A Survey of Location Prediction on Twitter

Xin Zheng, Jialong Han, and Aixin Sun

Abstract—Locations, e.g., countries, states, cities, and point-of-interests, are central to news, emergency events, and people's daily lives. Automatic identification of locations associated with or mentioned in documents has been explored for decades. As one of the most popular online social network platforms, Twitter has attracted a large number of users who send millions of tweets on daily basis. Due to the world-wide coverage of its users and real-time freshness of tweets, location prediction on Twitter has gained significant attention in recent years. Research efforts are spent on dealing with new challenges and opportunities brought by the noisy, short, and context-rich nature of tweets. In this survey, we aim at offering an overall picture of location prediction on Twitter. Specifically, we concentrate on the prediction of user home locations, tweet locations, and mentioned locations. We first define the three tasks and review the evaluation metrics. By summarizing Twitter network, tweet content, and tweet context as potential inputs, we then structurally highlight how the problems depend on these inputs. Each dependency is illustrated by a comprehensive review of the corresponding strategies adopted in state-of-the-art approaches. In addition, we also briefly review two related problems, i.e., semantic location prediction and point-of-interest recommendation. Finally, we make a conclusion of the survey and list future research directions.

Index Terms—Twitter, tweets, home location, tweet location, mentioned location, location prediction

1 Introduction

The last decade has witnessed an unprecedented proliferation of online social networks. Those include general-purpose platforms like Twitter and Facebook, location-based ones like Foursquare and Gowalla, photo-sharing sites like Flickr and Pinterest, as well as other domain-specific platforms such as Yelp and LinkedIn. On these platforms, users may establish online friendship with others sharing similar interests. Users may also share with online friends their daily lives in forms of texts, photos, videos, or check-ins.

Among all online social networks, Twitter is characterized by its unique way of following friends and sending posts. On the one hand, Twitter friendships are not necessarily mutual. For example, users may "follow" celebrities without requiring them to follow back. On the other hand, textual posts on Twitter, *a.k.a.* tweets or microblogs, are limited to 140 characters. Users are encouraged to post frequently but casually about anything, such as moods, activities, opinions, local news, etc.

Users, online friendships, and tweets make Twitter a virtual online world. This virtual world intersects with the real world, where locations acting as intermediate connections. Twitter users have long-term residential addresses. Their home locations cause them to notice, get interested, and tweet news or events around their daily activity regions. With increasing popularity of GPS-enabled devices such as smartphones and tablets, users may casually attach real-time locations when sending out tweets. Users may also mention locations in their tweets, e.g., cities they previously lived in, or restaurants they want to try. In this survey, we concentrate on the above three types of Twitter-related locations, namely *user home location*, *tweet location*, and *mentioned location*. Knowing physical locations involved in Twitter helps us to understand what is happening in real life, to bridge the online and offline worlds, and to develop applications to support real-life demands, among many applications. For example, we can monitor public health of residents [1], recommend local events [2] or attractive places [3] to tourists, summarize regional topics [4], and identify locations of emergency [5] or even disasters [6].

Although Twitter users may casually reveal locations either manually or with the help of GPS, location information on Twitter are far from complete and accurate. Cheng et al. [7] find that only 21 percent of users in a U.S. Twitter dataset provide residential cities in their profiles, while 5 percent give coordinates of their home addresses. Despite the low availability, Hecht et al. [8] report that self-declared home information in many user profiles are inaccurate or even invalid. Hecht et al. [8] and Ryoo et al. [9] observe that only 0.77 and 0.4 percent of tweets have location information attached in their datasets, respectively. Similar percentages are also reported by Bartosz et al. [10] and Priedhorsky et al. [11]. Therefore, completing Twitter-related locations acts as the prerequisite for many other studies and applications, and is worth careful investigation.

The problem of predicting locations associated with objects has been termed as geolocation or geocoding, and

- *X. Zheng is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, and SAP Research and Innovation Singapore, SAP Asia Pte Ltd, Singapore 119968. E-mail: xzheng008@e.ntu.edu.sg.*
- *J. Han is with the Tencent AI Lab, Shenzhen, Guangdong, China. E-mail: jialonghan@gmail.com.*
- *A. Sun is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798. E-mail: axsun@ntu.edu.sg.*

Manuscript received 4 July 2017; revised 23 Nov. 2017; accepted 6 Feb. 2018. Date of publication 20 Feb. 2018; date of current version 3 Aug. 2018. (Corresponding author: Xin Zheng.)
Recommended for acceptance by Y. Chang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2807840

studied for Wikipedia [12], [13], [14], web pages [15], [16], and general documents [17]. The recognition and disambiguation of mentioned entities[1] in formal documents, or entity recognition [18] and linking [19], are also extensively investigated for decades. Various text processing techniques have been proposed to address these problems. Intuitively, recognition and disambiguation of Twitter-related locations should also depend heavily on tweet texts. Users living in certain cities may discuss local landmarks, buildings and events, possibly with dialects or slang. Tweets sent out from certain locations may explicitly mention them in the text, or implicitly include some relevant words. However, the characteristics of Twitter pose emerging challenges for these existing research problems in new problem settings. On the one hand, users often write tweets in a very casual manner. Acronyms, misspellings, and special tokens make tweets noisy, and techniques developed for formal documents are error-prone on tweets. The limit of 140-character also makes tweets short, which may not be easily understood by readers who are unaware of tweets' context. On the other hand, compared with formal documents, Twitter users contribute their online friendships and profiles explicitly. They also intentionally or unintentionally attach geo-tags to tweets. The richness of contextual information on Twitter enables new opportunities to relieve aforementioned challenges.

Given the above significance, necessity, challenges, and opportunities, Twitter-related location prediction problems have received much attention in the literature, and even been proposed as one of the shared tasks in the 2nd Workshop on Noisy User-generated Text (W-NUT).[2] To the best of our knowledge, no previous survey focuses extensively on exactly the same scope. Imran et al. [20] have done a comprehensive study on tracking and analyzing mass emergency with social media data. Their focus is multifaceted, which not only involves locations but also has temporal and event aspects. Melo et al. [21] review various techniques for geolocating ordinary documents, but the unique challenges and opportunities of Twitter are not touched. Ajao et al. [22] conduct a smaller scale survey which addresses the most similar scope as we are aware of. However, they only clarify possible input and output of location prediction problems on Twitter. Detailed techniques are discussed with minimal efforts. Nadeau et al. [18] and Shen et al. [19] concentrate on named entity recognition and linking, respectively. They are related to one of the three problems in this survey, i.e., mentioned location prediction. Besides, their focuses are on general entities and documents, while we specially target the intersection of the location domain and Twitter platform.

In this survey, we aim at completing an overall picture of location prediction problems on Twitter. In Section 2, we brief the input, output, and evaluation metrics of Twitter-based location prediction. In Sections 3, 4 and 5, we detail previous efforts on each problem. By highlighting the role of each input, we systematically summarize essentials of previous works on each prediction problem. In Section 6, we brief two additional location-related problems. Though attracting less attention or not as relevant, these two problems complement the three major problems and the scope of this survey. Finally, we conclude the survey and discuss future research directions.

1. A named entity is a real-world object; examples are persons, organizations, or locations.

2. The workshop also provides an evaluation dataset which we call W-NUT(http://noisy-text.github.io/2016/index.html).

2 Problem Overview

This survey focuses on location prediction problems on Twitter. In this section, first, we give an overview of the Twitter platform. By introducing Twitter usage from an ordinary user's point of view, we summarize Twitter dataset from three perspectives i.e., content, network, and context. Next, we discuss three geolocation problems of general interest. Those prediction problems rely on the above information as major input. Finally, we briefly review evaluation metrics for the aforementioned prediction problems.

2.1 An Overview of Twitter

As one of the most popular online social network, Twitter constantly accumulates large volume of heterogeneous data at a high velocity. Those include 1) short and noisy tweets posted by users, 2) a massive Twitter network established among users, and 3) rich types of contextual information for both users and tweets. Such information serves as input and enables the study of a few geolocation problems. In this section, we briefly describe the three types of information.

2.1.1 Tweet Content

A *tweet* is a piece of user-generated text with its length up to 140 characters. It may describe anything a user wants to post, e.g., her mood or events happening around her. Besides original posts, a user may also *retweet* others' tweets she reads. Tweets and retweets from a user will be pushed to her *followers'* (see definition in Section 2.1.2) Twitter interface for them to read. When composing tweet contents, a user may include *hashtags*, which are words or unspaced phrases starting with "#". Finally, one can also *mention* another user's name by a preceding "@" in tweet content. A mentioned user will be notified, and may start a *conversation* with the mentioning user through subsequent mentions.

2.1.2 Twitter Network

Besides posting tweets, a user may subscribe others' tweets by *following* them. If user u_i follows u_j, we call u_i the *follower*, and u_j the *followee*. Note that following relationships are unidirectional, i.e., u_i following u_j does not necessarily mean u_j following u_i. When the direction of a following relationship is not the major concern, we regard u_i and u_j as *friends*. If it happens that u_i and u_j follow each other, we say u_i and u_j are *mutual friends*. We refer to all 'following' relationships as *Twitter friendship*, or *friendship* when the context is clear.

Note that Twitter friendship does not imply friendship in real life. It is often the fact that celebrities do not follow back most of their ordinary followers. Moreover, even two distant strangers may become mutual friends by chance. However, it is observed that friends in real life tend to mention each other frequently online [23], [24], [25], [26]. When introducing the studies on clues that imply real-life friendship, we consider both following and mentioning actions between Twitter users in a uniform manner, and refer to the resulted network as *Twitter network*.

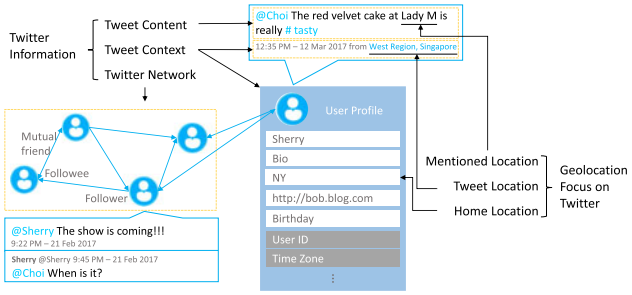


Fig. 1. An illustration of tweet content, tweet context, and Twitter network, and the three types of locations: home location, tweet location, and mentioned location in Twitter.

2.1.3 Tweet Context

A tweet is more than a piece of short text. When a tweet is sent out, it is attached with its posting *timestamp*. Moreover, with the prevalence of GPS-enabled devices like smartphones and tablets, users may optionally publish their current locations as *geo-tags*[3] on tweets. Finally, users may complete their profiles to include information like home cities, timezones, and personal websites. We note that all above information provide context helping us better understand tweets. A user's daily-life tweets can be interpreted more precisely, if all such information are available. Because timestamps, geo-tags, and user profiles serve as contextual information for tweets, we refer to them as *tweet context*.

2.2 Location Prediction Problems on Twitter

In this survey, we focus on predicting three types of Twitter-related locations, namely *home location*, *tweet location*, and *mentioned location*. For each type of location, we give its definition and show how it is represented. We also briefly discuss how to set up ground truth for each task.

2.2.1 Home Location Prediction

Home locations refer to Twitter users' long-term residential addresses. The prediction of home locations enables various applications, e.g., local content recommendation, location-based advertisement, public health monitoring, and public opinion polling estimation. According to specific requirements of applications, home locations may be represented at different levels of granularity. Generally, there are three categories of home location granularity:

- *Administrative regions*, i.e., countries, states, or cities where users stay.
- *Geographical grids*, i.e., the earth is partitioned into cells of equal or varying sizes,[4] and home locations are represented by the cells they fall in.
- *Geographical coordinates*, i.e., homes are represented by their latitudes and longitudes. Coordinates may be self-reported or converted from administrative regions or cells by taking their centers.

Ground truth home locations may be collected from users' self-declared profiles. For example, in Fig. 1, the user reports that she lives in NY (New York). Due to possible privacy concerns, empty and noisy information appears in user profiles. Some studies also aggregate geo-tags attached with users' tweets as their ground truth home locations. Possible aggregating approaches include:

- The most frequent city involved in the geo-tags.
- The first valid geotag, and convert it to an administrative region, a grid, or coordinates.
- The geometric median[5] of the geo-tags.

For the sake of evaluation, a uniform level of granularity should be decided and fixed for an application. However, to achieve maximum coverage of ground truth, user profiles and geo-tag aggregations could be utilized in combination.

2.2.2 Tweet Location Prediction

Tweet location means the place where a tweet is posted. By inferring tweet locations, we may draw a more complete picture of a user's mobility. Different from home locations, which are collected from both user profiles and geo-tags, tweet locations are generally based on geo-tags of tweets. Due to the original views of tweet locations, point-of-interests (POIs in short) or coordinates are broadly adopted as representations of tweet locations, instead of administrative regions or grids.

2.2.3 Mentioned Location Prediction

When writing tweets, users may mention the names of some locations in tweet contents. Mentioned location prediction may facilitate better understanding of tweet contents, and benefit applications like location recommendation and disaster & disease management. In this survey, we involve two sub-tasks of mentioned location prediction:

- *Mentioned location recognition*, i.e., extract text fragments in a tweet that refer to location names.
- *Mentioned location disambiguation*, i.e., identify what locations those fragments refer to by resolving them to entries in a location database.

Due to the inherent noise and ambiguity of tweet language, ground truth of mentioned locations largely rely on human annotations. To represent location mentions in tweets, BIO or BILOU[6] labeling schemes are widely adopted. For both sub-tasks, the granularity of locations involve both administrative regions and POIs. When a pre-defined location database is employed, the granularity generally respects that of the database.

2.3 Twitter Inputs for Location Prediction Problems

All the three types of information on Twitter, i.e., content, network, and context, are commonly adopted to solve the three location prediction problems, i.e., the prediction of home location, tweet location and mentioned location. This is because multiple data source could help to enrich

3. Geo-tags may be in the form of point-of-interests (e.g., a hotel or a shopping mall) or simply geographical coordinates (latitudes and longitudes).

4. Equal-sized cells are achieved by uniformly binning latitudes and longitudes [12]. The major drawback is that rural areas are over-represented at the expense of urban areas. Therefore, quad-tree [27] or k-dimensional tree (k-d tree) [13], [14], [28] are adopted to achieve varying-sized cells with better resolutions on populated areas.

5. The geometric median of a point set S is the point in S which has minimal average distance to the other points.

6. BIO stands for the **B**eginning, **I**nside, and **O**utside of a location mention in a sentence. BILOU additionally annotates the **L**ast word of a multi-word mention, as well as all **U**nit-length mentions.

the available information, so that to relieve data sparsity issue on Twitter. However, for different geolocation problems, the ways to utilize the input data are different. We will discuss the differences at the end of each section.

2.4 Evaluation Metrics

In this section, we review common evaluation metrics adopted in the literature. Depending on the representations of predicted and ground-truth locations that are fed to the evaluation stage, common metrics could be categorized as *distance-based* or *token-based*. In the distance-based point of view, locations are represented by their geographical coordinates. Token-based metrics treat locations as discrete symbols, e.g., country, city, grid, POI. Next, we formulate both of them and demonstrate their usage scenarios.

2.4.1 Distance-Based Metrics

In home location or tweet location prediction, we aim at making predictions for each user or tweet. For unified notations, let s be a user or tweet, and S be the set of all users or tweets for prediction. A system is expected to predict a location $l(s)$ for each s. The prediction $l(s)$ is expected to coincide with or be close to the ground truth location $l^*(s)$. Whatever granularity we adopt, all ground-truth and predicted locations could be converted to coordinates. *Error Distance* (*ED* for short) is then defined as the Euclidean distance between ground-truth and predicted coordinates:

$$\mathrm{ED}(s) = \mathrm{dist}(l(s), l^*(s)).$$

Since evaluations are conducted on a collection of users or tweets, we may take the mean or median of all error distances to end up with corpus-level metrics. This results in i.e., *Mean Error Distance* and *Median Error Distance*:

$$\mathrm{MeanED} = \frac{1}{|S|}\sum_{s\in S}\mathrm{dist}(l(s), l^*(s)),$$
$$\mathrm{MedianED} = \underset{s\in S}{\mathrm{median}}\{\mathrm{dist}(l(s), l^*(s))\}.$$

When wildly inaccurate predictions occur, Median Error Distance is usually less sensitive than Mean Error Distance. Therefore, Mean Error Distance is preferred by some studies. Instead of Mean Error Distance, some studies [29], though very few, employ *Mean Squared Error* as below:

$$\mathrm{MSE} = \frac{1}{|S|}\sum_{s\in S}\mathrm{dist}^2(l(s), l^*(s)).$$

The only difference between Mean Squared Error and Mean Error Distance is the former takes square of Error Distance.

Besides Mean and Median Error Distance, there is another widely-adopted corpus-level metric called *Distance-based Accuracy*, or *Acc@d* for short. Given a predefined threshold d of error distance, any prediction whose error distance does not exceed d is regarded as "tolerably correct". The Acc@d metric over the corpus is then defined as the proportion of tolerably correct predictions:

$$\mathrm{Acc@d} = \frac{|\{s \in S : \mathrm{ED}(s) \le d\}|}{|S|}. \quad (1)$$

The commonly adopted distance threshold d is 100 miles, or 161 km [30], [31].

2.4.2 Token-Based Metrics

Alternatively, token-based metrics treat locations as discrete symbols, e.g., country, city, grid, POI. Though geographical information is not taken into consideration, token-based metrics allow for more general usage scenarios.

For the three geolocation problems, the simplest token-based metric is Accuracy. Let $l(s)$ and $l^*(s)$ be the predicted and ground-truth locations for a user, a tweet, or a recognized location mention s. Note that their administrative-region or POI representations are kept. A prediction is deemed correct only if it coincides with the ground-truth. *Accuracy* is then defined as the ratio of correct predictions within S:

$$\mathrm{Acc} = \frac{|\{s \in S : l(s) = l^*(s)\}|}{|S|}.$$

In some cases, a system may give a ranking list $L(s)$ of predicted locations instead of one. A straightforward approach is to treat the top location as the only prediction and resort to Accuracy. However, this approach ignores other predictions in the list, which may also be useful when fed to downstream applications or users. In light of this, *Ranking-based Accuracy*, or *Acc@k* is designed. A ranking list is considered "correct" if the ground-truth location lies within the top-k results $L_k(s)$. Acc@k is then defined as the proportion of "correct" lists:

$$\mathrm{Acc@k} = \frac{|\{s \in S : l^*(s) \in L_k(s)\}|}{|S|}.$$

Finally, we note that the geolocation systems may not be able to make predictions in some cases. For example, in home and tweet location predictions, some systems cannot assign locations if insufficient information is given [25], [29], [32]. In mentioned location disambiguation, systems may not find appropriate entry for a given location mention. In such cases, Precision, Recall and F_1 are adopted as metrics. Given a user, a tweet, or a recognized location mention s, let $l(s) = \mathrm{null}$ if the system cannot make any prediction. The *Precision* over the evaluation corpus S is defined as the ratio of correct predictions among all predictions:

$$\mathrm{Precision} = \frac{|\{s \in S : l(s) = l^*(s)\}|}{|\{s \in S : l(s) \neq \mathrm{null}\}|}.$$

Meanwhile, *Recall* is defined similarly as Accuracy, i.e.,

$$\mathrm{Recall} = \frac{|\{s \in S : l(s) = l^*(s)\}|}{|S|}.$$

After Precision and Recall are defined, F_1 is the harmonic mean of Precision and Recall:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

Finally, we note that Precision, Recall and F_1 are applicable and are actually widely adopted for mentioned location recognition. When evaluating location recognition results,

mentioned fragments should be regarded as "tokens". A predicted fragment is deemed correct if its left and right boundaries coincide with those of a ground-truth fragment, respectively. Precision is then defined as the ratio of correctly predicted fragments over all predicted fragments. Recall is the proportion of correctly predicted fragments among all ground-truth fragments. Accordingly, their harmonic mean is defined as F_1.

3 Home Location Prediction

Knowing home locations of Twitter users enables many applications, such as local content recommendation, location-based advertisement, public health monitoring, public opinion polling, etc. However, because it is optional for Twitter users to complete their profiles, Twitter users' home locations are mostly absent or noisy. Therefore, many research efforts have been spent on predicting users' home locations. In most studies, home locations are predicted at city-level, and sometimes at state or country level. In this section, we detail them based on different inputs, namely tweet content, Twitter network, and tweet context. Note that many studies simultaneously involve multiple inputs, especially the first two. In this case, they will be mentioned multiple times, where assumptions and techniques regarding different inputs are discussed in the corresponding sections.

3.1 Inference Based on Tweet Content

Users' home locations could be casually revealed by certain words in tweet content. For example, people in Houston would talk about Houston Rockets more than users in New York. Residents from Texas usually use dialect "howdy" and those from Philadelphia often call themselves "phillies". Thus, the underlying challenge for content-based home location prediction is to precisely link users to locations via those indicative words.

Previous studies on content-based home location prediction could be divided in two classes: word-centric and location-centric. Word-centric method is to estimate the probability of a location l given words w in text, or $p(l|w)$; while location-centric method focuses on the probability of generating a tweet d at a given location $p(d|l)$. Next, we will detail the two kinds of studies respectively.

3.1.1 Word-Centric Methods

In the beginning of Section 3.1, we mentioned two examples about location-indicative words in users' tweets. Word-centric methods aim at identifying and exploiting such words to predict users' home locations. Not all words are location-indicated. For example, words like "downtown" and "OMG" are used everywhere on Twitter. Therefore, only *local words*, i.e., words that show strong locality, should be involved. Besides, the location information implied by local words, or their *spatial word usage*, should be learnt from data before making predictions. Next, we describe how both tasks are achieved in the literature.

Identifying Local Words. In information retrieval literature, a commonly adopted practice is to eliminate *stop words* like "a", "the", etc., from documents before indexing them for retrieval. As for tweets, it is often the case that location-irrelevant words like "downtown" and "OMG" appear more frequently than "howdy" and "phillies" like words. They will lead home location prediction results to random if indiscriminately taken into consideration. Unlike eliminating predefined list of stop words, we usually resort to eliminate location-irrelevant words, i.e., identify and keep local words. Since local words are not enumerable like stop-words in most applications, a large amount of research efforts are spent on identifying local words, either unsupervised or supervised.

Unsupervised local word identification methods aim at statistical measures that are directly computable on the data and are indicative of a word's locality. Laere et al. [33] propose two types of local word selection methods. One leverages Kernel Density Estimation [34] which spatially smooth term occurrences, and the other is based on Ripley's K statistic [35] which measures term's geographical deviation. Inspired by Inverse Document Frequency (IDF) in information retrieval, Ren et al. [36] and Han et al. [28] propose Inverse Location Frequency (ILF) and Inverse City Frequency (ICF), respectively, to measure the locality of words. Their assumption is that local words should be distributed in fewer locations and have larger ILF and ICF values. Besides IR-based measures, some studies also resort to measures that have information theoretic interpretations, e.g., information gain and maximum entropy in [28], and K-L divergence in [27]. Their assumption is that the distributions of local words should be more biased than ordinary ones. Noted that Yamaguchi et al. [27] deal with streaming tweets which could update users' home location according to newly posted tweets. In [8], Hecht et al. propose a CALGARI score for words, which is similar to information theory based measures. Mahmud et al. [37] apply a series of heuristic rules to select local words. Han et al. [38] report a comparison of statistical-based, information theory-based and heuristic-based methods on local words selection.

On the other hand, supervised methods are also considered in a number of studies. In [1], Cheng et al. view the problem of local word identification as a classification problem. First, they fit the geographical distribution of each word with spatial variation model by Backstrom et al. [39]. The spatial variation model assumes that each word has a geographical center, a center frequency C, and a dispersion ratio α. The probability of seeing the word at a location with distance d to the center is proportional to $Cd^{-\alpha}$. In simple words, this model specifies a one-peak distribution at the center with exponential decay. After the model is fit, the parameters are used as word features. Second, they manually labeled 19,178 words in a dictionary as either local or non-local. Finally, they train a classification model and apply it to all other words in the tweet dataset. Ryoo and Moon [9] apply the above method [39] to a Korean tweet dataset, and achieve satisfactory results.

Modeling Spatial Word Usage. After identifying local words, the next problem is how to use them to predict users' home locations. Most studies model this problem in a probabilistic manner. Researchers propose probabilistic models to characterize the conditional distribution of users' home locations w.r.t. their tweets contents, then decompose and concretize the model to make predictions.

A representative probabilistic model is introduced in Cheng et al. [1]. The distribution of user u's home location l given her tweet contents $S(u)$ is decomposed as

$$P(l|u) \propto \sum_{w \in S(u)} P(l|w)P(w).$$

Here only local word w are considered, and $P(w)$ denotes the probability of w over the entire corpus. After the decomposition, major efforts are spent on estimating the location distribution $P(l|w)$ of word w, or *spatial word usage*. It is reported that estimating $P(l|w)$ directly from the corpus is inferior. The reason is that some w may be unobserved in less populated locations, which does not mean that the location is irrelevant to w. To relieve this sparsity problem, smoothing techniques need to be involved. A special type of spatial words is location names in tweets. Li et al. [40] observe that the probability of tweeting venue names is location-based at some time, while it is also random at other time. Thus they make it a two level estimation. A Bernoulli distribution is adopted to estimate whether a location name is posted randomly or based on location, following which a multinomial distribution is used to estimate the probability of tweeting the venue name from each location.

In the same work [1], Cheng et al. propose several *explicit* smoothing methods. The first method, Laplace smoothing (or add-one smoothing), increase word w's count in all locations by one before normalizing it to produce a distribution. This method ensures that all locations get positive probabilities. However, it does not involve the geographical information in l. They further propose another two smoothing methods, namely, state-level smoothing and grid-based neighborhood smoothing. In those methods, a fixed portion of per-state or per-cell word counts are evenly distributed to only locations in the same state or cell, instead of all locations on the map. In [36], Ren et al. also consider an explicit smoothing technique called circular-based neighborhood smoothing. On the other hand, some parameterized spatial word usage models, once fitted, have *implicit* smoothing effects. In [1], Cheng et al. treat the fitted spatial variation model in [39] as a smoothed distribution. In an extension work [7], Cheng et al. generalize the one-peak model [39] by wave-like smoothing to allow multi-peaks for words distributions. In the influence-based social closeness models [30] (see Section 3.2.2), Li et al. treat friends followed and location names mentioned by users uniformly, and use Gaussian models to fit their geographical usage. Instead, Chang et al. [41] use Gaussian mixture models to fit spatial word usage. Their model also allows multi-peaks and is implicitly smoothed.

3.1.2 Location-Centric Methods

Word-centric methods characterize local words distributions and infer locations from them. Some other studies adopt different methods that give locations more centric roles.

A few studies adopt classification-based approaches to home location prediction. They treat users' statistics about local words as features, and all candidate locations as classification labels. Hecht et al. [8] select top 10,000 words with highest CALGARI scores as local words. Users are then represented as 10,000-dimensional term frequency vectors, and fed into a multinomial Naive Bayes classifier for training and home location prediction. Similarly, Rahimi et al.'s [42] apply logistic regression on users' TF-IDF vectors. Instead of selecting local words as features, they subject to a sparse l_1 regularization penalty [43]. Similarly, Cha et al. [44] use sparse coding and dictionary learning techniques for word feature selection. In [37], Mahmud et al. adopt a hierarchical ensemble algorithm to train two-level classifier ensembles on the granularity of timezone-city or state-city. In their extension work [45], they also propose identifying and removing travelling people from training data to improve the performance of home location classifiers. A person is considered travelling if any two of her tweets were sent from locations with distance above 100 miles. Wing and Baldridge [13] also resort to hierarchical classification [46]. Instead of adopting administrative partitions directly, they use k-d tree to achieve adaptive grids in their hierarchy. This leads to better granularity for populated regions, and avoids unnecessarily over-representing less populated areas.

There are also studies that adopt information-retrieval-based approaches to home location prediction. They treat locations as pseudo-documents that consist of all tweets whose users live here. Given the pseudo-document of a user whose home location is to be predicted, locations with the most similar pseudo-documents are retrieved as prediction results. Specifically, Wing et al. [12] adopt a grid representation of locations. They estimate a language model [47] for each grid with its pseudo-document. Good-Turing smoothing [48] is applied to smooth the probability of unseen words. Kullback-Leibler divergence is adopted as the similarity measure between location documents and user documents. In their subsequent work [13], they resort to adaptive grids as in [14]. When geo-coordinates need to be reported instead of grids, they find that reporting the centroid of user locations in the grid yields better precision than reporting mid-points of the grid.

Besides traditional methods, some recent works also explore deep learning models to tackle home location prediction. By extending their previous work [49], Miura et al. [50] propose a more sophisticated model. They order a user's messages chronologically and apply sequential model RNN to encode the content. In virtue of attention mechanism, a global message representation which addresses important information could be obtained. Similar process is also applied on context, i.e., location description and timezone. The combination of the three representations is then fed to a softmax layer to predict home location. Rahimi et al. [51] apply a multilayer perceptron (MLP) with one hidden layer to classify users' home locations. They adopt l_2 normalized bag-of-words representation of a given user's tweet contents as input. The output is a predefined discretized region generated by either a k-d tree or k-means.

3.2 Inference Based on Twitter Network

Besides posting tweets, other major activities that users involve in on Twitter are to establish following relationship and interact with friends. Like their tweet contents, users' social relationships may reveal their home locations as well. In Section 3.2.1, we review some friendship-based methods, where friends are assumed to have smaller home location distances. Moreover, it is also argued in studies that social-closeness, which is based on friendship, interactions, and other implicit signals, are more reliable for estimating home distances than sole friendship. These studies are reviewed in Section 3.2.2. Finally, when multiple users' home

locations are unknown and to be predicted, their home locations are not independent because they are directly or indirectly interlinked through the Twitter network. This dependency cannot be captured by *local* inference methods that predict one home location at a time. In Section 3.2.3, we demonstrate how *global* inference methods are applied in some studies.

3.2.1 Friendship-Based Methods

In social science, the assumption of *homophily* [52] suggests that similar people make contacts at a higher rate than dissimilar ones. Given the task of predicting home locations based on Twitter network, a quick intuition may be that one's home location is very likely to coincide with her friends' home locations. In the preliminary model of [36], Ren et al. assume that the higher proportion of a user's friends live at a location, the higher probability for the user to stay at the same location. Davis et al. [53] employ a similar approach to that of [36], except that they only consider mutual friendship. Rodrigues et al. [54] model home location prediction with the Potts model [55], which aims to maximize global home co-location between mutual friends. One drawback of the above three approaches is that they do not use the coordinates of home locations of a user's friends. Locations are treated as a discrete set of objects, while the distance between them is ignored.

One of the earliest attempts to model friendship and home location distance is made by Backstrom et al. [56]. Although this study is conducted on Facebook, we include it in this survey because of its impacts on later Twitter-based studies. The authors analyze a large number of Facebook users with known home locations and their friendships. They try to fit the probability of two users being friends w.r.t. their home distance with the following curve

$$P(u_i, u_j \text{ are friends } | dist(u_i, u_j) = x) = a(b + x)^{-c}, \quad (2)$$

and find that $c = 1$ produces a good fit. In other words, the probability of friendship is inversely proportional to home distance (with intercept b). Based on this model, given friends of a user and their home locations, the most probable home location for the user could be found, by maximizing the probability of generating all seen friendship links.

The aforementioned three methods all depend user home proximity solely on *direct* friendship. In other words, they implicitly assume that friendship observed on an online social network implies real off-line friendship, and thus close home distance. This may be far from true. In [57], Kong et al. find that a pair of friends has 83 percent of chance to live within 10 kilometers if their common friends account for more than half of their friends, respectively. The chance decreases to 2.4 percent if the common friend ratio is limited to 10 percent. This implies that rich *indirect* friendships on Twitter may better indicate off-line friendship between two users, and thus their home location proximity. As is also observed by Kossinets et al. [58], if two users a and b have relationship with many third users, a and b may possibly have a relationship. Inspired by this, Kong et al. improve the model in [56] by considering cosine similarity between two users' friend collections in Eq. (2). Rout et al. [59] also relate the probability a user lives in a city to the distribution of indirect friendships between the user and her friends at the location. Miura et al. [50] encode user friendship information into a neural network model. Different from the other works, they separate users in connected network and their corresponding cities, and assign them user embeddings and city embeddings respectively. An attention mechanism is applied on the addition of user and city embeddings to draw useful information on home location prediction.

3.2.2 Social-Closeness-Based Methods

In the previous section, we discussed several friendship-only methods, which only involve friendships available in the Twitter network. However, it may harm home prediction if we depend home distance purely on direct friendship on Twitter. Studies report that the inverse proportion model [56] in Eq. (2) on Facebook does not hold for Twitter. For example, McGee et al. [23] observe that friendship probability w.r.t home distance on Twitter roughly satisfy a bimodal distribution. One peak is around 10 miles, and the other is far away. Similar observations are also made by Scellato et al. [68] and Volkovich et al. [69] on other social networks. Investigations in [57] and [59] indicate that *social closeness*, or how familiar two users are to each other in real life, is a better indicator of home proximity. Therefore, many subsequent works are dedicated to going beyond online friendship and estimating social closeness instead.

In Twitter network, *mention* is another form of user interaction. When users mention each other or have conversation with each other, the two users are believed to have closer relationship or share similar interest. Such kind of 'friendship' is valuable in home location prediction. McGee et al. [23] make an analysis on 104,214 Twitter users with home located inside US. They find that besides mutual friendship through following, users' actions of mentioning and actively chatting with each other also indicate their home proximity. In a subsequent work [24], McGee et al. confirmed similar observations by examining a larger dataset. They also make more observations: 1) if the followed user account is a protected account[7] (typically an ordinary person), the two users are geographically close; and 2) local newspaper accounts are close to their followers. By treating geographical proximity as ground truth social closeness, McGee et al. trained a decision tree to assign social closeness between different users to ten quantiles with the above cues as features. They further use home distance in each social closeness quantile to fit Eq. (2), one model for each quantile.

Similar to McGee et al. [23], [24], Compton et al. [25] also exploit mentions between users. They build a user mention graph and optimize unknown home locations such that users mentioning each other are located as close as possible. Jurgens [26] also considers bidirectional mention relationship instead of friendship. Rahimi et al. [42] find that bidirectional mention are too rare to be useful. They adopt unidirectional mention as undirected edge.

Besides mentions and conversations as social closeness indicators, some studies also suggest *influence* to be another, but negative factor of social closeness. For example, a user

7. A protected account means that others need to get permissions to follow it, and its friend list and tweets are not public.

in Chicago may follow Lady Gaga in New York and President Obama in Washington. The establishment of such following relationship is not a result of social closeness between the user and the celebrities, but caused by the celebrities' social influence. The intuition in this example has been supported by a few studies. By analyzing a large Twitter dataset, Kwak et al. [60] find that users with fewer than 2,000 mutual friends (thus unlikely to have large influence) are more likely to be geographically close to most of them. In McGee's work [24] described earlier in this section, they also discover that a user u's friend who has many friends and followers tend to be further away from u.

In [30], Li et al. construct a user influence model to capture the above intuitions. Specifically, they model a user's influence as a bivariate Gaussian distribution centered at her location, with the variance of the distribution interpreted as her influence scope. The probability of user u_i following u_j is measured by the probability density of u_j's influence distribution at u_i's home location. Finally, all unknown home locations and influence scopes are treated as parameters and learnt from the data by Maximum Likelihood Estimation (MLE). Similarly, Yamaguchi et al. [62] propose a *landmark*-based home location prediction technique. Here, a landmark is a user with a lot of friends living in a small region. They argue that landmark friends are reliable cues to infer a user's home location. In this sense, landmarks are actually non-celebrities with small influence. In an extension [40] of their earlier work [30], Li et al. extend home location prediction to multiple location profiling. The motivation is that many people may have home cities, as well as working and college cities that may not coincide with their homes. They may not only follow friends living nearby and celebrities far away, but also colleagues and classmates in her working and college cities, respectively.

3.2.3 Local versus Global Inference

Given that users are connected by the Twitter network, predicting their home locations is technically different from a typical prediction task where objects to be classified/scored are independent. For most studies reviewed above, we only describe how to conduct *local* inference, i.e., predict a user's home location based on one- or two-hop friendship or mentioning. Even if friendship-based and social-closeness-based features are carefully designed, one may still face many problems when implementing a home location predictor. What if all friends of the current user have unknown home locations? Whether and how should an inferred home location be updated when the user's friends' home locations are updated via inference? In this section, we review some studies on how they deal with the above problems and how *global* inference is carried out.

The easiest global inference approach would be to apply local inference *iteratively* on users with unknown home locations (i.e., label propagation [70]). In each iteration, a user's home is temporally guessed through their friends with known or inferred locations. A few studies adopt this approach [26], [42], [56], [63]. However, it is also reported in [59] that simple iterations may reduce prediction accuracy. The authors find that iteratively making prediction causes the population distribution to be flatter, which contradicts with the common sense that most people live in densely populated areas. Therefore, they stick to local inference. In [57], Kong et al. conduct a variation of iterative inference called *confidence-based* iteration. The idea is to estimate a confidence for each home location guess, and only pass those with high confidence to the next iteration. Finally, it is worth noting that some studies define an explicit global objective function (or joint distribution) to optimize. Their inference methods are thus naturally global. Rahimi et al. [63] also find that label propagation would be biased by highly-connected nodes (i.e., celebrities with large amount of followers), and the nodes that are not connected to any labeled nodes could not be inferred. Therefore, they remove celebrities by identifying the number of mentions based on a graph constructed by mention relationship. For nodes with no labeled neighbors, they estimate the labels by the content-based method proposed in [42]. In [30], Li et al. derive from their global likelihood a two-stage iterative maximization method. Both unknown locations and influence scopes (recall in Section 3.2.2) are updated in each iteration. Compton et al. [25] directly optimize their objective function by parallel coordinate descent [71]. On the other hand, Rodrigues et al. [54] and Li et al. [40] resort to Gibbs sampling [72] to infer parameters in their joint distributions.

3.3 Inference Based on Tweet Context

In Section 2.1.3, we categorize various information associated with tweets as tweet context. Among them, tweet posting time and self-declared user profiles like locations and time zones are mainly employed information to help predict home location.

Mahmud et al. [37], [45] takes tweet posting time into consideration. In their dataset, all posting times are recorded in GMT. After binning a GMT day into time slots of equal length, users are viewed as distributions of tweet posting times. Since users in different time zones exhibit time shifts in their distribution, a time-zone classifier is trained with the distribution as features. Such classifications reveal the time zones of users and could provide a broad range of users' locations. In the work of Han et al. [31], [38], the authors observe that self-declared locations and time zones, as free texts, are not always accurate. For example, informal abbreviations like "mel" (for Melbourne) may occur. Therefore, besides tweet contents, they also include all four-grams of self-declared locations and time zones as features to train a home location classifier. Efstathiades et al. [65] simply utilize a probabilistic model based on the temporal distribution of geo-tags associated with tweets to estimate user home location and work place. The method is based on their observation that tweeting activity during rest time (i.e., late in the night) is more likely to be generated from "home" location, while during working time posting activity is mostly likely to be generated from "work" location. Poulston et al. [66] also leverage geo-tags, but they find that users usually have several active regions. Simply adopting the median as home location is not appropriate. Thus they cluster the geo-tags first, and the group with highest number of posts is considered as "home cluster". The geometric median of all points in "home cluster" is taken as home coordinate. Similarly, Cheng et al. [73] also group user's geo-tags into squares and the one with most number of geo-tags is regarded as the center. Instead of

taking geometric median directly, they repeat the process within the center area with finer cells until the square size is smaller than a predefined size. The final center is considered as the user's home location. By leveraging neural network model together with mixture density network, Rahimi et al. [67] convert two-dimensional geo-tags into continuous vector space and take them as input.

3.4 Summaries and Discussions

In this section, we review literatures on user home location prediction. We summarize the studies listed in Table 1. Techniques for home location prediction rely equally on tweet content and Twitter network. For tweet content, word-centric approaches are characterized by two components, i.e., local word identification and spatial word usage modeling. Location-centric approaches, on the other hand, cast the problem to classification or ranking problems. For Twitter network, dependencies between users' home locations are explained by their friendship and interactions. Global inference approaches are involved to solve the collective inference problem. Finally, tweet contexts like posting time and self-declared profiles are also involved in some studies.

Finally, we note that a systematic experimental comparison is conducted by Jurgens et al. [74]. The competing methods include Backstrom et al. [56], Kong et al. [57], Li et al. [30], [40], Mcgee et al. [24], Rout et al. [59], Davis et al. [53], Jurgens [26] and Compton et al. [25]. Their dataset consists of 1.3 billion tweets, 15 million users, and 26 million following relationships. Both self-declared home location and aggregation of geo-tags have been adopted as ground truth. Readers can refer to this experimental comparison for detailed results.

4 Tweet Location Prediction

According to an analysis by Java et al. [75], users' primary aims of sending out tweets are to share or to seek information. For example, one may tweet about a restaurant where she is enjoying delicious food. Such information will help promote the restaurant, if its name is clearly associated with the tweet as a tag. One may also send tweets saying she is lost when looking for a building. In this case, a tag on where the tweet is posted may enable her friends to give her precise directions. Unfortunately, it is reported that less than 1 percent of tweets have explicit geo-tags [76]. Therefore, predicting tweet location has received considerable attention.

At the first glance, tweet location prediction seems to be very similar to home location prediction. The "only" difference seems to be their inputs: for home location prediction we have all tweets from a user, while for tweet location prediction we are given only one tweet. In this section, we review literatures on tweet location prediction. We will also spend efforts to highlight different properties of the two problems, as well as different emphasis resulted on specific techniques.

4.1 Inference Based on Tweet Content

Due to similar problem definitions, tweet location and home location predictions share many common techniques on handling tweet content. For example, word-centric and location-centric methods, which we reviewed for home location prediction, are also observed in studies on tweet location prediction. We will detail those works in Section 4.1.1. Moreover, we will also review some topic-model-based approaches in Section 4.1.2, which are (most of the time) specially designed for tweet location prediction.

4.1.1 Word- or Location-Centric Methods

As summarized in Section 3.1, word-centric methods for home location prediction [30], [40], [41] are characterized by modeling spatial word usage. To alleviate the data sparsity issue, Gaussian or Gaussian mixture models are used to achieve smoothed word usage distributions [30], [40], [41]. Similarly, in [11], Priedhorsky et al. also employ Gaussian mixture models for tweet location prediction. However, they concentrate on modeling the spatial usage of not only words, but also n-grams. The reason lies in that, for tweet location prediction, we have only one tweet as the input. This information is much more limited than that for home location prediction, where a large number of tweets from a user are provided. Therefore, it is worthwhile to exploit the input with reasonable redundancy. In experiments, they find that their models are improved by including rare n-grams, even those occurring just three times. Flatow et al. [32] also resort to modeling spatial n-gram usage with Gaussian models. Similar to the idea of local words, they prefer *geo-specific* n-grams, i.e., those whose tweets are mostly located in a small eclipse on the map. Alternatively, Chong and Lim [77] apply a learning to rank method which encodes tweet content by a smoothed probability estimation that a word occurs at a venue. In their following work [78], word importance for different locations is distinguished. Since single tweet is short and of little information, they borrow the idea of query expansion and add words from the user's related historical tweets as supplement information. This is based on the assumption that users tend to visit same or related locations because of habits or constrains.

As for location-centric methods, previous studies also involve information-retrieval-based solutions. Kinsella et al. [79] treat both tweets and locations as Dirichlet-smoothed [80] unigram language models. The probability of a location language model generating a tweet, or the KL-divergence between language models of a tweet and a location, are adopted as location ranking functions. Li et al. [81] also employ an information-retrieval-based approach with KL-divergence as the retrieval function. For locations with few tweets, they augment their language models with web pages retrieved through their names. Similarly, Lee et al. [82] resort to user tips posted on the Foursquare pages of locations to construct language models for those locations. Besides Laplace smoothing (or add-one smoothing), they also try absolute discounting and Jelinek-Mercer smoothing [83] to deal with unseen words, but no performance gain is observed. Liu and Huang [84] apply Hidden-Markov-based model to infer tweet location on city-level. The observations are language models for each city based on geo-tagged tweets and the states are corresponding cities.

We also note that a few tweet location prediction studies involve classification-based approaches. Hulden et al. [85] classify tweet text into discretized cell grids with words as features. A data sparsity issue appears when grid size becomes too small. To deal with this problem, they apply a

TABLE 1
Summary of Studies on Home Location Prediction

Work	Input	Method	Dataset	Ground Truth	Granularity	Metrics
[36]	Content, network	Hybrid	Data from **[1], [60]**	Most frequent geo-tagged city, location profile	City, town	**MeanED, Acc@k, Acc**
[28]	Content	Word-centric	Data from [14], geo-tagged tweets	Most frequent geo-tagged city	City	MedianED, Acc, Acc@d, MeanED
[31]	Content, context	Hybrid	Data from [28], geo-tagged tweets	Most frequent geo-tagged city	City	MedianED, Acc, Acc@k
[27]	Content	Word-centric	Tweets	Location profile	Grid	Precision, Recall, MedianED
[8]	Content	Word-centric	Tweets	Location profile	Country, state	Acc
[37]	Content, context	Word-centric	Geo-tagged tweets	The earliest geo-tagged city	City, state, time-zone	Recall, Acc@d
[38]	Content, context	Classification	Data from [14], [28], tweets	The earliest geo-tagged city, most frequent geo-tagged city	Country, city	MedianED, Acc, Acc@d
[1], [7]	Content	Word-centric	Geo-tagged tweets	Most frequent geo-tagged city	City	MeanED, Acc@k, Acc
[9]	Content	Word-centric	Tweets	Median geo-tagged coordinates	Coordinates	MeanED
[40]	Content, network	Hybrid	Tweets	Location profile	City	Acc@k
[30]	Content, network	Hybrid	Tweets	Location profile	City	Acc, MeanED
[41]	Content	Word-centric	Data from [1]	Most frequent geo-tagged city	City	Acc, MeanED
[42]	Content, network	Hybrid	Data from [14], [28], [61]	The earlist geo-tagged coordinates, coordinates of the most frequent geo-tagged city	Coordinates	Acc@d, MeanED, MedianED
[44]	Content	Location-centric	CMU GeoText data	Geo-tag	Coordinates	MeanED, MedianED
[45]	Content, context	Location-centric	Geo-tagged tweets	The ealiest geo-tagged city	City	Recall, Acc
[13]	Content	Location-centric	Data from [14], [28]	Coordinates of the earlist tweet, coordinates of the most frequent geo-tagged city	Grid	MeanED, Acc@d, MedianED
[12]	Content	Location-centric	Wikipedia, data from [61]	Geo-tag	Grid	MeanED, MedianED
[14]	Content	Location-centric	Data from [61], geo-tagged tweets	The earlist geo-tagged coordinates	Grid	Acc@d, MeanED, MedianED
[50]	Content, network, context	Hybrid NN	**Data from [14]** and **W-NUT**	The earlist geo-tagged coordinates, majority vote of the closest city center	**City**	**MedianED, Acc, Acc@d, MeanED**
[51]	Content	MLP	Data from **[14], [28], [61]**	The earlist geo-tagged coordinates, coordinates of the most frequent geo-tagged city	**Grid**	**Acc@d, MeanED, MedianED**
[53]	Network	Friendship-only	Tweets	Most frequent geo-tagged city, location profile	City	Precision, Recall
[54]	Content, network	Friendship-only	Geo-tagged tweets	Most frequent geo-tagged city	City	Precision, Recall, F_1, Acc
[57]	Network	Friendship-only	Tweets, Gowalla check-in	Most frequent check-in, location profile	Coordinates	Acc, MeanED
[59]	Network	Friendship-only	Tweets	Location profile	City	Acc@d, MeanED
[24]	Network	Social-closeness based	Geo-tagged tweets	Median geo-tagged coordinates	Coordinates	Acc@d, MeanED
[25]	Network	Social-closeness based	Geo-tagged tweets	Location profile, median geo-tagged coordinates	Coordinates	Recall, MeanED, MedianED
[26]	Network	Social-closeness based	Geo-tagged tweets, Foursquare data	Location profile, median geo-tagged coordinates	Coordinates	MedianED
[62]	Network	Social-closeness based	Data from [30]	Location profile	Coordinates	Acc@d, Recall, F_1, MedianED, MeanED
[63]	Content, network	Hybrid	Data from [14], [28], [61]	The earlist geo-tagged coordinates, coordinates of the most frequent geo-tagged city	Coordinates	Acc@d, MeanED, MedianED
[61]	Content	Geo-topic	Geo-tagged tweets	The earliest geo-tagged city	State	MeanED, MedianED
[64]	Content	Geo-topic	Data from [61]	The earliest geo-tagged city	Coordinates	MeanED, MedianED
[65]	Context	Probabilistic	Data from [61], geo-tagged tweets	Location profile, work place on LinkedIn	POI	MeanED, Acc@d, Acc
[66]	Context	Clustering	Geo-tagged tweets	Manual label	Coordinates	MeanED, Acc
[67]	Context	NN with mixture density network	Data from [14], [61]	The earlist geo-tagged coordinates	Coordinates	Acc@d, Meand ED, MedianED

Works in bold are state-of-the-art methods based on the corresponding metrics and data. The same notations are used in the following tables.

Gaussian kernel to estimate the prior probability of each cell and the conditional probability of each word given a cell. Besides unigrams, Dredze et al. [86] also extract bigrams from tweet content, together with features derived from Twitter contexts, and feed them to a classifier. Cao et al. [87] employ both tweet content and social relationship features to classify tweet text to locations at fine-grained POI level. Another work [88] we are aware of aims at predicting location types, e.g., railway station, cinema or supermarket, rather than exact locations for tweets. The underlying reason may be again due to the large number of fine-grained tweet locations. For user home prediction, the number of classes, i.e., cities, are manageable under the multi-class classification framework. Some works even alleviate the class number issue by hierarchical classification [13], [37], [45]. However, the class number is simply unaffordable for tweet location prediction, given that there may be hundreds of thousands of POIs in a city. Iso et al. [89] adopt Neural Network model to predict tweet location. They utilize convolutional mixture density network which is fed by tweet content, to estimate the parameters of Gaussian mixture model, and employ the mode value of estimated density as the predicted coordinates for tweets. They claim that different loss functions do affect model performance.

4.1.2 Geo-Topic-Model-Based Methods

As effective approaches to unsupervised text mining, topic models have been extended to account for texts with geographical information like blogs [94], [95]. Such models are also expanded to tweets and used for geolocation on tweets due to their generative nature. Topic models could integrate different aspects related to locations as latent variables into a unified model, which could make information interact with each other, as we call them geo-topic-model-based method.

Eisenstein et al. [61] extend traditional topic models by "corrupting" conventional topics and produce location-varied topics. For example, "NBA" and "Kobe" may be representative words in "basketball" topic produced by conventional models. By sampling from a Gaussian distribution centered at the "basketball" topic vector, the corrupted "basketball" topic for Boston may also include "Celtics" (a Boston-based team) while slightly changing other word frequencies. In their subsequent work, Eisenstein et al. [64] propose a Sparse Additive GEnerative model (SAGE). The model is capable of supporting the idea of location-based topic corruption in [61]. It also enables sparsity and simplicity in model inference. An issue in these works [61], [64] lies in the special way they preprocess tweets. They concatenate each user's tweets into a long tweet, and use the first valid geographical coordinates as the location of the long tweet. We note that the two works are actually for home location prediction, and introduced here for the sake of a complete review of topic-model-based methods.

By leveraging SAGE model [64], Hong et al. [90] construct a model that takes region, topic and users' interests into consideration. Different from [61], [64], they respect the original view of tweets and model locations in a per-tweet manner. They assume tweet location depends on the user's geographical interest distribution. The topic of a tweet then depends on the user's topical interest, as well as local topics. Words in the tweet are finally generated by the chosen topic as well as a "local words" distribution. Instead of modeling users' geographical interest as a multinomial distribution, Chen et al. [91] introduce user interest as a latent variable and construct a *location function*, e.g., eating, shopping, or health, all of which are as bridges to link users and locations. Each user has an interest distribution over location functions, which affect tweets generation. Yuan et al. [2] propose an intermediate variable called *regions* between users and tweet locations. For example, a user may have a "work" region and a "home" region, which are Gaussian distributions centered at her work place and home address, respectively. Suppose the user is at her work region and wants to eat, i.e., choosing "eating" from her topical interests. She will pick a restaurant near her work place and write a tweet about eating and the work region, tagged with the name of the restaurant.

4.2 Inference Based on Twitter Network

Compared with home locations, tweet locations are usually described at a much finer granularity, i.e., POI-level rather than city-level, and are highly dynamic. Besides, tweets are usually short and noisy which increase the difficulty of predicting tweet location. To enrich available information, some works also try to align with friendship network.

In Sadilek's work [92], the dynamic input comes from real-time locations of a user's friends, and her own historical locations. To study the correlation between the trajectories of friends and the auto-correlation within one's trajectory, they accumulate over ten thousands of users, each with more than one hundred geo-tagged tweets. A Dynamic Bayesian Network (DBN) is trained on the location sequence of each user, with her friends' locations, the time of the day, and the day of the week as features. One interesting aspect of their model is that it not only models the attractive force between friends' locations but also captures other non-linear patterns. For example, two co-workers in the same store may have a day shift and a night shift. In this case, given enough historical data, their model can predict that one is at home given that the other is working in store. Chong and Lim [78] find that users with more similar tweet content history may be more similar in their venue visitation history. Collaborative filtering is adopted to propagate visitation information to users without location visiting history based on the similarity of historical tweet content. They provide us a new view that useful information can be obtained even from users without following or followed relationship.

4.3 Inference Based on Tweet Contexts

Tweet posting times are indicative of users' home locations, where a user is characterized by a distribution of posting times [37], [45]. Unlike home locations, for tweet location prediction we only access a tweet's posting time rather than a distribution. However, a time stamp may also be informative if enough historical data for locations are provided. For example, tweet posting histories may suggest that a club tends to be tweet-active at night, while a park tends to receive more tweets on weekends. Inspired by this, Li et al. [81] keep tweet time distributions for locations at three different scales of periods, i.e., day, week, and month. Given a tweet with a timestamp, probabilities of the three distributions generating the timestamp are linearly combined to give preferences between locations. In the geographic topic model of [2], Yuan et al. adopt two scales of time periods,

TABLE 2
Summary of Studies on Tweet Location Prediction

Work	Input	Model	Dataset	Ground Truth	Granularity	Metrics
[11]	Content	Word-centric	Data from [61], geo-tagged tweets	Geo-tag	Coordinates	MeanED, Precision, Recall
[32]	Content	Word-centric	**Geo-tagged tweets**	Geo-tag	Coordinates	**MeanED, Precision, Recall, F_1**
[77]	Content, Context	Ranking	Foursquare data, tweets	Foursquare check-ins	POI	MRR[8]
[78]	Content, network	Naive Bayes model	Foursquare data, tweets	Foursquare check-ins	POI	MRR, VMMR[9]
[79]	Content	Location-centric	Geo-tagged tweets	Geo-tag	Country, state, city, zip-code	Acc, Acc@k
[81]	Content, context	Location-centric	Geo-tagged tweets	Geo-tag	POI	Acc@k
[82]	Content	Location-centric	Foursquare data, geo-tagged tweets	Geo-tag	POI	Precision, Recall
[84]	Content	Location-centric	Geo-tagged tweets	Geo-tag	City	MeanED, MedianED, Acc
[85]	Content	Classification	Data from [61], geo-tagged tweets	The earliest geo-tagged coordinates, geo-tag	Coordinates	MeanED, MedianED
[86]	Content, context	Classification	Geo-tagged tweets	Geo-tag	Country, city	Acc, Acc@d, MedianED
[87]	Content, network	Classification	Geo-tagged tweets, Foursquare data	Geo-tag	POI	Acc@k, MeanED
[88]	Content	Classification	Geo-tagged tweets	Human label	POI	Precision, Recall, Acc
[89]	Content	Convolutional Mixture Density Network	Geo-tagged tweets	Geo-tag	Coordinates	MeanED, MedianED
[90]	Content	Geo-topic	Geo-tagged tweets	Geo-tag	Coordinates	MeanED
[91]	Content	Geo-topic	Geo-tagged Weibo data	Human label	POI	Acc, MeanED
[2]	Content, context	Geo-topic	Data from **[61]**, geo-tagged tweets	Geo-tag	Coordinates	**Acc, MeanED**
[92]	Network	Dynamic Bayesian network	Geo-tagged tweets	Geo-tag	Coordinates	Acc@d
[29]	Content, context	Stacking	Geo-tagged tweets	Geo-tag	Coordinates	MSE, MedianED, MeanED, Recall
[93]	Content, context	Classification	Geo-tagged tweets	Location check-ins	Location category	Acc

namely day (weekday/weekend) and time of the day. Given a user, the generative model first decides whether on weekdays or weekends to send the tweet according to her preference. Then the daytime is drawn from her preference distribution, which is also conditioned on the day variable. Finally, the user decides which region to go to and send a tweet about. Dredze et al. [86] take both time zone and tweet posting time as features for a classifier. They find the cyclical temporal patterns do have effects on prediction results.

Schulz et al. [29], on the other hand, accumulate tweet location indicators from user profiles. Possible indicators may be users' self-declared home locations, websites, and timezones, as well as location names mentioned in the tweet. By querying multiple databases,[10] those indicators are resolved to polygon-shaped administrative regions, with resolution confidences being heights of the polygons. Those polygons are finally stacked up [17] to produce a spatial distribution of possible tweet locations. In experiments, they find that such a multi-indicator approach is more robust than single-indicator approaches, which is error-prone due to ambiguity. Chong and Lim [77] provide another angle to utilize the context information and observe that both venues' active time and users' visiting place histories could help on tweet location prediction. They investigate venues' active time and estimate the probability that a location is popular given a time by a smoothed kernel density estimation method. Besides, they find an average user is spatially focused because she is usually constrained by geographical, social or personal factors. Thus, they encode this idea into the estimation of the probability that a user visits a location.

4.4 Summaries and Discussions

As listed in Table 2, we review literatures on tweet location prediction. Besides the fact that techniques for both tweet location and home location predictions emphasize much on employing tweet content, we also discuss several differences between home location and tweet location predictions. We list them below for a concise summary:

- Except studies with distance-based evaluations, home locations are predicted at coarse granularities like city, while tweet locations at a finer POI-level.
- Home location prediction relies equally on Twitter network and tweet content; but few studies utilize Twitter network to predict tweet locations.

8. Mean Reciprocal Rank.
9. Macro-averaged version of Mean Reciprocal Rank.
10. Those include GADM database of Global Administrative Areas (http://www.gadm.org), ThematicMapping (http://thematicmapping.org/downloads/world borders.php), and IANA Time Zone Database (http://efele.net/maps/tz/world/).

- Classification-based approaches are common for home location prediction, which is not the case for tweet location prediction.
- When employing posting time information, users are viewed as time distributions, while tweets are essentially time stamps. This may leads to different location ranking functions.

Despite the above differences, we note that the two problems not always clearly separated. Studies like [13], [14], [44], [61], [64] concatenate a users' tweets into one document, and use the first available geo-tag as the ground-truth location. We note that a geo-tag chosen this way may not necessarily to be the user's home location. Since users are not explicitly modeled, their techniques could be used for both prediction tasks. On the other hand, [2], [90], [91] explicitly model users' interests over locations, location functions, and regions. These models may only be used for tweet location prediction, but better exploitation for the specific problem and data could be expected from them.

5 Mentioned Location Prediction

Users occasionally send tweets to comment on a restaurant, a shopping mall, or a cinema, by treating Twitter as a life-logging platform. When parades or disasters take place, numerous tweets may be sent out by users to inform others about the events. Besides attaching geo-tags to those tweets, users may also reveal the relevant locations by mentioning their names in tweets. Preprocessing on the location names are crucial steps to accumulating information for, and performing subsequent analysis on, users and events [20], [96]. There are two steps for mentioned location processing: 1) recognition: to label text chunks which are potential location mentions, and 2) disambiguation: to map recognized location mentions to the right entries in a location database.

For well-formatted documents (e.g., news), the entity recognition [18] and disambiguation [19] problems have been investigated for decades. It is well received that the *variability* and *ambiguity* of entity mentions are two major difficulties for entity recognition and linking. Here variability means an entity may be mentioned in various surface forms, and ambiguity means one mention may refers to multiple entities. Unfortunately, the two difficulties are actually rendered more challenging by the noisy and short nature of tweets. In this section, we review recognition and disambiguation efforts for location mentions in tweets. We highlight how the two problems are made worse in the tweet scenario, and how they are dealt with by existing studies. Note that, we may not limit in studies solely on location entities. Recognition and disambiguation efforts of other types of entities in tweets will also be included in our survey, as long as they are inspiring to, and experimentally involve, mentioned location prediction.

5.1 Inference Based on Tweet Content

Like in ordinary documents, recognizing and disambiguating mentioned locations in tweets are generally based on tweet content, and are carried out in a pipelined manner. On the one hand, words like "Street" and "at" may suggest inner and outer boundaries of location mentions. On the other hand, other words in the mention context may provide clues for disambiguating the mentions. We will introduce previous works on both tasks in Sections 5.1.1 and 5.1.2, respectively. We also note that some studies propose joint approaches to couple the two tasks. They will be reviewed in the end of Section 5.1.2.

5.1.1 Mentioned Location Recognition

For Named entity recognition (NER) in formal documents, state-of-the-art machine learning algorithms like conditional random fields [113] have been designed. Equipped with comprehensive linguistic features like Part-of-Speech (POS) tags and capitalizations, they could achieve satisfactory performance [114]. Based on those algorithms and features, off-the-shelf NER tools like *StanfordNER*[11] and *OpenNLP*[12] are also developed and released.

When faced with noisy and short tweets, traditional NER features and tools are both at risk of deteriorated performance. For example, consider a typical tweet saying "shopping @ orchard st". Because of the informal writing, common clues indicating "Orchard Street" as a location mention in formal documents, like "at" ("@"), "street" ("st"), and capitalizations ("Orchard" instead of "orchard"), are all absent. Ritter et al. [97] rebuild the entire NER pipeline for tweets. They use Brown clustering [115] to identify word variations clusters (e.g., "at" and "@"). A dedicated classifier is also trained to recognize whether each capitalization in a tweet is informative. Similarly, Liu et al. [98], [99] train a tweet normalization model to correct informal words (e.g., "gooood" to "good") before performing NER. Noticing that words like "orchard" may be hard to label within the given short tweet, they train a k-nearest-neighbor word classifier to inform the NER classifier with global information, i.e., how the word is labeled in other tweets. Li et al. [100] investigate a novel streaming setting for tweet NER. They exploit the gregarious property of entity mentions to differentiate valid mentions from non-entity segments. Their approach also inherently addresses the short tweet problem.

Besides the above tweet NER attempts for general entities, there are also a few studies specially on location entities as shown in Table 3. Those studies are characterized by the use of location gazetteers, e.g., *Geonames*[13] [101], [106] and *Foursquare* [104], [105]. Malmasi et al. [101] do not involve CRF in their location mention recognizer. They simply use an off-the-shelf dependency parser to exact all noun phrases, and conduct fuzzy matching with *Geonames*. Their matching criteria take patterns of addresses and POIs into consideration. Zhang et al. [106] rely on a location mention recognizer they build in [102]. A gazetteer-based location parser, a CRF-based recognizer, and a rule-based street/building parser are used in conjunction to achieve best recall. A similar combination is also adopted by Gelernter et al. [103]. Li et al. [104], [105] observe that Twitter users often mention locations by abbreviations [116]. They opt to augment their Foursquare-based gazetteer with frequent-substring-based partial names.

5.1.2 Mentioned Location Disambiguation

Given location mentions recognized in a document, location disambiguation (i.e., linking) [19] refers to resolving those

11. http://nlp.stanford.edu/software/CRF-NER.shtml
12. http://opennlp.apache.org
13. http://www.geonames.org/

TABLE 3
Summary of Studies on Named Entity Recognition

Work	Input	Method	Dataset	NER type	Metrics
[97]	POS tagging, shallow parsing, capitalization	CRF	Tweets, Freebase	Location, person, etc.	Precision, Recall, F_1
[98], [99]	Contextual, dictionary, orthographic, lexical	KNN, CRF	Tweets	Location, person, etc.	Precision, Recall, F_1
[100]	Dictionary, statistical	Dynamic programming	Microsoft Web N-Gram, tweets	Location, person, etc.	Precision, Recall, F_1
[101]	POS tagging	Rule-based matching	Tweets	Location	Precision, Recall, F_1
[102]	Lemma form, POS tagging, capitalization, dictionary, contextual, orthographic	Named location recognizer, street and building parser, NER	Tweets, GeoNames	Location	Precision, Recall, F_1
[103]	Orthographic	NER, gazetteer matching, lexico-semantic pattern recognition	NGA gazetteer, tweets	Location	Precision, Recall, F_1
[104], [105]	Lexical, contextual, grammatical, BILOU schema, geographical	CRF	**Tweets, Foursquare**	**Location**	**Precision, Recall, F_1**

The ground truth in these studies are all based on human annotation.

mentions to right entries in a location database. The challenge of this task lies in that different locations may have the same names. For example, at the coarse city-level granularity, "Washington" may refer to a state in the west of the U.S., as well as a city in the east. "Olympia" may refer to the capital city of Washington state, as well as an ancient Greek city. At a finer POI-level, chained restaurants, e.g., McDonald, may have many branches in a city.

For general entities in formal documents, traditional approaches [117], [118], [119] disambiguate one mention at a time. To exploit dependencies between mentions, pair-wise fashioned [120], [121] and global collective disambiguation approaches [122], [123], [124] are proposed. Those approaches assume that the disambiguation decisions for multiple mentions in the same document should be *coherent*. For example, if "Washington" and "Olympia" co-occur in the same tweet, they are more likely to refer to the U.S. state and its capital. As for mentioned locations in tweets, Zhang et al. [106] employ similar ideas in their study. They take the hierarchy structure of locations into consideration. Not only parent-child location pairs (e.g., "Washington" and "Olympia"), but also siblings in the location hierarchy (e.g., cities in the same state), are regarded as coherent. Ji et al. [107] investigate collectively disambiguating POI mentions in tweets. Their coherence measure is based on the average distance among chosen POIs for the recognized mentions. Different from [106], [107], Li et al. [108] advocate disambiguation coherence at user-level rather than tweet-level. They assume that mentioned locations in a user's tweets are generally inside her living city. They first identify the living city by aggregating candidate locations for the mentions, and then refine those candidates with the living city. Shen et al. [109] also conduct collective disambiguation at user-level by modeling user interests. However, their method is aimed for general entities.

In conventional studies, mentioned location disambiguation is based on the output of recognition in a pipeline manner. If fed with wrong outputs, e.g., mentions with inaccurate boundaries, the disambiguation component may fail due to inability of finding candidates in the database. Motivated by this, recent studies [107], [110] suggest enabling information to flow in both directions between the two components. If the disambiguation component suffers from no candidates or low confidence, it may give feedbacks to the recognition component to correct the input mentions. In [110], Guo et al. leverage structural SVMs [125] to jointly optimize mention recognition and disambiguation. Both recognition features (e.g., capitalization) and disambiguation features (e.g., entity popularity) are integrated to train the structural SVM. Similarly, Ji et al. [107] jointly consider both types of features in a structural prediction framework. They resort to beam search [126] to look for the best combination of recognition and disambiguation decisions.

5.2 Inference Based on Twitter Network, Tweet Context

Like home and tweet location prediction, user friendship and contextual information could also be explored for mention disambiguation.

In [111], Hua et al. assume that the more a user is influenced by others mentioning an entity, the more likely she will mention the same entity. Specifically, they adopt an incremental disambiguation approach. In the offline stage, they preprocess a large number of tweets with [109] as a base system. Such preprocess enables them to estimate friendship-based user interest for entities in the online stage. When a candidate entity e is considered for a mention in user u's tweet, they look for other users who once mentioned e. An entity e is preferred if its users have good reachability to u in the friendship network. Besides friendship network, they also exploit time stamps of tweets. Due to their incremental disambiguation framework, they could estimate *entity recency* when a new tweet comes. Given a time stamp, the recency for an entity e is defined by the number of tweets mentioning e in the last time window of predefined length. They further use personalized PageRank [127] to propagate entity recency on the Wikipedia network to account for related entities. Finally, recently hot entities are rewarded when disambiguating mentions.

Fang et al. [112] consider both geo-tags and time stamps of tweets in mention disambiguation. An entity prior w.r.t.

TABLE 4
Summary of Models on Location Mention Linking

Work	Input	Model	Dataset	Ground Truth	Metrics
[106]	Content	Classification	Geo-tagged tweets	Human label	Precision, Recall
[107]	Content	Structured perceptron with multi-view learning	**Tweets**	Human label	**Precision, Recall, F_1**
[108]	Content	Ranking	Foursquare data, Geo-tagged tweets	Geo-tag	Precision, Recall, F_1
[109]	Content	Graph-based	Tweets	Human label	Acc
[110]	Content	Structural SVM	Tweets, some data from [97]	Human label	Precision, Recall, F_1
[111]	Network, context	Ranking	**Tweets**	NER identified by [100]	**Acc**
[112]	Content, context	Probabilistic	Geo-tagged tweets	Human label	Precision, Recall, F_1

All the studies are on the granularity of POI level.

time and location is estimated and used to replace the coarse-grained global entity popularity. Note that [111], [112] aim for general entities, not limiting to locations. When only locations are considered, the interaction between locations and timezones may enable interesting approaches. In Zhang et al.'s work [106], they attempt to disambiguate location mentions with time stamps. They observe that tweet traffic is fairly low between 2am-5am on weekdays. When there are several candidate locations (e.g., "Olympia"), they carefully choose one to avoid timezones that place the time stamp in the low traffic window.

5.3 Summaries and Discussions

In this section, we review literatures on mentioned location prediction as summarized in Table 4. Like tweet locations, mentioned locations also depend heavily on tweet content, and slightly on Twitter network and tweet context. However, we note that mentioned locations does not necessarily imply tweet locations (e.g., "going to Tokyo tomorrow"). In [128], Antoine et al. use a large volume of tweets to analyze the differences between mentioned locations and tweet locations. Moreover, due to the definitions, their ground truths are collected differently. Ground truths for tweet location prediction are obtained by referring to geo-tags of tweets. Mentioned locations, however, are mostly identified though human annotation [129].

Like predicting home and tweet locations, mentioned location prediction also suffers from the noisy and short nature of tweets. When adopting recognition and disambiguation approaches for formal documents, it is common to involve tweet- and location-specific techniques/information.

Finally, there are a few experimental analysis on tweet NER that are worth noting. Gelernter et al. [130] perform an error analysis on *StanfordNER* for recognizing locations in tweets. They do not retrain *StanfordNER* with labeled tweets, but use the off-the-shelf version. Lingad et al. [6] compare a few NER tools on disaster related Twitter data, e.g., *StanfordNER*, *OpenNLP*,[14] *Yahoo! PlaceMaker*,[15] and *TwitterNLP* [97]. They find that retrained *StanfordNER* outperforms the other competitors. Liu et al. [131] also make a similar comparison between *LER* proposed by themselves and other tools. Besides *StanfordNER* and *TwitterNLP*, they also include *GeoLocator* [103], and *UnlockText*.[16] Derczynski et al. [132] compare tweet NER performances of several systems, but they do not restrict to location entities.

6 Other Related Problems

In this section, we review two other problems related to location prediction on Twitter, namely *semantic location prediction* and *point-of-interest recommendation*. We will also try to highlight their differences in terms of definitions, ground truths, and solutions.

6.1 Semantic Location Prediction

In Section 4, we show that many studies depend tweet locations heavily on tweet content. The underlying assumption is that, if a tweet semantically talks about a location, it is likely to be posted at the venue. However, people could talk about New York where they visited before but currently locate in Japan. Thus, semantic locations and tweet locations may not always coincide. Therefore, some studies focus on predicting semantic locations instead of tweet locations.

Dalvi et al. [133] investigate matching users' tweets to restaurants in *Yahoo! Local*.[17] Those tweets may talk about dishes, service, or ambience of certain restaurants. They assume that each user has a latent location, and that they are likely to talk about nearby restaurants. When talking about restaurants, users follow a restaurant-specific bigram language model. To evaluate their model, they manually annotate hundreds of tweets, where candidate restaurants are suggested by a base system in their previous work. Zhao et al. [134] study matching tweets to general POIs on *Foursquare*. Different from [133] and other studies on tweet location prediction, they assume that geo-tags of tweets are known and given as input. Nearby locations with compatible keywords are preferred in the matching. By introducing dummy locations, their model is capable of identifying the "no semantic location" cases. Evaluations are conducted with thousands of manually annotated tweets.

To sum up, this line of work is characterized by the need of manually annotated ground truth due to the subjective definition of semantic location. We note that manual

14. http://opennlp.apache.org
15. http://developer.yahoo.com/geo/placemaker/
16. http://edina.ac.uk/unlock/texts/
17. http://local.yahoo.com, though it is offline now.

annotations take much more efforts to obtain than geo-tags. Dalvi et al. [133] and Zhao et al. [134] only involve hundreds or thousands of annotated tweets for evaluation respectively. This could explain why this problem attracts less attention than the three major tasks introduced above.

6.2 Point-of-Interest Recommendation

Due to its content-centric nature, Twitter is regarded by users as an ideal platform to share events, emotions, and opinions. Meanwhile, location-based social networks (LBSNs) like Foursquare, Gowalla, Brightkite, and Yelp concentrate more on POI-centric information. Besides establishing online friendships, they encourage users to check in, rate, and comment on POIs, as well as keep their information up to date. The popularity of LBSNs has given rise to abundant studies on POI recommendation.

Due to its popularity [135], Foursquare is adopted by many studies [3], [136], [137], [138], [139], [140], [141], [142], [143], [144] as data source. However, Foursquare APIs do not allow access to users' check-in history as reported in many studies. Luckily, when checking in on Foursquare, users may optionally allow Foursquare to send tweets like "I'm at [POI] [Foursquare URL of POI]." By monitoring Twitter streams, researchers manage to accumulate sufficient check-in data for POI recommendation. This might be the most significant connection between this line of study and Twitter. In the following, we clarify the differences between POI recommendation and main tasks in this survey.

Judging from the names, POI recommendation focuses on locations at fine-grained POI level. Moreover, it aims at suggesting POIs that users have never been to, instead of locations that they have connection with [145]. A user does not need to write a tweet to get suggested places to visit. Recommendations are made based on the user's and others' historical data, including check-ins, ratings, and comments, as well as context like the current time and user location. Finally, evaluation methods are also different: for each user in test set, visited POIs after some checkpoint time or selected samples are masked, predicted, and evaluated.

In terms of solutions, POI recommendations are generally based on collaborative filtering framework. Although user friendship, content and context are also exploited, they mostly come from LBSNs rather than Twitter. For friendship, Ye et al. and Gao et al. [136], [137], [138] employ Foursquare friendship network, while Ying et al. and Cho et al. [146], [147] rely on Gowalla and Brightkite networks. Yang et al. [139] claim that Foursquare friendship is not public,[18] and turn to Twitter network. As for content, check-in tweets do not provide as much textual information as ordinary tweets. However, several Foursquare-based studies manage to explore user comments [139], [140] and POI tags/descriptions [141], [142] in recommendation. Hasan et al. [148] find that the time of visiting different places depends on types of activities. Such spatio-temporal context is also involved in many other investigations [3], [137], [143].

This section is only aimed at clarifying connections and differences between LBSN-based POI recommendation and Twitter-based location prediction. Due to the scope of this survey, we only involve a small portion of recommendation studies. Readers may refer to [145], [149], [150] for extensive surveys and [151] for an experimental evaluation.

18. By the time we finish this survey, authorizations from Foursquare users are needed to access their friends via API. However, one can view any user's friend list via a browser.

7 Conclusion and Future Work

In this survey, we review and summarize techniques of three geolocation problems on Twitter: home location, tweet location, and mentioned location. Compared with similar problems on formal documents, i.e., document geolocation and named entity recognition & disambiguation, geolocation problems on Twitter face unique challenges and opportunities. The challenges generally arise from the noisy and short nature of tweet content. The opportunities, on the other hand, are enabled by the massive Twitter network and rich tweet context.

All the three prediction problems rely heavily on tweet content. For home and tweet location prediction, techniques could be categorized to the following two classes:

- Word-centric methods. They are characterized by identifying local words and modeling spatial word usage. Statistical, information theory and heuristic rule-based methods are designed to select location indicative words without supervision. Researchers also consider supervised ways to identify local words based on manual features and annotations. When modeling spatial word usage, direct estimations from data may suffer from sparsity problem. Therefore, multiple smoothing techniques are proposed.
- Location-centric methods. They are characterized by constructing pseudo-documents or classifiers for locations. Pseudo-documents construction are essential for information-retrieval-inspired approaches. Similar to spatial word usage, language models for pseudo-documents also require smoothing. However, geographical smoothing techniques, e.g., Gaussian model and grid-based smoothing, are not applicable. For tweet location prediction, classification methods are rarely adopted because it is usually at fine-grained POI level.

As for mentioned location, efforts on recognition address the noisy-content challenge by sophisticated features and comprehensive gazetteers. Collective disambiguation is employed to relieve the information scarcity brought by short tweets. Jointly optimizing both recognition and disambiguation components is also advocated in some studies.

As a significant feature of the platform, Twitter network plays a key role in home location prediction. Various hypotheses have been made on the connections between friendship and home proximity. Inspired by Backstrom et al. [56], many works try to formulate the relationship between the probability of friendship and home location distance. However, the indication is not very strong on Twitter. To fix this issue, social-closeness-based methods are proposed to differentiate noisy friendship. Explicit factors like friends with interactions are employed as useful information to predict home proximity. Implicit factors like influence scope are captured by sophisticated models. Finally, we note that Twitter network causes the predictions for different users to depend on each other. Therefore, it is necessary to involve global inference approaches.

Though short in length, tweets are accompanied with rich context. Those include timestamps and geo-tags associated with tweets, as well as various attributes in user profiles. Among them, temporal information like tweet timestamps and user-declared timezones are effective in implying tweet and home locations at coarse-grained granularity. Geo-tags and timestamps are also proven informative for disambiguating mentioned locations and other types of entities. Finally, we relate semantic location prediction for tweets and LBSN-based POI recommendation. We note that spatio-temporal factors are modeled in a more sophisticated manner in LBSN-based POI recommendation.

Geolocation is not only tackled on Twitter, but also many other platforms like Facebook [56], Foursquare [26], Gowalla [57], etc. The prediction models proposed based on Twitter can also be adapted to other social media sites, while might require some changes. But before considering model adaptations, we need to be clear on whether the three geolocation problems on Twitter, i.e., prediction of home location, tweet location and mentioned location, are applicable to the target platform or not. For example, tweet and mentioned location prediction on some image and video sharing platforms like Instagram and Pinterest may not be applicable. Next, the differences of available information, i.e., content, network, context, between the target platform and Twitter is another main consideration adapting the models on Twitter to other platforms. An example is that the friendship relationship on Facebook is bidirectional, but is unidirectional on Twitter.

At last, we would like to suggest some future directions. First, deep learning methods demonstrate great ability of learning feature representations automatically. A few recent works [50], [51] tried to apply neural network models directly to geolocation problems on Twitter and achieved some progress. Appropriate combination of Twitter properties and neural networks on geolocation deserves further research.

Second, most of current reviewed methods mainly focus on content information. The usage of network and context is not well investigated, especially for tweet and mentioned location prediction. In addition, the interactions among content, context, and network are not well analyzed. Most of current methods assume them to be independent features and combine them in a linear fashion. Joint modeling and exploiting of those factors could be a possible direction.

Third, data sparsity is a major issue for geolocation problem, especially for tweet and mentioned location prediction. Effective methods to augment useful information leave a big room to improve. Reliable images or cross-platform information might help to improve the performance. The exploration of appropriate approaches to leverage auxiliary knowledge also need more research.

Acknowledgments

Xin Zheng is in the SAP Industrial Ph.D Program, partially funded by the Economic Development Board and the National Research Foundation of Singapore. This work was partially supported by Singapore Ministry of Education Academic Research Fund MOE2014-T2-2-066.

References

[1] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating twitter users," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2010, pp. 759–768.

[2] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Who, where, when and what: Discover spatio-temporal topics for twitter users," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, 2013, pp. 605–613.

[3] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining user mobility features for next place prediction in location-based services," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 1038–1043.

[4] V. Rakesh, C. K. Reddy, D. Singh, and M. Ramachandran, "Location-specific tweet detection and topic summarization in twitter," in *Proc. Adv. Social Netw. Anal. Mining*, 2013, pp. 1441–1444.

[5] J. Ao, P. Zhang, and Y. Cao, "Estimating the locations of emergency events from twitter streams," in *Proc. Int. Conf. Inf. Technol. Quantitative Manage.*, 2014, pp. 731–739.

[6] J. Lingad, S. Karimi, and J. Yin, "Location extraction from disaster-related microblogs," in *Proc. World Wide Web Conf. Companion Volume*, 2013, pp. 1017–1020.

[7] Z. Cheng, J. Caverlee, and K. Lee, "A content-driven framework for geolocating microblog users," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, 2013, Art. no. 2.

[8] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: The dynamics of the location field in user profiles," in *Proc. Conf. Human Factors Comput. Syst.*, 2011, pp. 237–246.

[9] K. Ryoo and S. Moon, "Inferring twitter user locations with 10 km accuracy," in *Proc. World Wide Web Conf. Companion Volume*, 2014, pp. 643–648.

[10] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located twitter as proxy for global mobility patterns," *Cartography Geographic Inf. Sci.*, vol. 41, no. 3, pp. 260–271, 2014.

[11] R. Priedhorsky, A. Culotta, and S. Y. Del Valle, "Inferring the origin locations of tweets with quantitative confidence," in *Proc. ACM Conf. Comput. Supported Cooperative Work Social Comput.*, 2014, pp. 1523–1536.

[12] B. P. Wing and J. Baldridge, "Simple supervised document geolocation with geodesic grids," in *Proc. Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol.*, 2011, pp. 955–964.

[13] B. Wing and J. Baldridge, "Hierarchical discriminative classification for text-based geolocation," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 336–348.

[14] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised text-based geolocation using language models on an adaptive grid," in *Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn.*, 2012, pp. 1500–1510.

[15] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: Geotagging web content," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2004, pp. 273–280.

[16] W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh, "On assigning place names to geography related web pages," in *Proc. ACM/IEEE-CS Joint Conf. Digit. Libraries*, 2005, pp. 354–362.

[17] A. Woodruff and C. Plaunt, "GIPSY: Automated geographic indexing of text documents," *J. Amer. Soc. Inf. Sci.*, vol. 45, no. 9, pp. 645–655, 1994.

[18] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[19] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 443–460, Feb. 2015.

[20] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Comput. Surveys*, vol. 47, no. 4, 2015, Art. no. 67.

[21] F. Melo and B. Martins, "Automated geocoding of textual documents: A survey of current approaches," *Trans. GIS*, vol. 21, no. 1, pp. 3–38, 2017.

[22] O. Ajao, J. Hong, and W. Liu, "A survey of location inference techniques on twitter," *J. Inf. Sci.*, vol. 41, no. 6, pp. 855–864, 2015.

[23] J. McGee, J. A. Caverlee, and Z. Cheng, "A geographic study of tie strength in social media," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2011, pp. 2333–2336.

[24] J. McGee, J. Caverlee, and Z. Cheng, "Location prediction in social media based on tie strength," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2013, pp. 459–468.

[25] R. Compton, D. Jurgens, and D. Allen, "Geotagging one hundred million twitter accounts with total variation minimization," in *Proc. IEEE Int. Conf. Big Data*, 2014, pp. 393–401.
[26] D. Jurgens, "That's what friends are for: Inferring location in online social media platforms based on social relationships," in *Proc. Int. Conf. Weblogs Social Media*, 2013, pp. 273–282.
[27] Y. Yamaguchi, T. Amagasa, H. Kitagawa, and Y. Ikawa, "Online user location inference exploiting spatiotemporal correlations in social streams," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2014, pp. 1139–1148.
[28] B. Han, P. Cook, and T. Baldwin, "Geolocation prediction in social media data by finding location indicative words," in *Proc. Conf. Comput. Linguistics: Tech. Papers*, 2012, pp. 1045–1062.
[29] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser, "A multi-indicator approach for geolocalization of tweets," in *Proc. Int. Conf. Weblogs Social Media*, 2013, pp. 573–582
[30] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: Unified and discriminative influence model for inferring home locations," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, 2012, pp. 1023–1031.
[31] B. Han, P. Cook, and T. Baldwin, "A stacking-based approach to twitter user geolocation prediction," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Syst. Demonstrations*, 2013, pp. 7–12.
[32] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, "On the accuracy of hyper-local geotagging of social media content," in *Proc. ACM Conf. Web Search Data Mining*, 2015, pp. 127–136.
[33] O. V. Laere, J. A. Quinn, S. Schockaert, and B. Dhoedt, "Spatially aware term selection for geotagging," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 221–234, Jan. 2014.
[34] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. vol. 26, Boca Raton, FL, USA: CRC press, 1986.
[35] B. D. Ripley, *Spatial Statistics*. vol. 575, Hoboken, NJ, USA: Wiley, 2005.
[36] K. Ren, S. Zhang, and H. Lin, "Where are you settling down: Geo-locating twitter users based on tweets and social networks," in *Proc. Asia Inf. Retrieval Symp.*, 2012, pp. 150–161.
[37] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? inferring home locations of twitter users," in *Proc. Int. Conf. Weblogs Social Media*, 2012, pp. 511–514.
[38] B. Han, P. Cook, and T. Baldwin, "Text-based twitter user geolocation prediction," *J. Artif. Intell. Res.*, vol. 49, no. 1, pp. 451–500, 2014.
[39] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial variation in search engine queries," in *Proc. Conf. World Wide Web*, 2008, pp. 357–366.
[40] R. Li, S. Wang, and K. C.-C. Chang, "Multiple location profiling for users and relationships from social network and content," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1603–1614, 2012.
[41] H. W. Chang, D. Lee, M. Eltaher, and J. Lee, "@ phillies tweeting from philly? predicting twitter user locations with spatial word usage," in *Proc. Conf. Adv. Social Netw. Anal. Mining*, 2012, pp. 111–118.
[42] A. Rahimi, D. Vu, T. Cohn, and T. Baldwin, "Exploiting text and network context for geolocation of social media users," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2015, pp. 1362–1367.
[43] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statistical Soc.. Series B (Methodological)*, pp. 267–288, 1996.
[44] M. Cha, Y. Gwon, and H. T. Kung, "Twitter geolocation and regional classification via sparse coding," in *Proc. Int. Conf. Web Social Media*, 2015, pp. 582–585.
[45] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 47:1–47:21, 2014.
[46] C. N. Silla Jr and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining Knowl. Discovery*, vol. 22, no. 1–2, pp. 31–72, 2011.
[47] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1998, pp. 275–281.
[48] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3–4, pp. 237–264, 1953.
[49] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, "A simple scalable neural networks based model for geolocation prediction in twitter," in *Proc. Workshop Noisy User-Generated Text*, 2016, pp. 235–239.
[50] 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. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1260–1272.
[51] A. Rahimi, T. Cohn, and T. Baldwin, "A neural model for user geolocation and lexical dialectology," in *Proc. Annu. Meeting Assoc. Comput. Linguistics, Volume 2: Short Papers*, 2017, pp. 209–216.
[52] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociology*, vol. 27, no. 1, pp. 415–444, 2001.
[53] 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, 2011.
[54] E. C. Rodrigues, R. Assunção, G. L. Pappa, D. R. R. Oliveira, and W. M. Jr, "Exploring multiple evidence to infer users' location in twitter," *Neurocomputing*, vol. 171, no. C, pp. 30–38, 2016.
[55] S. Z. Li, *Markov Random Field Modeling in Image Analysis, ser. Advances in Pattern Recognition*. Berlin, Germany: Springer, 2009.
[56] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proc. Conf. World Wide Web*, 2010, pp. 61–70.
[57] L. Kong, Z. Liu, and Y. Huang, "SPOT: Locating social media users based on social network context," *Proc. VLDB Endowment*, vol. 7, no. 13, pp. 1681–1684, 2014.
[58] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Sci.*, vol. 311, no. 5757, pp. 88–90, 2006.
[59] D. Rout, K. Bontcheva, D. Preoţiuc-Pietro, and T. Cohn, "Where's @wally?: A classification approach to geolocating users based on their social ties," in *Proc. ACM Conf. Hypertext Social Media*, 2013, pp. 11–20.
[60] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proc. Conf. World Wide Web*, 2010, pp. 591–600.
[61] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2010, pp. 1277–1287.
[62] Y. Yamaguchi, T. Amagasa, and H. Kitagawa, "Landmark-based user location inference in social media," in *Proc. Conf. Online Social Netw.*, 2013, pp. 223–234.
[63] A. Rahimi, T. Cohn, and T. Baldwin, "Twitter user geolocation using a unified text and network prediction model," in *Proc. Meeting Assoc. Comput. Linguistics Joint Conf. Natural Language Process. Asian Fed. Natural Language Process.*, 2015, pp. 630–636.
[64] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1041–1048.
[65] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos, "Identification of key locations based on online social network activity," in *Proc. IEEE/ACM Conf. Adv. Social Netw. Anal. Mining*, 2015, pp. 218–225.
[66] A. Poulston, M. Stevenson, and K. Bontcheva, "Hyperlocal home location identification of twitter profiles," in *Proc. ACM Conf. Hypertext Social Media*, 2017, pp. 45–54.
[67] A. Rahimi, T. Baldwin, and T. Cohn, "Continuous representation of location for geolocation and lexical dialectology using mixture density networks," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2017, pp. 167–176.
[68] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo, "Socio-spatial properties of online location-based social networks," in *Proc. Int. Conf. Weblogs Social Media*, 2011, pp. 329–336.
[69] Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, and A. Kaltenbrunner, "The length of bridge ties: Structural and geographic properties of online social interactions," in *Proc. Int. Conf. Weblogs Social Media*, 2012, pp. 346–353.
[70] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CALD-02-107, 2002.
[71] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *Math. Program.*, vol. 156, no. 1–2, pp. 433–484, 2016.
[72] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, no. 1–2, pp. 5–43, 2003.
[73] Z. Cheng, J. Caverlee, H. Barthwal, and V. Bachani, "Who is the barbecue king of texas?: A geo-spatial approach to finding local experts on twitter," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 335–344.

[74] 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. Int. Conf. Web Social Media*, 2015, pp. 188–197.

[75] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: An analysis of a microblogging community," in *Proc. Workshop Knowl. Discovery Web Workshop Social Netw. Anal.*, 2007, pp. 118–138.

[76] M. Graham, S. A. Hale, and D. Gaffney, "Where in the world are you? geolocation and language identification in twitter," *The Prof. Geographer*, vol. 66, no. 4, pp. 568–578, 2014.

[77] W. Chong and E. Lim, "Exploiting contextual information for fine-grained tweet geolocation," in *Proc. Int. Conf. Web Social Media*, 2017, pp. 488–491.

[78] W. Chong and E. Lim, "Tweet geolocation: Leveraging location, user and peer signals," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 1279–1288.

[79] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in glasgow: Modeling locations with tweets," in *Proc. CIKM Workshop Search Mining User-Generated Contents*, 2011, pp. 61–68.

[80] C. Zhai and J. D. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2001, pp. 334–342.

[81] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson, "The where in the tweet," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2011, pp. 2473–2476.

[82] K. Lee, R. K. Ganti, M. Srivatsa, and L. Liu, "When twitter meets foursquare: Tweet location prediction using foursquare," in *Proc. Conf. Mobile Ubiquitous Syst.: Comput., Netw. Services*, 2014, pp. 198–207.

[83] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. Meeting Assoc. Comput. Linguistics*, 1996, pp. 310–318.

[84] Z. Liu and Y. Huang, "Where are you tweeting?: A context and user movement based approach," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2016, pp. 1949–1952.

[85] M. Hulden, M. Silfverberg, and J. Francom, "Kernel density estimation for text-based geolocation," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 145–150.

[86] M. Dredze, M. Osborne, and P. Kambadur, "Geolocation for twitter: Timing matters," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2016, pp. 1064–1069.

[87] B. Cao, F. Chen, D. Joshi, and P. S. Yu, "Inferring crowd-sourced venues for tweets," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 639–648.

[88] S. Hahmann, R. S. Purves, and D. Burghardt, "Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes," *J. Spatial Inf. Sci.*, vol. 2014, no. 9, pp. 1–36, 2014.

[89] H. Iso, S. Wakamiya, and E. Aramaki, "Density estimation for geolocation via convolutional mixture density network," *CoRR*, vol. abs/1705.02750, 2017, http://arxiv.org/abs/1705.02750

[90] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proc. Conf. World Wide Web*, 2012, pp. 769–778.

[91] Y. Chen, J. Zhao, X. Hu, X. Zhang, Z. Li, and T.-S. Chua, "From interest to function: Location estimation in social media," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 180–186.

[92] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends and following them to where you are," in *Proc. Conf. Web Search Data Mining*, 2012, pp. 723–732.

[93] A. Galal and A. El-Korany, "Enabling semantic user context to enhance twitter location prediction," in *Proc. Int. Conf. Agents Artif. Intell.*, 2016, pp. 223–230.

[94] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, "Mining geographic knowledge using location aware topic model," in *Proc. ACM Workshop Geographic Inf. Retrieval*, 2007, pp. 65–70.

[95] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proc. Conf. World Wide Web*, 2006, pp. 533–542.

[96] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," in *Proc. IEEE Conf. Visual Analytics Sci. Technol.*, 2011, pp. 181–190.

[97] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 1524–1534.

[98] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2011, pp. 359–367.

[99] X. Liu, F. Wei, S. Zhang, and M. Zhou, "Named entity recognition for tweets," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, 2013, Art. no. 3.

[100] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2012, pp. 721–730.

[101] S. Malmasi and M. Dras, "Location mention detection in tweets and microblogs," in *Proc. Conf. Pacific Assoc. Comput. Linguistics*, 2015, pp. 123–134.

[102] J. Gelernter and W. Zhang, "Cross-lingual geo-parsing for non-structured data," in *Proc. Workshop Geographic Inf. Retrieval*, 2013, pp. 64–71.

[103] J. Gelernter and S. Balaji, "An algorithm for local geoparsing of microtext," *GeoInformatica*, vol. 17, no. 4, pp. 635–667, 2013.

[104] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 43–52.

[105] C. Li and A. Sun, "Extracting fine-grained location with temporal awareness in tweets: A two-stage approach," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 7, pp. 1652–1670, 2017.

[106] W. Zhang and J. Gelernter, "Geocoding location expressions in twitter messages: A preference learning method," *J. Spatial Inf. Sci.*, vol. 9, no. 1, pp. 37–70, 2014.

[107] Z. Ji, A. Sun, G. Cong, and J. Han, "Joint recognition and linking of fine-grained locations from tweets," in *Proc. Conf. World Wide Web*, 2016, pp. 1271–1281.

[108] G. Li, J. Hu, J. Feng, and K.-l. Tan, "Effective location identification from microblogs," in *Proc. IEEE Int. Conf. Data Eng.*, 2014, pp. 880–891.

[109] W. Shen, J. Wang, P. Luo, and M. Wang, "Linking named entities in tweets with knowledge base via user interest modeling," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, 2013, pp. 68–76.

[110] S. Guo, M. Chang, and E. Kiciman, "To link or not to link? A study on end-to-end tweet entity linking," in *Proc. Conf. HLT-NAACL*, 2013, pp. 1020–1030.

[111] W. Hua, K. Zheng, and X. Zhou, "Microblog entity linking with social temporal context," in *Proc. ACM SIGMOD Conf. Manage. Data*, 2015, pp. 1761–1775.

[112] Y. Fang and M. Chang, "Entity linking on microblogs with spatial and temporal signals," *Assoc. Comput. Linguistics*, vol. 2, pp. 259–272, 2014.

[113] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[114] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. Conf. Comput. Natural Language Learn.*, 2009, pp. 147–155.

[115] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Comput. Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[116] M. D. Lieberman, H. Samet, and J. Sankaranarayanan, "Geotagging with local lexicons to build indexes for textually-specified spatial data," in *Proc. IEEE Int. Conf. Data Eng.*, 2010, pp. 201–212.

[117] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, et al., "Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation," in *Proc. Conf. World Wide Web*, 2003, pp. 178–186.

[118] R. Mihalcea and A. Csomai, "Wikify!: Linking documents to encyclopedic knowledge," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2007, pp. 233–242.

[119] D. Milne and I. H. Witten, "Learning to link with wikipedia," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 509–518.

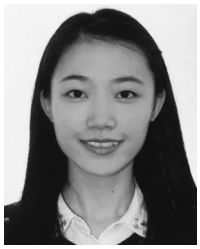
[120] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Proc. AAAI Workshop Wikipedia Artif. Intell.: An Evolving Synergy*, 2008, pp. 25–30.

[121] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of wikipedia entities in web text," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, 2009, pp. 457–466.

[122] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data," in *Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn.*, 2007, pp. 708–716.

[123] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 782–792.

[124] X. Han, L. Sun, and J. Zhao, "Collective entity linking in web text: A graph-based method," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 765–774.
[125] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, 2005.
[126] Y. Zhang and S. Clark, "Joint word segmentation and POS tagging using a single perceptron," in *Proc. Meeting Assoc. Comput. Linguistics*, 2008, pp. 888–896.
[127] G. Jeh and J. Widom, "Scaling personalized web search," in *Proc. Conf. World Wide Web*, 2003, pp. 271–279.
[128] É. Antoine, A. Jatowt, S. Wakamiya, Y. Kawai, and T. Akiyama, "Portraying collective spatial attention in twitter," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, 2015, pp. 39–48.
[129] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in twitter data with crowdsourcing," in *Proc. NAACL HLT Workshop Creating Speech Language Data Amazon's Mech. Turk.*, 2010, pp. 80–88.
[130] J. Gelernter and N. Mushegian, "Geo-parsing messages from microtext," *Trans. GIS*, vol. 15, no. 6, pp. 753–773, 2011.
[131] F. Liu, M. Vasardani, and T. Baldwin, "Automatic identification of locative expressions from social media text: A comparative analysis," in *Proc. Workshop Location Web*, 2014, pp. 9–16.
[132] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," *Inf. Process. Manage.*, vol. 51, no. 2, pp. 32–49, 2015.
[133] N. N. Dalvi, R. Kumar, and B. Pang, "Object matching in tweets with spatial models," in *Proc. Conf. Web Search Data Mining*, 2012, pp. 43–52.
[134] K. Zhao, G. Cong, and A. Sun, "Annotating points of interest with geo-tagged tweets," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2016, pp. 417–426.
[135] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *Proc. Int. Conf. Weblogs Social Media*, 2011, pp. 81–88.
[136] M. Ye, P. Yin, and W.-C. Lee, "Location recommendation for location-based social networks," in *Proc. SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 458–461.
[137] H. Gao, J. Tang, and H. Liu, "Exploring social-historical ties on location-based social networks," in *Proc. Int. Conf. Weblogs Social Media*, 2012, pp. 114–121.
[138] H. Gao, J. Tang, and H. Liu, "gSCorr: Modeling geo-social correlations for new check-ins on location-based social networks," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2012, pp. 1582–1586.
[139] D. Yang, D. Zhang, Z. Yu, and Z. Wang, "A sentiment-enhanced personalized location recommendation system," in *Proc. ACM Conf. Hypertext Social Media*, 2013, pp. 119–128.
[140] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 1721–1727.
[141] B. Liu and H. Xiong, "Point-of-interest recommendation in location based social networks with topic and location awareness," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 396–404.
[142] B. Liu, Y. Fu, Z. Yao, and H. Xiong, "Learning geographical preferences for point-of-interest recommendation," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2013, pp. 1043–1051.
[143] C. Cheng, H. Yang, M. R. Lyu, and I. King, "Where you like to go next: Successive point-of-interest recommendation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2605–2611.
[144] D. Yao, C. Zhang, J. Huang, and J. Bi, "SERM: A recurrent model for next location prediction in semantic trajectories," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 2411–2414.
[145] H. Gao, "Personalized POI recommendation on location-based social networks," Ph.D. dissertation, School of Computing, Informatics, and Decision Systems Engineering, Arizona State Univ., Tempe, Arizona, 2014.
[146] J. J.-C. Ying, E. H.-C. Lu, W.-N. Kuo, and V. S. Tseng, "Urban point-of-interest recommendation by mining user check-in behaviors," in *Proc. ACM SIGKDD Workshop Urban Comput.*, 2012, pp. 63–70.
[147] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, 2011, pp. 1082–1090.
[148] S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proc. ACM SIGKDD Workshop Urban Comput.*, 2013, pp. 6:1–6:8.
[149] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: A survey," *GeoInformatica*, vol. 19, no. 3, pp. 525–565, 2015.
[150] S. Zhao, I. King, and M. R. Lyu, "A survey of point-of-interest recommendation in location-based social networks," *CoRR*, vol. abs/1607.00647, 2016, http://arxiv.org/abs/1607.00647
[151] Y. Liu, T. Pham, G. Cong, and Q. Yuan, "An experimental evaluation of point-of-interest recommendation in location-based social networks," *Proc. VLDB Endowment*, vol. 10, no. 10, pp. 1010–1021, 2017.



Xin Zheng received the BE degree from Xiamen University, China, in 2014. She is a PhD candidate of the School of Computer Science and Engineering, Nanyang Technological University, Singapore. She is also a research associate with SAP Research and Innovation Singapore, SAP Asia Pte Ltd. Her research interests include information retrieval, text mining, and summarization on social media.

Jialong Han received the BE degree from the Renmin University of China, in 2010, and the PhD degree from the Renmin University of China, in 2015, under the supervision of Prof. Ji-Rong Wen. He is a researcher in the Tencent AI Lab. Prior to joining Tencent, he worked as a postdoctoral research fellow in the School of Computer Science and Engineering, Nanyang Technological University. His research interests include graph data mining and management, as well as their applications on knowledge graphs.

Aixin Sun received the BASc (First Class Honours) and PhD degrees in computer engineering from the School of Computer Science and Engineering, Nanyang Technological University, Singapore, in 2001 and 2004, respectively. He is an associate professor in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research areas include text mining, social computing, multimedia, and digital libraries.