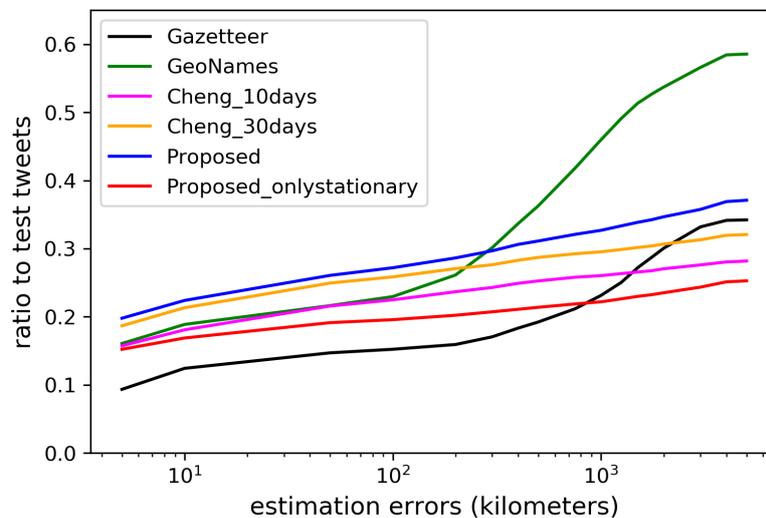


Figure 15. Tweet location estimation results from day 1 to day 10.

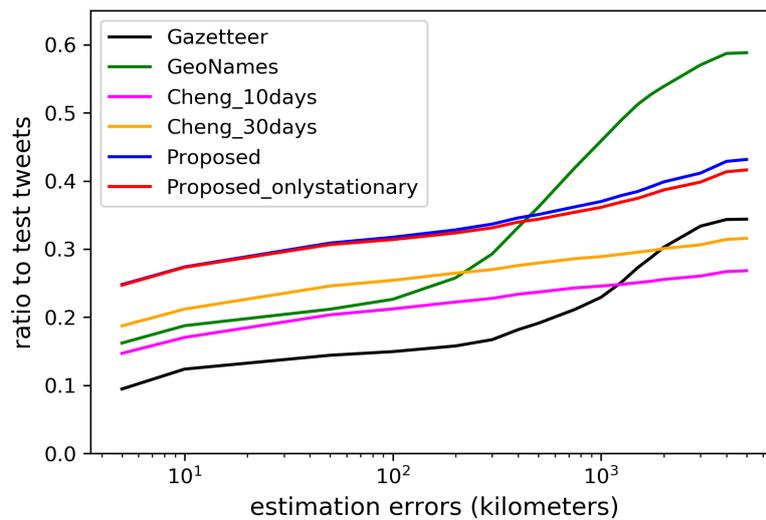


Figure 16. Tweet location estimation results from day 11 to day 30.

Table 4 shows the examples of the local words contained in each dictionary. Place names are contained in Gazetteer and GeoNames, but some of them can also be used in any location, such as *park* and *mountain*, which degraded the estimation accuracy. The dictionaries constructed from geotagged tweets contain diverse types of local words, such as unofficial place names, local specialties, and events, but not the general words contained in the Gazetteer and GeoNames. Especially, our proposed method can iteratively collect many minor local words, including new events, while forgetting local words representing old events.

Table 4. Examples of local words in different types of database (Gazetteer, GeoNames, Cheng_30days, and Proposed after 30 days).

Database	Local Words
all	pittsburgh, virginia, oakland, san diego, boston, nashville, hollywood, astoria, harlem, new york, brooklyn, manhattan, miami beach
only in GeoNames	able, aid, art, back, beauty, book, bridge, case, check, click, dad, dog, dream, fact, fall, fashion, flower, free, friday, game, good, hand, health, heart, kid, lady, land, life, light line, link, little, lord, love, meter, model, monday, music, need, next, night, place, ready, right, run, school, service, smile, spa, stage, station, stop, summer, sweet, table, thankful, thursday, top, traffic, vegan, view, wait, water, way, weather
Gazetteer and GeoNames	accident, ball, beach, blue, casa, chance, cool, day, early, friend, garden, goodnight, guy, happy, home, honor, house, jet, joy, long, loving, lucky, man, many, mountain, national park, nice, north, ocean, park, point, power, skyline, starbuck, story, sun, surprise, time, trail, west, wood
only in Cheng	firstdayofschool, 11 September, wewillneverforget, art festivalny fashion week preview, annual world aid day luncheon, formation world tour los angele, joe bear foundation golf tournament
Cheng and Proposed	worldtradecenter, wtc, statueofliberty, lady liberty, twin tower, downtown atlanta, splash mountain, 29 rooms, 49 er, 911 memorial, asu sun devil stadium, big apple, staten island ferry, crocodile, hhn26, washington state fair, comic con, great american beer festival, usopen, albuquerque international balloon fiesta, america chicago marathon
Proposed (stationary)	downtown grand junction, reunion tower, abraham lincoln memorial, cafe henrie, stone street, hu kitchen, amazon corporate headquarters, black cat alley, 101 coffeeshop, 102nd floor observatory, arnold arboretum of harvard university, domaine carneros winery
Proposed (temporary)	army tenmiler, atlanta pride parade, beach marathon, bolton fair, chalktoberfest, honkfest, joshua tree music festival, louisiana comic con, moon river music festival, rise festival, san francisco fleet week

4.3. Visualization of the Collected Geographic Information

Once we collect the geographic information composed of local words l_k , their associated sets of N_k geotags $G_k = \{x_{k,n} | n = 1, \dots, N_k\}$, the types of the local words: stationary or temporary, and a set of tweets $S_k = \{w_{k,n} | n = 1, \dots, N_k\}$ in a database, what is where in the real world can be automatically visualized interactively in web browsers by using D3.js [48], as shown in Figure 17. The visualization of the extracted local words can be found in <http://www2c.comm.eng.osaka-u.ac.jp/~lim/index.html>. Some tweets cannot be embedded due to the deactivation or privacy setting changes of accounts.

Firstly, when we access the web page, the list of local words can be queried from the database by the types or the extracted days of the local words. Then, when we select a word l_k from the list of the retrieved local words $L = \{l_k | k \in \mathbb{N}\}$, its location is overlaid on a map as a heatmap by using its set of geotags G_k . Further, its set of tweets S_k are also presented with the images V_k if any, which are embedded using Twitter API [49]. For example, Figure 17 shows an example when we select the local word *sleater kinney*. From the locations and tweets, we can know the local word *sleater kinney* is the name of a rock and roll band and they performed at the Riot Fest, which was held at the Douglas Park in Chicago, IL, on 18 September 2016.

Figures 18 and 19 show the locations of some stationary and temporary local words only in our constructed geographic dictionary with their images. The images in Figure 18 show what each stationary local word represents. The images show that the local words in the green boxes, such as *carterlakenationalpark*, *custertestatepark*, *chicagobotanicgarden*, and *birmingham botanical garden*, represent parks and gardens, and those in the blue boxes, such as *denver bronoco mile*, *abshire stadium*, *kemper museum*, and *bowlero los angeles*, represent stadiums, a museum, and a local bowling alley. The images for the local words in the orange boxes, such as *44 restaurantbar*, *biscuitbitch*, *atro coffee*, *10 belowicecream*, *abracadabar*, and *lobstah*, show some examples of food served at each restaurant, bar, etc., or of local foods. Further, the temporary local words usually represent different events happening at different locations each day. The images and temporary local words in Figure 19 show examples of events happened during the last 3 days. For example, a marathon *mychicagomathon*, a contest *nycomiccon2016*,

a festival *greatamericabeerfestival*, and a hurricane *huracanmathew* happened on the 28th day, a speech by *amy cudy*, festivals *avofest2016* and *rise festival*, and an sports event at *spence park* happened on the 29th day, and marathons *eastbay510k* and *armytenmiler* and festivals *chalktoberfest* and *rise festival* happened on the 30th day at the locations shown on the map. Figure 20 shows an example of temporary local words which change their locations on different days. *beyoncé formation world tour* is the title of a concert tour during which Beyoncé performed at different locations at different days. Even though the same local word was posted several times at different locations at different days, our proposed method can extract these types of local words at their suitable timing by repeatedly examining their spacial locality within different time windows.

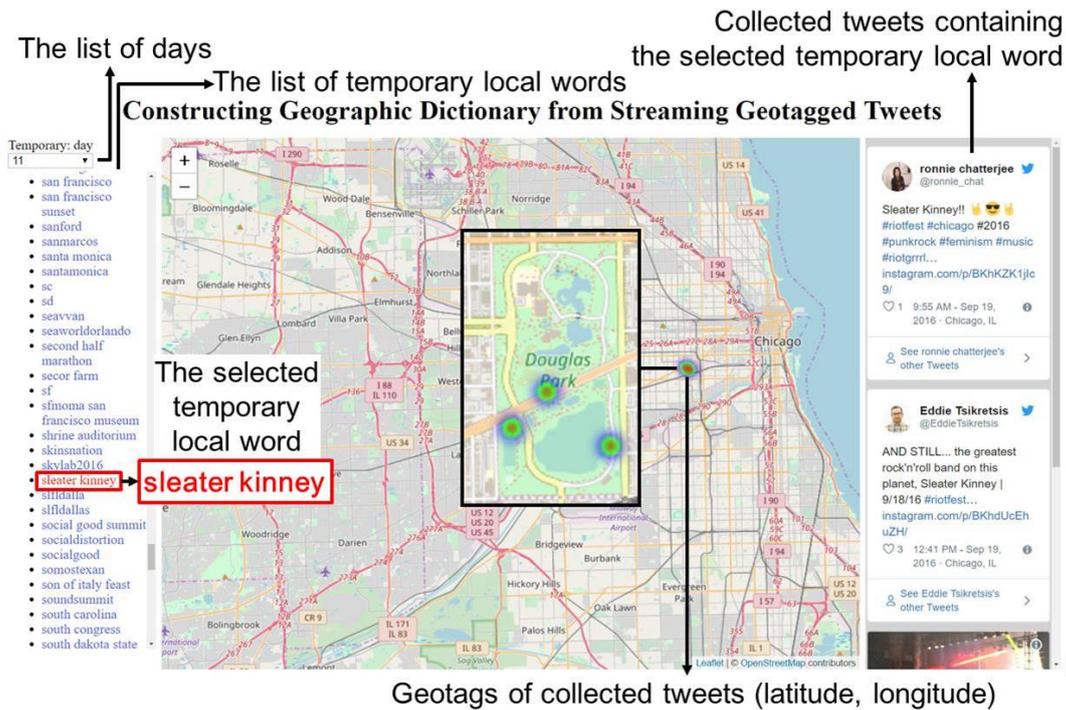


Figure 17. Visualization of collected geographic information.



Figure 18. Examples of stationary local words only in our geographic dictionary.

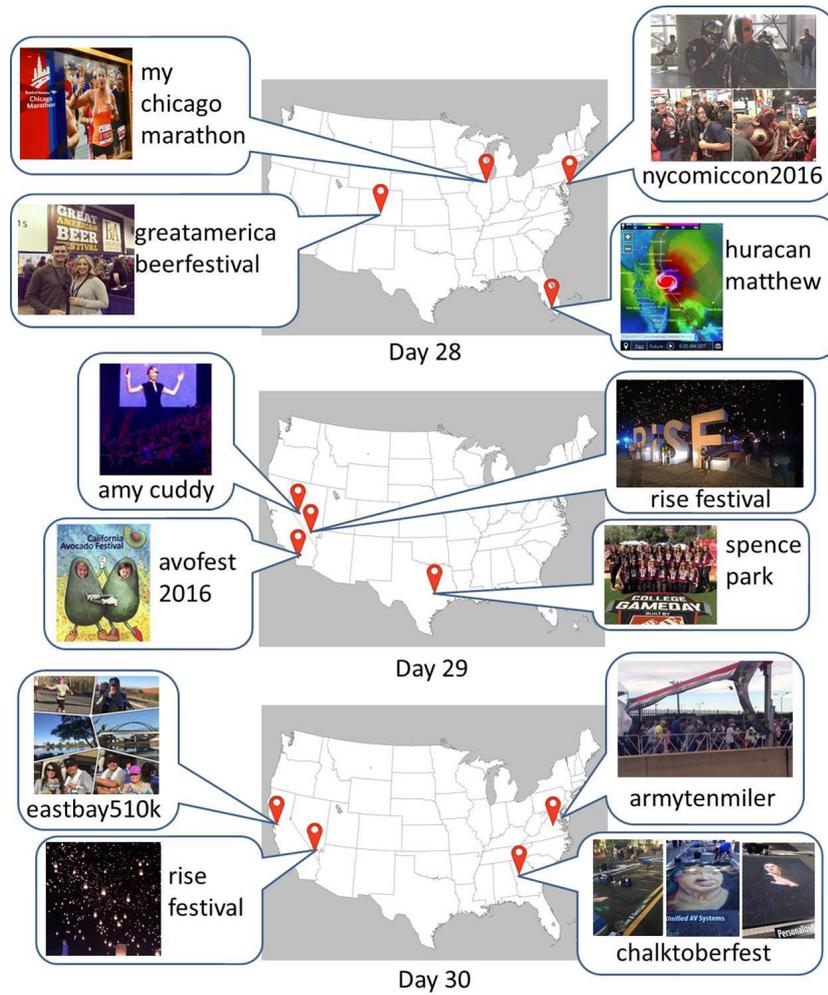


Figure 19. Examples of the temporary local words only in our geographic dictionary on different days.

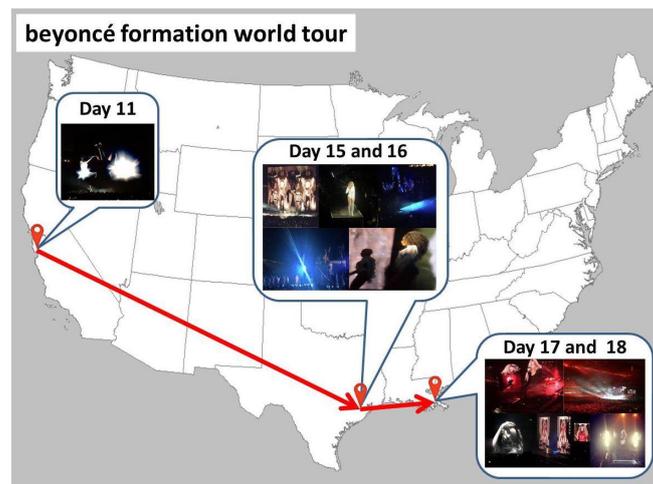


Figure 20. The locations for the temporary local word *beyoncé formation world tour* were correctly updated in our geographic dictionary.

5. Conclusions

This paper proposed a method for iteratively collecting geographic information from streaming geotagged tweets. By separately recording and examining the usage history of each word, the proposed method can determine suitable time window for each word to examine its spatial locality, so that various types of local words, ranging from well-known places to more minor local places, local products, and events, are extracted, while old local words are forgotten. Further, the geotagged tweets from bot accounts are filtered out in order to obtain a more accurate spatial distribution for each word. The experiments with over 6 million geotagged tweets posted from the United States during one month show that the geographic information related to over 61,000 local words were collected after a month. The diversity and accuracy of the collected local words were verified by comparing with existing geographic dictionaries constructed by experts, crowdsourcing, and automatically from geotagged tweets by focusing on a specific time interval. The collected information can also be visualized so that what is where in the real world can be easily observed. Although only a part of tweets are geotagged, non-geotagged tweets can further be used by inferring their locations based on the extracted local words in order to more rapidly increase the number of local words. Further, in our constructed geographic dictionary, some local words can be related, for example, different local words can represent same facilities, events, etc. Analyzing the relations of the local words by using the collected geographic information in order to further organize the dictionary would also be our future work.

Author Contributions: Conceptualization, J.L. and N.N.; Methodology, J.L. and N.N.; Software, J.L. and N.N.; Validation, J.L., N.N., K.N. and N.B.; Formal analysis, J.L. and N.N.; Investigation, J.L. and N.N.; Data curation, J.L.; Writing—original draft, J.L.; Writing—review and editing, N.N., K.N., and N.B.; Visualization, J.L. and N.N.

Funding: This research was funded by Japan Society for the Promotion of Science KAKENHI: Grant Number 26330137 and 16H06302.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Census U.S. Gazetteer. Available online: <https://www.census.gov/programs-surveys/geography/geographies/reference-files/gazetteer.html> (accessed on 18 April 2019).
2. GeoNames. Available online: <http://www.geonames.org> (accessed on 18 April 2019).
3. OpenStreetMap. Available online: <https://www.openstreetmap.org> (accessed on 18 April 2019).
4. Al-Olimat, H.S.; Thirunarayan, K.; Shalin, V.L.; Sheth, A. Location Name Extraction from Targeted Text Streams using Gazetteer-based Statistical Language Models. In Proceedings of the International Conference on Computational Linguistics, New Mexico, NM, USA, 20–26 August 2018; pp. 1986–1997.
5. Middleton, S.E.; Kordopatis-Zilos, G.; Papadopoulos, S.; Kompatsiaris, Y. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Trans. Inf. Syst.* **2018**, *36*, 40:1–40:27. [CrossRef]
6. Gritta, M.; Pilehvar, M.T.; Limsopatham, N.; Collier, N. What’s Missing in Geographical Parsing? *Lang. Resour. Eval.* **2018**, *52*, 602–623. [CrossRef]
7. Flickr. Available online: <https://www.flickr.com/> (accessed on 15 April 2019).
8. Twitter. Available online: <https://twitter.com/> (accessed on 15 April 2019).
9. Ahern, S.; Naaman, M.; Nair, R.; Yang, J. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. In Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries, Vancouver, BC, Canada, 18–23 June 2007; pp. 1–10.
10. Crandall, D.; Backstrom, L.; Huttenlocher, D.; Kleinberg, J. Mapping the World’s Photos. In Proceedings of the ACM International Conference on World Wide Web, Madrid, Spain, 20–24 April 2009; pp. 761–770.
11. Skovsgaard, A.; Sidlauskas, D.; Jensen, C.S. A Clustering Approach to the Discovery of Points of Interest from Geo-Tagged Microblog Posts. In Proceedings of the IEEE International Conference on Mobile Data Management, Brisbane, Australia, 14–18 July 2014; pp. 178–188.

12. Al-Ghossein, M.; Abdessalem, T. SoMap: Dynamic Clustering and Ranking of Geotagged Posts. In Proceedings of the International Conference on World Wide Web, Montréal, QC, Canada, 11–15 April 2016; pp. 151–154.
13. Vu, D.D.; To, H.; Shin, W.Y.; Shahabi, C. GeoSocialBound: An Efficient Framework for Estimating Social POI Boundaries using Spatio-Textual Information. In Proceedings of the International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, San Francisco, CA, USA, 26 June–1 July 2016; pp. 1–6.
14. Gao, S.; Janowicz, K.; Couclelis, H. Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-based Social Networks. *Trans. GIS* **2017**, *21*, 446–467. [[CrossRef](#)]
15. Hu, Y.; Gao, S.; Janowicz, K.; Yu, B.; Li, W.; Prasad, S. Extracting and Understanding Urban Areas of Interest using Geotagged Photos. *Comput. Environ. Urban Syst.* **2015**, *54*, 240–254. [[CrossRef](#)]
16. Spyrou, E.; Korakakis, M.; Charalampidis, V.; Psallas, A.; Mylonas, P. A Geo-Clustering Approach for the Detection of Areas-of-Interest and Their Underlying Semantics. *Algorithms* **2017**, *10*, 35. [[CrossRef](#)]
17. Kuo C.-L.; Chan, T.-C.; Fan, I.-C.; Zipf, A. Efficient Method for POI/ROI Discovery Using Flickr Geotagged Photos. *Int. J. -Geo-Inf.* **2018**, *7*, 121. [[CrossRef](#)]
18. Zheng, X.; Han, J.; Sun, A. A Survey of Location Prediction on Twitter. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1652–1671. [[CrossRef](#)]
19. Laere, O.V.; Quinn, J.; Schockaert, S.; Dhoedt, B. Spatially Aware Term Selection for Geotagging. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 221–234. [[CrossRef](#)]
20. Bo, H.; Cook, P.; Baldwin, T. Text-based Twitter User Geolocation Prediction. *J. Artif. Intell. Res.* **2014**, *49*, 451–500.
21. Bo, H.; Cook, P.; Baldwin, T. Geolocation Prediction in Social Media Data by Finding Location Indicative Words. In Proceedings of the International Conference on Computational Linguistics, Mumbai, India, 8–15 December 2012; pp. 1045–1062.
22. Chang, H.-W.; Lee, D.; Eltaher, M.; Lee, J. @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, 26–29 August 2012; pp. 111–118.
23. Rattenbury, T.; Naaman, M. Methods for Extracting Place Semantics from Flickr Tags. *Trans. Web* **2009**, *3*, 1–30. [[CrossRef](#)]
24. Intagorn, S.; Lerman, K. A Probabilistic Approach to Mining Geospatial Knowledge from Social Annotations. *Sigspat. Spec.* **2012**, *4*, 2–7. [[CrossRef](#)]
25. Hecht, B.; Hong, L.; Suh, B.; Chi, E.H. Tweets from Justin Bieber’s Heart: The Dynamics of the “Location” Field in User Profiles. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; pp. 237–246.
26. Ryoo, K.; Moon, S. Inferring Twitter User Locations With 10 km Accuracy. In Proceedings of the ACM International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; pp. 643–648.
27. Cheng, Z.; Caverlee, J.; Lee, K. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceedings of the ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, 26–30 October 2010; pp. 759–768.
28. Watanabe, K.; Ochi, M.; Okabe, M.; Onai, R. Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In Proceedings of the ACM International Conference on Information and Knowledge Management, Glasgow, UK, 24–28 October 2011; pp. 2541–2544.
29. Chen, L.; Roy, A. Event Detection from Flickr Data Through Wavelet-based Spatial Analysis. In Proceedings of the ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 523–532.
30. Boettcher, A.; Lee, D. EventRadar: A Real-Time Local Event Detection Scheme Using Twitter Stream. In Proceedings of the IEEE International Conference on Green Computing and Communications, Besancon, France, 11–14 September 2012; pp. 358–367.
31. Feng, W.; Han, J.; Wang, J.; Aggarwal, C.; Huang, J. STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream. In Proceedings of the IEEE International Conference on Data Engineering, Seoul, Korea, 13–16 April 2015; pp. 1561–1572.
32. Abdelhaq, H.; Sengstock, C.; Gertz, M. EvenTweet: Online Localized Event Detection from Twitter. In Proceedings of the International Conference on Very Large Data Bases, Trento, Italy, 26–30 August 2013; pp. 1326–1329.

33. Zhang, C.; Zhou, G.; Yuan, Q.; Zhuang, H.; Zheng, Y. GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 513–522.
34. Zhang, C.; Lei, D.; Yuan, Q.; Zhuang, H.; Kaplan, L.; Wang, S.; Han, J. GeoBurst+: Effective and Real-Time Local Event Detection in Geo-Tagged Tweet Streams. *ACM Trans. Intell. Syst. Technol.* **2018**, *9*, 34:1–34:24. [[CrossRef](#)]
35. Zhang, S.; Cheng, Y.; Ke, D. Event-Radar: Real-time Local Event Detection System for Geo-Tagged Tweet Streams. *arXiv* **2017**, arXiv:1708.05878.
36. Yamaguchi, Y.; Amagasa, T.; Kitagawa, H.; Ikawa, Y. Online User Location Inference Exploiting Spatiotemporal Correlations in Social Streams. In Proceedings of the ACM International Conference on Information and Knowledge Management, Shanghai, China, 3–7 November 2014; pp. 1139–1148.
37. Kamimura, T.; Nitta, N.; Nakamura, K.; Babaguchi, N. On-line Geospatial Term Extraction from Streaming Geotagged Tweets. In Proceedings of the IEEE International Conference on Multimedia Big Data, Laguna Hills, CA, USA, 19–21 April 2017; pp. 322–329.
38. Li, C.; Sun, A. Fine-Grained Location Extraction from Tweets with Temporal Awareness. In Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Australia, 6–11 July 2014; pp. 43–52.
39. Ajao, O.; Hong, J.; Liu, W. A Survey of Location Inference Techniques on Twitter. *J. Inf. Sci.* **2015**, *41*, 855–864. [[CrossRef](#)]
40. Chi, L.; Lim, K.H.; Alam, N.; Butler, C.J. Geolocation Prediction in Twitter Using Location Indicative Words and Textual Features. In Proceedings of the Workshop on Noisy User-Generated Text, Osaka, Japan, 11 December 2016; pp. 227–234.
41. Ozdakis, O.; Ogüztüzün, H.; Karagoz, P. A Survey on Location Estimation Techniques for Events Detected in Twitter. *Knowl. Inf. Syst.* **2017**, *52*, 291–339. [[CrossRef](#)]
42. Ozdakis, O.; Ramampiaro, H.; Nøvåg, K. Spatial Statistics of Term Co-occurrences for Location Prediction of Tweets. In Proceedings of the European Conference on Information Retrieval, Grenoble, France, 26–29 March 2018; pp. 494–506.
43. Eric, B. A Simple Rule-based Part of Speech Tagger. In Proceedings of the Conference on Applied Natural Language Processing, Trento, Italy, 31 March–3 April 1992; pp. 152–155.
44. TermExtract. Available online: <http://gensen.dl.itc.u-tokyo.ac.jp/pytermextract/> (accessed on 18 April 2019).
45. Achananuparp, P.; Hu, X.; Shen, X. The Evaluation of Sentence Similarity Measures. In Proceedings of the International Conference on Data Warehousing and Knowledge Discovery, Turin, Italy, 2–5 September 2008; pp. 305–316.
46. Stopwords ISO. Available online: <https://github.com/stopwords-iso/stopwords-iso> (accessed on 18 April 2019).
47. Backstrom, L.; Kleinberg, J.; Kumar, R.; Novak, N. Spatial Variation in Search Engine Queries. In Proceedings of the International Conference on World Wide Web, Beijing, China, 21–25 April 2008; pp. 357–366.
48. D3.js. Available online: <https://d3js.org/> (accessed on 18 April 2019).
49. Embedded Tweets. Available online: <https://developer.twitter.com/en/docs/twitter-for-websites/embedded-tweets/overview> (accessed on 18 April 2019).

