Semantic Annotation of Mobility Data using Social Media

Fei Wu[†1] Zhenhui Li[†2] Wang-Chien Lee[‡3] Hongjian Wang[†4] Zhuojie Huang[§5]

[†]College of Information Sciences and Technology
[‡]Department of Computer Science and Engineering
[§]GeoVISTA Center
Pennsylvania State University, University Park, PA, USA
{[1]fxw133, [2]jessieli, [4]hxw186 }@ist.psu.edu, [3]wlee@cse.psu.edu, [5]seenhzj@gmail.com

ABSTRACT

Recent developments in sensors, GPS and smart phones have provided us with a large amount of mobility data. At the same time, large-scale crowd-generated social media data, such as geo-tagged tweets, provide rich semantic information about locations and events. Combining the mobility data and surrounding social media data enables us to semantically understand *why a person travels to a location at a particular time* (e.g., attending a local event or visiting a point of interest). Previous research on mobility data mining has been mainly focused on mining patterns *using only the mobility data*. In this paper, we study the problem of using social media to annotate mobility data. As social media data is often noisy, the key research problem lies in using the right model to retrieve only the relevant words with respect to a mobility record. We propose frequency-based method, Gaussian mixture model, and kernel density estimation (KDE) to tackle this problem. We show that KDE is the most suitable model as it captures the locality of word distribution very well. We test our proposal using the real dataset collected from Twitter and demonstrate the effectiveness of our techniques via both interesting case studies and a comprehensive evaluation.

Categories and Subject Descriptors

I.7.m [**Computing Methodologies**]: DOCUMENT AND TEXT PROCESSING

Keywords

Heterogeneous data; Mobility data; Microblogs; Semantics; Annotation; Keneral Density Estimation;

1. INTRODUCTION

With the rapid advancement of positioning technology and wide availability of mobile devices, we can now easily collect large-scale mobility data from mobile users. Given the location history of a mobile user, one of the most fundamental and important questions one may ask is: *What is the* ***purpose*** *for this person to visit a certain location at a particular time*? In other words, we wish to *understand the* ***semantics*** *of a person's mobility data*. Generally, the semantics could be the landmark information (e.g., a museum or a shopping mall) or information about the events attended (e.g., basketball game, movies or exhibition). The semantics provide us with richer and much more interpretable information about a mobile user, therefore can greatly benefit many applications such as advertisement targeting, personalized recommendation, and future movement prediction.

Mining mobility data has gained increasing interests recently. Researchers have studied various mobility patterns, such as frequent patterns [21], periodic behaviors [18], representative behaviors [13, 10], activity recognition [19, 15, 27, 31, 37]. However, all these patterns and activities are extracted *purely from the mobility data*. While they are useful for understanding the inherent moving behaviors of a mobile user, they do not provide any *contextual semantics* required for understanding the intended activities of the user.

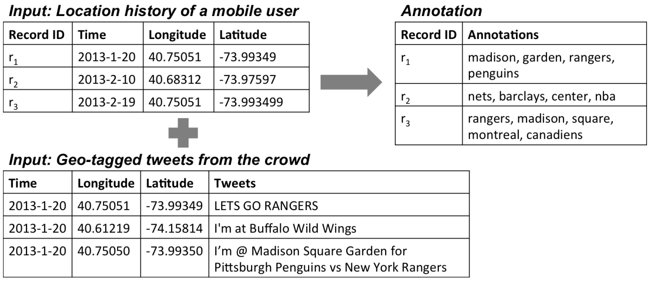
To understand the contextual semantics of the mobility data, it is necessary to harness *external surrounding locational data*. The external information considered in previous work [3, 1, 27, 32] is mostly *static*, such as landmarks [3, 1], landscapes [27], land-use categories [32]. The previous studies assume a spatial point (e.g., a location) always carries the same semantics regardless of the time. The static annotation will miss the important *dynamic event information*. For instance, Madison Square Garden, a well-known multi-purpose arena in New York City, holds both Knicks[1] and Rangers[2] games as well as many other events such as concerts and exhibitions. Different people (and even the same person) may visit Madison Square Garden for different purposes at different times. Consequently, simply using "Madison Square Garden" to annotate a user's location record could fail to reveal the complete purpose of his visit. Therefore, it is critical to examine the *dynamic events* associated with this location and use them for annotation.

Motivated by the example above, in this paper, we study the problem of annotating a user's location history with dynamic semantics, that reflect a user's interest or purpose at a particular location and time. To achieve our goal, an issue we face is to find the sources for time-sensitive information related to locations. Fortunately, popular location-based social media services, such as Twitter and Foursquare, facili-

[1] Knicks is a basketball team based in New York City

[2] Rangers is an ice hockey team based in New York City

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741675.



Input: Location history of a mobile user

Record ID	Time	Longitude	Latitude
r_1	2013-1-20	40.75051	-73.99349
r_2	2013-2-10	40.68312	-73.97597
r_3	2013-2-19	40.75051	-73.993499

Input: Geo-tagged tweets from the crowd

Time	Longitude	Latitude	Tweets
2013-1-20	40.75051	-73.99349	LETS GO RANGERS
2013-1-20	40.61219	-74.15814	I'm at Buffalo Wild Wings
2013-1-20	40.75050	-73.99350	I'm @ Madison Square Garden for Pittsburgh Penguins vs New York Rangers

Annotation

Record ID	Annotations
r_1	madison, garden, rangers, penguins
r_2	nets, barclays, center, nba
r_3	rangers, madison, square, montreal, canadiens

Figure 1: An example of semantic annotation on location history using social media.

tate mobile users to voluntarily generate enormous amount of spatiotemporal text. Such human-powered sensing data brings us rich and comprehensive information of dynamic local events.

In this paper, our goal is to annotate the location history of a mobile user using the spatiotemporal documents collected from social media, such as geo-tagged tweets. As illustrated in Figure 1, we assume two inputs: (i) the location history of a mobile user, represented by a sequence of $\langle time, location\rangle$ pairs; and (ii) a set of documents from social media with the time and location information attached. For each entry in a user's location history, we aim to derive a list of words that best describes the purpose of the user's visit to that location at that time. We name this word list as an *annotation document*. For example, in Figure 1, location record r_1 is annotated with words related to a Rangers' game with Penguins at Madison Square Garden. The annotation documents of a user can latter be used to create a profile of that user. For example, Figure 1 shows that the user has interests in the hockey team Rangers and the basketball team Brooklyn Nets.

However, the use of time-sensitive social media data for dynamic annotation comes with a price. Specifically, in the previous studies on static annotation [3, 1, 27, 32], a set of *clean, predefined and fixed* attributes is often provided as an input. Therefore, one can simply annotate the *closest* location attribute to a given location. But for our problem, we need to automatically and dynamically model the spatial distributions of words from social media and use only the words that are relevant to local venues and events for annotation. This brings several challenges in practice. First, social media data (e.g., tweets) are extremely *noisy*, containing numerous irrelevant documents about casual chatting with friends, retweeting news, and posting personal opinions. Second, as there are many landmarks and various events nearby, the word frequency in the neighborhood is often dominated by a small number of popular venues and events. Therefore, in order to discover the true semantics of a querying location record, we need to find a suitable model to identify the relevant words.

In this paper, we first look into two simple methods for annotation: a frequency based method and a Gaussian mixture model based method. We will discuss the limitations of these methods and justify the reason of choosing a more suitable model - Kernel Density Estimation (KDE) model for our problem. In essence, KDE model well captures both the *locality* and the *relevance* of the words w.r.t. a given location record. Further, as the estimated spatial distribution by the KDE model is controlled by a bandwidth parameter h, we will analyze several options of choosing the parameter h. Finally, we conduct extensive experiments to demonstrate the effectiveness of our proposed method.

In summary, the main contributions of this paper are:

- We study an interesting problem of inferring a user's interests and purposes from his location history using external contextual information (e.g., tweets). To the best of our knowledge, we are the first to use time-sensitive social media data to generate dynamic annotations for the mobility data.
- We propose to study different methods and identify the most suitable model for semantic annotation using noisy and dynamic social media data.
- The effectiveness of our method is demonstrated using both quantitative evaluations and case studies on a large geo-tagged tweet data.

The rest of the paper is organized as follows. We first review the related work in Section 2. The semantic annotation problem is then formulated in Section 3. We describe our proposed method in Section 4. Case studies and quantitative evaluations on real data are presented in Section 5. We further discuss an application of our method to user interest profiling in Section 6, and conclude the paper in Section 7.

2. RELATED WORK

Mobility Pattern Mining. In literature, numerous methods have been proposed to extract patterns from the mobility data. Representative works include stop and move detection [3, 2, 24], activity recognition (e.g., biking and walking) [19, 15, 27, 31, 37], significant place extraction [39, 24, 38, 7] and frequent regular pattern discovery [21, 13, 18, 10], just to name a few. These works mainly focus on the inherent *trajectory patterns*. While the patterns describe the mobility records, they do *not* explore external source to discover the contextual semantics.

Static Annotation. To mine contextual semantics of mobility data, existing methods have been using various types of *static information* including landmarks [3, 1], landscape and environment [27], and land-use categories [32, 34]. In these papers, a location (i.e., a point, a region, or a road segment) is associated with a set of *predefined and fixed* attributes, such as a landmark (e.g., "Eiffel tower") or a land type (e.g., residential area or business center). A location on the trajectory is then annotated using the attributes of nearby locations [3, 27]. Yan et al. [32, 34, 33] extend the point-based annotation to three kinds of objects: points, lines, and regions based on spatial join, using direction, distance, and topological spatial relations such as intersection. In addition to landmarks and land-use categories, Yan et al. [32, 33] further consider transportation modes to determine the type of POIs for trajectory annotations via a hidden Markov model. While the contextual semantics used in these studies are static, the contextual semantics in our problem need to be *dynamically extracted* from social media data and contain richer location information.

Local Word Detection. There have been extensive studies in mining geo-tagged social media data. One line of research work that is highly relevant to our problem is local word detection. The general premise is that local words should have concentrated spatial distributions around their location centers. Based on this observation, Backstrom et al. [4] proposes a spatial variation model to detect local words. This method has been used in [9] to detect local words in tweets. Meanwhile, [22] uses spatial discrepancy to detect spatial bursts. In these work, to determine whether the word is local or not, each word is assigned with a locality score. In our problem setting, local word detection may serve as a filter to select only important local words for annotation. However, as shown in our experiment results in Section 5.5, this filtering step is *not necessary* and could even be *erroneous.*

Microblogs Summarization. Our problem is also related to the microblogs summarization problem. On documents, researchers have developed summarization methods based on word frequency [26, 30], cluster of sentences [25, 29], and graph of sentences [14, 23]. As microblogs are short and contain informal use of the language, methods based on word frequency have been shown to perform the best [17, 8]. However, the summarization methods do not model the spatial characteristics of words, which is essential in our problem. While the concept of using frequency may be applied, we will discuss the problem with of frequency based methods in Section 4.1 and experiments.

3. PRELIMINARIES

3.1 Problem Definition

In this paper we consider location data of one mobile user as a set of spatiotemporal points, $U = \{r_1, r_2, ..., r_n\}$, where each record $r_i = (loc_i^U, t_i^U)$. The location loc_i^U is a geographic coordinate, and t_i^U is the timestamp. Location data can be collected from a variety of platforms such as mobile phones and web services.

We consider the external context data as a set of spatiotemporal documents $\mathcal{D} = \{d_1, d_2, ..., d_m\}$. Each document d_j can be represented as a (W_j, loc_j^D, t_j^D) tuple, where W_j is the set of words in d_j, loc_j^D is a geographic coordinate, and t_j^D is a timestamp. We define $\mathcal{W} = \bigcup_j W_j$ as the set of all words (uni-gram) in $\mathcal{D}$. Examples of external context documents are geo-tagged tweets or other spatiotemporal documents from location-based services (e.g., Flickr). For simplicity, we only consider uni-gram in this paper. However, our model is also applicable for N-grams.

Given the location history of a user U and a source of external context documents $\mathcal{D}$, our goal is to obtain an annotation document that is potentially associated with a mobility record $r_i \in U$. Formally, an annotation document for a mobility record r_i is a set of relevant words, $\mathcal{A}(r_i) = \{(w, s_w(r_i)) | w \in \mathcal{W}, s_w(r_i) > \theta\}$, where $s_w(r_i)$ is a function that measures the relevance of a word w to the record r_i.

3.2 Problem Analysis

The key research problem is to find a relevance function $s_w(r_i)$ for a given location record r_i. A straightforward solution is to use the words frequently appearing near the location of r_i for annotation. However, setting the distance threshold (for defining near) is non-trivial. Another issue with this approach is that all the words within the distance threshold are treated equally regardless of their distances to the location record r_i.

To avoid setting a hard distance threshold, an alternative idea is to model the distribution of words in order to annotate the words based on their probabilities (or densities) at the given location r_i. One frequently-used model for spatial distribution in practice is the Gaussian mixture model [10]. While we have adopted it as the second proposed method, we suspect that Gaussian model may suffer from several potential drawbacks. First, the number of components in Gaussian mixture model may vary considerably across different words. For example, there could be 3 museums but 10 parks in a city. Second and more importantly, the true distribution of a word may not necessarily follow a mixture of Gaussian. In fact, the social media data generated by the crowd are constrained by the underlying city maps and natural landscapes, such as road networks, downtowns, lakes, and mountains, thus may deviate significantly from a Gaussian distribution.

We propose to use kernel density estimation (KDE) to model the distribution of words. KDE has been applied to location-based social networks [36], check-in data [16], human mobility [20], epidemiology [5], ecology [18], and marketing [12]. In particular, [36, 20] have demonstrated the advantages of KDE over Gaussian model under their problem settings. We propose to adapt KDE, aiming to capture both the locality of a word distribution and the relevance of a word to a given location record r_i. In the following section, we provide a detailed discussion of our proposed methods.

4. SEMANTIC ANNOTATION METHODS

The essence of the annotation problem boils down to measuring the *locality* of a word with respect to the location of a record r_i. The locality can be represented by a local density measure.

The annotation should be time-sensitive, as the same location may hold different events at different time. Therefore, for a location record $r_i = (loc_i^U, t_i^U)$, we consider the documents within a time window τ of t_i, i.e., $[t_i - \tau, t_i + \tau]$. We define the set of documents that fall into this time window as $\mathcal{D}_i = \{d_j | t_i - \tau \leq t_j^D \leq t_i + \tau\}$. Accordingly, the word set for documents in this time window is defined as $\mathcal{W}_i = \bigcup_{d_j \in \mathcal{D}_i} W_j$. For each $w \in \mathcal{W}_i$, we further let $L_i(w)$ be the set of locations of all tweets in $\mathcal{D}_i$ which contain the word w: $L_i(w) = \{loc_j^D | w \in W_j, d_j \in \mathcal{D}_i\}$. In this section, we only discuss annotation problem within one time window.

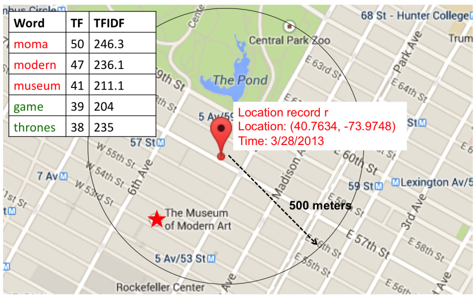
4.1 Frequency Based Method

One simple but intuitive measure is to count the occurrence of word w near the given location record. More specifically, given a mobility record r_i, the score $s_w^F(r_i)$ is defined as:

$$s_w^F(r_i; L_i(w), \delta) = |\{loc_j^D \in L_i(w) : dist(loc_i^U, loc_j^D) < \delta\}|,$$

where δ is a spatial threshold and $dist()$ is a distance function between two geographic points.

It is clear that the frequency measure can not find words that are specific to the locations. To alleviate this problem, we weight the word frequency by the inverted document frequency. The resulting relevance score inherits the same idea as the *tf-idf* score which is widely used in information retrieval for evaluating word importance over a collection of documents. Mathematically, given a mobility record r_i, the



Word	TF	TFIDF
moma	50	246.3
modern	47	236.1
museum	41	211.1
game	39	204
thrones	38	235

Figure 2: An example illustrating the problem of frequency based methods. The true user's intention of this location record is to attend the Game of Thrones event. But since MOMA is a more popular venue nearby, frequency-based methods will incorrectly use words "moma" and "modern" for annotation.

score $s_w^{tf-idf}(r_i)$ is defined as follows:

$$s_w^{tf-idf}(r_i; L_i(w), \delta) = s_w^F(r_i;\delta) \cdot \log \frac{|\mathcal{D}|}{|\{d_j \in \mathcal{D} : w \in W_j\}|}.$$

However, there are two problems with the above measure. First, it does not consider the *distance* from the user's location record to the center of the word. As a result, it favors nearby popular events, i.e., those events located within the spatial proximity. We illustrate this problem of $s^F(r_i;\delta)$ and $s^{tf-idf}(r_i;\delta)$ in Figure 2 with a real example. In this example, the user was attending the Game of Thrones (a TV series) exhibition. The original tweet is: "with @jedafrank (@ Game of Thrones Exhibition w/ 14 others) http://t.co/U5ztm1RWRu". The event location is very close to the Museum of Modern Art (MOMA). If we look at all the nearby tweets by setting the distance threshold $\delta = 500$ (meters), the highest ranked words are "moma" and "modern", since MOMA is a very popular location.

Second, the frequency based measure needs a threshold δ, which is not easy to set. Choosing a large threshold δ may include irrelevant words, whereas choosing a small threshold may result in a loss of useful information.

4.2 Gaussian Model

Using word probability addresses the problem of a hard threshold. The word probability at one location can be obtained from a two-dimensional Gaussian model fitted using the locations of all the occurrences of word w (i.e., $L_i(w)$). Therefore, we can define the relevance score as follows:

$$\begin{aligned} s_w^G(r_i; L_i(w)) &= f^G(r_i; \mu_w, \Sigma_w) \\ &= \frac{1}{2\pi|\Sigma_w|^{\frac{1}{2}}} e^{-\frac{1}{2}(loc_i^U - \mu_w)'\Sigma_w^{-1}(loc_i^U - \mu_w)}, \end{aligned}$$

where μ_w is a two-dimensional mean vector, and Σ_w is a 2×2 covariance matrix. In practice, a finite mixture of C Gaussian densities are often used for modeling multimodal distributions. The Gaussian mixture model (GMM) has been previously applied to model human mobility [10], as well as served as the underlying generative model to detect spatially related words [35]. Using the Gaussian mixture model, we can define the relevance score as follows:

$$s_w^{GMM}(r_i; L_i(w)) = \sum_{k=1}^{C} \pi_w^k f^G(r_i; \mu_w^k, \Sigma_w^k),$$

where μ_w^k, Σ_w^k, and π_w^k are the mean, covariance matrix, and weight of the k-th component, respectively, and $\sum_{k=1}^{C} \pi_w^k = 1$.

The mixture model still has several drawbacks. First, the assumption that the underlying probability is a mixture of Gaussians may not hold. Figure 3(a) show 100 tweet locations in NYC mentioning word "museum". It is easy to see the density of the word occurrences exhibits complex variations on the map . We use Gaussian mixture model to fit the points by setting component number $C = 2$. As shown in Figure 3(b), the density contours estimated according to the Gaussian mixture model tend to be over-smoothed. Second, even if the underlying distribution of a word follows the Gaussian mixture model, choosing the number of components is non-trivial. More importantly, the number of components for each word may vary significantly. Manually setting number of components for each word is not feasible.

4.3 Kernel Density Estimation

From the above discussion, we conclude that a good annotation framework should satisfy two main properties. First, it should model the effect of *distance*. Second, a good word density estimation at a location should only depend on the data points that are *local* to the location. With these properties in mind, we propose to use Kernel Density Estimation (KDE) methods to model the spatial density of the word occurrences.

KDE is a non-parametric model for estimating density from sample points. Following the kernel density model, we define the relevance score for a word w at r_i as follows:

$$s_w^{KDE}(r_i; L_i(w), H) = \frac{1}{|L_i(w)|} \sum_{loc_j^D \in L_i(w)} K_H(loc_i^U - loc_j^D).$$

Here, we define

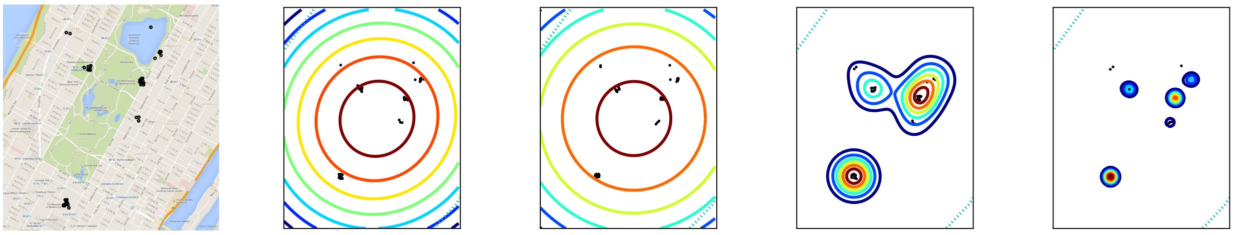
$$K_H(x) = |H|^{-1/2} K(H^{-1/2}(x)), \qquad (1)$$

where $K(x)$ is a kernel function of choice and H is a bandwidth matrix. The bandwidth matrix H has a strong influence on the estimated density. For our problem, we assume that the dimensions of the geographic coordinate are independent of each other. Further, we treat the two dimensions equally and use the same bandwidth h for both dimensions, i.e., $H = hI$ where I is the identity matrix. It is obvious that the estimated density will be sharply peaked around the sample points when h is small and overly smoothed when h is large.

In practice, the method is known to be less sensitive to the choice of kernel function than the bandwidth matrix H [11]. Further, it is desirable that a word occurrence contributes equally to all locations that are at the same distance from it. Therefore, we let $K(x)$ be the Gaussian kernel with matrix I as the covariance matrix. We can re-write Eq. (1) as:

$$K_H(x) = \frac{1}{2\pi|H|^{-\frac{1}{2}}} e^{-\frac{1}{2}x'H^{-1}x} = \frac{1}{2\pi h} e^{-\frac{1}{2h}x'x}.$$

Indeed, the KDE method satisfies the properties we have for a good annotation framework. First, the kernel density



(a) Locations on map (b) GMM ($C = 2$) (c) KDE ($h = 0.01$) (d) KDE ($h = 0.003$) (e) KDE ($h = 0.001$)

Figure 3: Example of different models on word distribution for word "museum".

function $K_H(loc_i^U - loc_j^D)$ quantifies how much each document $d_j \in \mathcal{D}_i$ contributes to r_i. It is easy to see that the function decays as a function of distance between r_i and d_j. Second, the parameter h controls the range of effect one sample point has on the geographic space. By tuning h, we can make sure that each sample only contributes to its nearby locations. Next we discuss how to determine the bandwidth parameter h for our annotation problem.

4.3.1 Determining the Bandwidth h

It is well known that the resulting density estimate can be highly sensitive to the value of the bandwidth h, producing densities sharply peaked when h is too small, and producing an overly smooth estimate when h is too large. Figure 3(c)-(d) show the density maps with different values of h.

In general, there are two classes of approaches to selecting the bandwidth parameter h: reference rule approach and data-driven approach. While these approaches typically aim to find a fixed h, it is possible to estimate an adaptive bandwidth h which depends on the density at each sample point. Below we introduce these approaches in detail.

Reference rule approach. The *reference rule* approach derives h from assumptions of the underlying distribution. One commonly used reference rule, namely, *the Scott's rule of thumb*, is derived from the assumption that the underlying density being estimated is Gaussian. For location data, we assume that the longitude and latitude are independent. The Scott's rule can be written as:

$$H_w = n^{-2/(d+4)}\Sigma_w,$$

where Σ_w is the variance matrix of location of all the occurrences of word w (i.e., $L_i(w)$) and d is the dimension ($d = 2$ for our problem). The estimate is optimal (i.e., minimizes the mean integrated squared error) if the true underlying density is Gaussian. However, as we pointed out earlier, this assumption may not necessaries be true at a fine-granularity in our problem.

Cross validation approach. From a data-driven perspective, cross validation can be applied to choose h_w which fits the data best. In general, cross validation fits the model to a part of the data (i.e., training data), and then evaluates the goodness of the model on the rest of the data (i.e., testing data). The model that has the best performance on the testing data are picked based on our evaluation metric. We use log-likelihood as the metric to choose a proper h_w. The log-likelihood is given by:

$$\mathcal{L}(h_w) = \frac{1}{|L_i^{test}(w)|} \sum_{l \in L_i^{test}(w)} \log s_w^{KDE}(l; L_i^{train}(w), h_w),$$

where $L_i^{train}(w)$ and $L_i^{test}(w)$ are partitions of $L_i(w)$. Larger value of $\mathcal{L}(h_w)$ indicates a better goodness of fit.

KDE with adaptive bandwidth. The bandwidth parameter h_w picked via aforementioned approaches is fixed. Another variation of KDE is to use an adaptive bandwidth, where h_w depends on each sample point. Breiman et al. [6] suggested adapting h_w to each sample point $loc_j^D \in L_i(w)$, and set it to be the distance between loc_j^D and its k-th nearest neighbor, where the optimal value of k can be determined via cross validation.

5. EXPERIMENT

In this section, we conduct both quantitative evaluations and case studies to verify the effectiveness of the proposed method on real datasets.

5.1 Datasets

We use three datasets of geo-tagged tweets from three major cities in U.S.A, i.e., New York City, Chicago, and Los Angeles. The statistics of the datasets is summarized in Table 1. Each tweet is of the form $\langle timestamp, userid, latitude, longitude, content\rangle$. We split the spatiotemporal documents by each day, as period for most events spans a day. In addition, events span multiple days can also be observed within each day.

City	#tweets	Time range
New York City (NYC)	15,612,712	11/2012-7/2013
Chicago (CHI)	11,269,220	10/2011-7/2013
Los Angeles (LA)	10,989,333	11/2012-7/2013

Table 1: Statistics of datasets.

To generate the other input of our method, i.e., the **location history** of a user, we gather all the pairs of GPS coordinate (longitude and latitude) and timestamp from the geotagged tweets of the user. We consider those check-in points as stop point. In practice, we may also use different sources to obtain the location history data, such as mobile services or GPS devices. For such densely sampled raw trajectory, a stopping point detection method can be applied first [38]. In this paper, we use the location history extracted from the geo-tagged tweets because the content of the corresponding

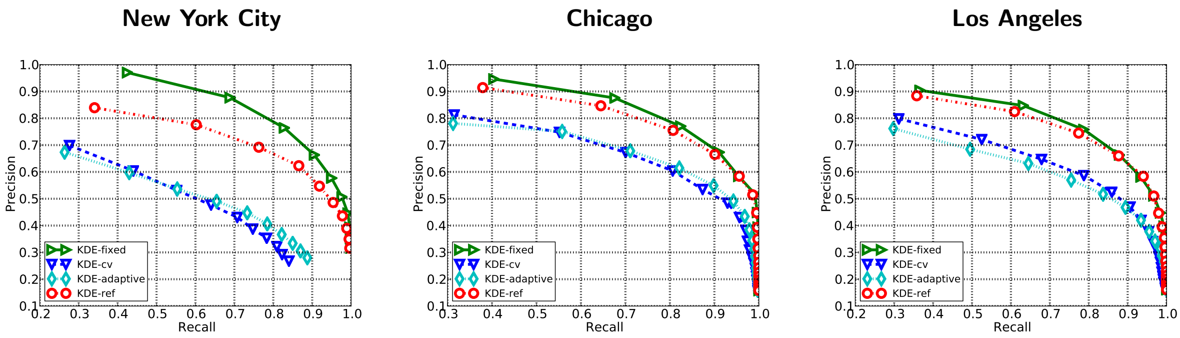


Figure 4: Precision-recall curves of different methods for choosing bandwidth h.

tweets provides a means for evaluation. We emphasize that *these tweets are excluded from the external contextual documents and are used for evaluation purpose only.*

5.2 Evaluation Method

Ground truth annotation dataset. Given a location record r of a user, the ground truth annotation document should consist of the relevant words that describe the true intentions of this user visiting location r. Since most of the tweets are not location-specific, our first step toward building a ground truth annotation dataset is to identify the location records that are relevant to some local events. To this end, we use the tweets that are posted through Foursquare check-ins. Foursquare is a location-based social network, where users can check-in at venues and events. When a user checks-in at a venue/event, it automatically generates a post containing the name of the venue/event along with a clause indicating the number of users that check-in at the same time, e.g., " w/ 400 others". Below we show an exemplar check-in tweet:

> "Daddy's home!!!! LETS GO RANGERS (@ Madison Square Garden for Pittsburgh Penguins vs New York Rangers w/ 90 others)"

We gather check-in tweets that have more than 50 users checking-in at the same time, as those check-ins are more likely to contain event information. We manually filter the irrelevant words in each tweet and use the remaining words to construct the ground truth annotation document. For example, for the exemplar tweet above, the relevant words are: { "rangers", "madison", "square", "garden", "pittsburgh", "penguins"}.

For experiments in this paper, we have prepared 1,540 ground truth annotation documents for New York City, 697 for Chicago, and 623 for Los Angeles.

Evaluation metric. We use precision and recall as the metric to evaluate our annotation methods. The precision and recall are commonly used in information retrieval systems to evaluate the quality of the search result. Given a location record r with ground truth annotation document $\mathcal{G}(r) = \{w_{g1}, w_{g2}, ..., w_{gt}\}$, and an annotation document $\mathcal{A}(r)$ obtained by a method under evaluation, the precision and recall are defined as follows:

$$P(\mathcal{A}(r)) = \frac{|\mathcal{G}(r) \bigcap \mathcal{A}(r)|}{|\mathcal{A}(r)|},$$

$$R(\mathcal{A}(r)) = \frac{|\mathcal{G}(r) \bigcap \mathcal{A}(r)|}{|\mathcal{G}(r)|}.$$

Since each method returns a ranked word list, we only keep top-k words in the list as the annotation document $\mathcal{A}(r)$. By varying k, we can plot a precision-recall curve for each method.

5.3 Determining Bandwidth h

In this section, we study different approaches to determining the bandwidth h in the KDE model. Note that, the choice of h relates to the granularity of the annotation. For our task, we aim to find a small h that gives us annotations at a fine granularity. For applications that need more coarse-level annotations (e.g., annotation by names of the cities), a larger h may compensate for the data sparsity issue and provide more robust result.

In addition to the three approaches introduced in Section 4.3.1, namely the reference rule (KDE-ref), cross validation (KDE-cv), and adaptive bandwidth (KDE-adaptive), we also consider a fourth option which uses a fixed h for all the words (KDE-fixed).

For KDE-adaptive and KDE-cv, we split the datasets as 70% for training and 30% for testing. We select the parameter h_w for each word that yields the highest log-likelihood on the test data. For KDE-fixed, we empirically choose $h = 10^{-4}$ as it performs the best in practice among a set of different values. Here we emphasize that for KDE-fixed, all the words share the same bandwidth h, whereas for all the other methods we compute a h_w for each word.

Figure 4 shows the precision-recall curves for all methods. It is easy to see that for all three cities, KDE-ref and KDE-data perform much worse than KDE-fixed and KDE-ref. One possible explanation is that the bandwidth h_w obtained by maximizing the log-likelihood $\mathcal{L}(h_w)$ is often unreliable when the samples are sparse [11, 20], which is indeed the case for many words in our dataset. In such cases, the estimated h_w tends to take a large value, resulting in an over-smoothed density estimate.

In addition, we have observed that the value of bandwidth h_w estimated by KDE-ref takes value around 10^{-4} for most of the words. This explains why KDE-ref and KDE-fixed have very similar results. This also suggests that it is not necessary to estimate h_w for each word separately for our problem. Thus we simply set $h = 10^{-4}$ for the remaining experiments.

5.4 Compare with Other Annotation Methods

In this section, we compare the performance of our proposed measures s^{KDE}, s^F, s^{tf-idf}, and s^{GMM}. We name those measures, KDE, FREQ, TFIDF, and GMM, respectively.

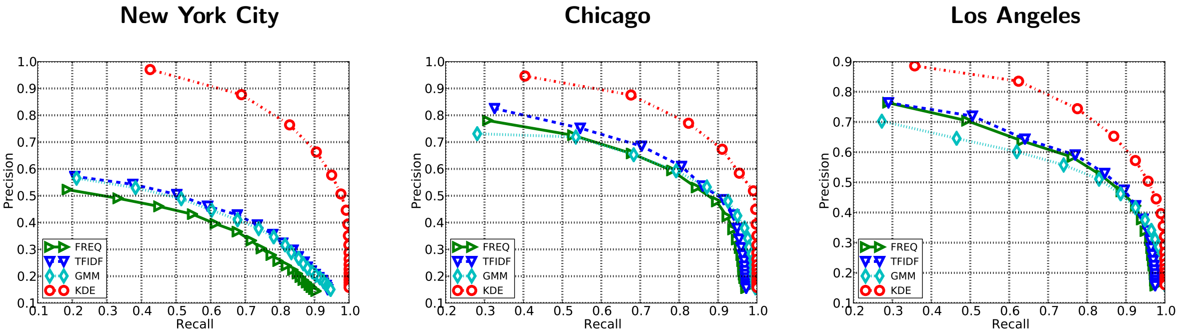
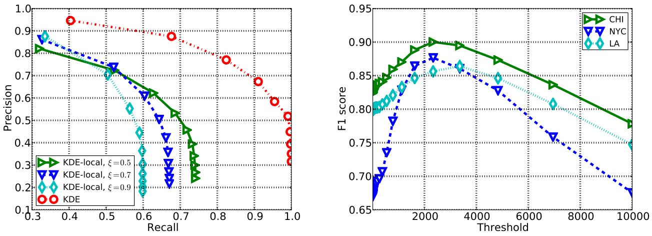


Figure 5: Comparison of Precision-recall curves for methods FREQ,TFIDF,GMM and KDE on three datasets.



(a)

(b)

Figure 6: (a) Comparison of Precision-recall curves of using threshold to filtering non-local words first on the Chicago dataset. (b) F_1 score w.r.t. to threshold.

We set $\delta = 1$ (kilometer) for FREQ and TFIDF, and $C = 5$ for GMM, as those parameters return the best result. We run this experiment on all three datasets.

Figure 5 summarizes the performance of four methods. We can see that KDE clearly outperforms other three methods. Especially for NYC dataset, KDE has a precision close to 0.95, when maintain a recall of 0.4. Meanwhile, all other methods only have precision value less than 0.6 under the same recall value. For CHI and LA datasets, KDE consistently has better precision over the other three methods. The KDE method has better performance on NYC dataset, as most large events are located at Manhattan area of New York City. There are more irrelevant tweets near places holding events. As a result, FREQ, TFIDF, and GMM perform poorly, as the measures they used are not local enough. In CHI and LA datasets, many events are held at less populated areas. FREQ, TFIDF, and GMM are less affected by irrelevant tweets.

5.5 Comparison with Filtering non-local Words

As we discussed earlier, local word detection method can be applied as a pre-processing step to filter non-local word first. However, for our annotation problem this pre-processing step can be *erroneous*. In this section, we compare KDE with KDE using local word detection method.

To model locality of words (query), Backstrom et al. [4] propose a model based on the intuition that a local word should have a high local focus. In addition, the frequency of this word should drop rapidly as the distance to the center increases, reflecting a strong association to the center location. The model uses a fast decaying function in the form of $Cd^{-\alpha}$ to capture the intuition, where d is the distance to its center. The center corresponds to the location where the word appears most frequently. The parameter is determined using the following likelihood function:

$$f(C,\alpha) = \sum_{l_i \in L(w)} \log C d_{l_i}^{-\alpha} + \sum_{l_i \notin L(w)} \log(1 - C d_{l_i}^{-\alpha}).$$

The C captures the center frequency and α measures the speed of decaying. We use the center of the most frequent grid as the word center and follow the center finding step as suggested by [9]. Then, a grid search is used to determine C and α that maximize the likelihood function. We set a threshold ξ on α to filter non-local words.

Figure 6(a) shows the precision-recall curve as we vary ξ from 0.5 to 0.9. It is clear that as the threshold ξ increases the performance decreases. The result indicates that many words describing the user intentions have a low α. For example, word "knicks" (name of a professional basket ball team) will be frequently mentioned at the arena where the game is held at a game day. However, many users may watch the game at some local bars and tweet about the event. The word density will have multiple peaks. Therefore, using a single modal model to fit the occurrences will result in a small α. For our annotation problem, words occurrences often express such multiple modality at a fine-granularity.

5.6 Threshold Study

In real applications, we can set a threshold parameter θ for KDE score to pick relevant words for annotation. We want to study what is the optimal threshold and whether this optimal value is consistent over different datasets. For a given threshold, we calculate the average F_1 score for the location records: $F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. Figure 6(b) shows the F_1 score on three datasets. For CHI, NYC, and LA datasets, KDE method achieves the best performance at $(F_1 = 0.90, \theta = 2300)$, $(F_1 = 0.87, \theta = 2300)$, and $(F_1 = 0.90, \theta = 3300)$, respectively. When the threshold is in the range of $[2000, 3000]$, the F_1 score is greater than 0.85 on the three datasets. The results suggest that a reasonable range for selecting θ should fall in the range of $[2000, 3000]$. We set $\theta = 2000$ for the following case studies.

5.7 Case Studies

In this section, we perform several case studies over three different cities to demonstrate the effectiveness of our method in generating time- and space-sensitive semantic *annotations* for the mobile users' location records. For comparison, we

(a) Annotation documents using tweets (our method)

r_1: (40.7505, -73.9934), 1/20/13

Word	Relevance Score
madison	99778
garden	95544
square	76263
penguins	49841
rangers	48560

r_2: (40.6831, -73.9760), 2/10/13

Word	Relevance Score
barclays	86572
center	64970
spurs	40781
nets	34046
antonio	92677

r_3: (40.7505, -73.9934), 3/24/13

Word	Relevance Score
madison	57783
garden	57440
square	45425
penn	25696
rangers	10096

r_4: (40.7634, -73.9748), 3/28/13

Word	Relevance Score
exhibition	26179
thrones	26102
game	23634
day	1002
others	978

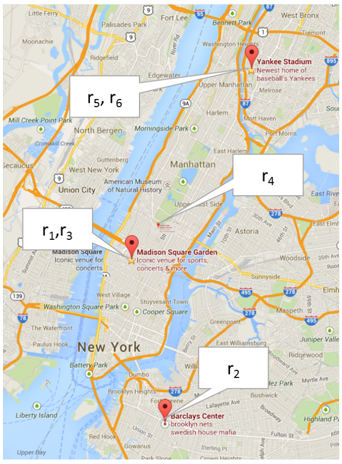
r_5: (40.8295, -73.9270), 4/18/13

Word	Relevance Score
yankee	104122
stadium	104087
diamondbacks	42027
yankees	40528
arizona	33577

r_6: (40.8295, -73.9269), 6/5/13

Word	Relevance Score
yankee	96293
stadium	91471
yankees	28911
indians	27710
cleveland	24590

(b) Locations on the map


r5, r6
Yankee Stadium
Newest home of baseball's Yankees
Manhattan
r4
r1,r3
Madison Square Garden
Iconic venue for sports, concerts & more
New York
r2
Barclays Center
brooklyn nets - swedish house mafia

(c) Annotation by static POI information (method for comparison)

r_1 , r_3: (40.7505, -73.9934)

POI	Distance
Madison Square Garden	3
Auntie Anne's Pretzels	9
Forest Electric Corporation	16
Auntie Anne's Pretzels	21
Event Bar	31

r_2: (40.6831, -73.9760)

POI	Distance
Barclays Center	72
Burlington Coat Factory	79
Jean-Baptiste Gary E MD	84
National Vision	98
Wellness Center	99

r_5 , r_6: (40.8259, -73.9269)

POI	Distance
New York Yankees	34
Yankees Audi Club	37
Johnny Rockets	64
Yankee Stadium	78
Apple Bank For Saving	129

r_4: (40.7634, -73.9748)

POI	Distance
Manhattan Dental Health	9
Och-Ziff Capital Management Group	23
Oz Management Group	23
Summit Rock Advisors	23
Ingres Corporation	23

(d) Tweets as ground truth:

r_1: Daddy's home!!!! LETS GO RANGERS (@ Madison Square Garden for Pittsburgh Penguins vs New York Rangers) [pic]: http://t.co/cUYIVcRx
r_2: I'm at Barclays Center for San Antonio Spurs vs Brooklyn Nets (Brooklyn, NY) w/ 69 others http://t.co/UgJxCvlX
r_3: Let's do it. (@ Madison Square Garden - @thegarden for @washcaps vs @NYRangers w/ 44 others) http://t.co/oaaE4f1t9n
r_4: with @jedafrank (@ Game Of Thrones Exhibition w/ 14 others) http://t.co/U5ztm1RWRu
r_5: I'm at Yankee Stadium - @mlb for Arizona Diamondbacks vs New York Yankees (Bronx, NY) w/ 129 others [pic]: http://t.co/MnHVcn170P
r_6: I'm at Yankee Stadium - @mlb for Cleveland Indians vs New York Yankees (Bronx, NY) w/ 143 others http://t.co/KTidl9wtEv

Figure 7: Case study for user #901.

also show annotation results using nearby POI information. The POI information is gathered via Google Places API[3].

Case User #901. In this case, we examine the location history of user #901, who lives in Staten Island (based on his user profile) and visits New York City frequently. We pick six location records from this user. Figure 7(a) shows the top-5 words extracted by our method for each record.

For the location records $\mathbf{r_5}$ and $\mathbf{r_6}$, our method ranks the words "yankee", "stadium" and "yankees" as the top three words with the relevance scores higher than 90. Figure 7(d) shows tweets from this user at those location records. We can see that this user was indeed watching Yankees' games on these two days, as he tweeted about the games. In addition, the location information of these two records on the map (Figure 7(b)) shows that the user was physically present at the Yankee Stadium, instead of watching the games at home.

Similarly, for location records $\mathbf{r_1}$ and $\mathbf{r_3}$, we can infer from the tweets in Figure 7(d) and locations in Figure 7(b) that the user was attending Rangers' hockey game at Madison Square Garden. Our method correctly ranks the words "madison", "garden", "square" and "rangers" at the top of the list. Further, word "penguin" from $\mathbf{r_1}$ corresponds to the team that Rangers was competing with on that day. We also note that, in overall, these words have lower scores than those related to the Yankees, because Rangers' games are relatively smaller events compared to Yankees' games.[4]

In addition, we find that this user went to see the NBA game between San Antonio Spurs and Brooklyn Nets on 02/10/2013 at Barclays Center ($\mathbf{r_2}$) and the Game of Thrones exhibition on 3/28/2013 ($\mathbf{r_4}$).

Figure 7(c) shows the annotation by static POI information. For records $\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}, \mathbf{r_5}, \mathbf{r_6}$, the static annotation correctly identifies location names. However, such annotation does not include dynamic event information, such as "rangers", "penguin", "spurs" and "nets" in $\mathbf{r_1}$ and $\mathbf{r_2}$.

Case User #115. Now we look at another user #115, who lives in New York City. We also pick six location records from this user and show the corresponding annotation documents in Figure 8(a). This user is also a fan of the Yankees, as indicated by record $\mathbf{r_4}$. For record $\mathbf{r_2}$ on 5/6/2013, our method ranks the words "art", "metropolitan" and "museum" as the top-3 words followed by the words "metgala" and "met". As the first three words indicate that this user was at the Metropolitan Museum of Art, the words "gala" and "metgala" further reveal the specific event this user was attending. As confirmed by the corresponding tweets in Figure 8(d) and the actual locations in Figure 8(b), this user indeed attended the Met Gala event, an annual affair to celebrate the opening of the Metropolitan Museum's fashion exhibit. Meanwhile, annotation by POI information cannot reveal such rich contextual information. As shown in Figure 8(c), the static POI annotation only identifies road and landmark names around the location. For record $\mathbf{r_5}$, our method detects a soccer game with "field", "citi", "israel" and "honduras" as the most relevant words. At the same time, annotation by POI information does not contain this event information as shown in Figure 8(c). Other than these big events, annotation using our method indicates the user's visit to Museum of Modern Art (MOMA) in $\mathbf{r_1}$ and an ice-skating event near Central Park in $\mathbf{r_3}$.

Case User #329. Finally, we study user #329 from Chicago. Similar results are observed and reported in Figure 9. The actual tweets and our annotations in Figure 9(a) reveal three activities of the user. $\mathbf{r_1}, \mathbf{r_3}$ records two different marathon that user #329 participated in. Although the en route check-ins do not have a specific POI, our method can associate the nearby tweets at that time and location and identify the "marathon" as a top word in annotation. Records $\mathbf{r_2}, \mathbf{r_6}$ are baseball matches of Chicago White Sox. The "United Center" and "hawks" in the annotations of $\mathbf{r_4}, \mathbf{r_5}$ reveal the

[3] https://developers.google.com/places/
[4] Yankee Stadium has a capacity of 50,291, whereas Madison Square Garden has a capacity of 18,200.

(a) Annotation documents using tweets (our method)

r_1: (40.7583, -73.9854), 5/5/13

Word	Relevance Score
modern	20607
art	16814
museum	15975
moma	3272

r_2: (40.778, -73.962), 5/6/13

Word	Relevance Score
metropolitan	54272
museum	46360
art	44129
punk	10156
metgala	7279

r_3: (40.7684, -73.9745), 5/13/13

Word	Relevance Score
wollman	12690
skating	10909
rink	10405
ice	10123
fox	4497

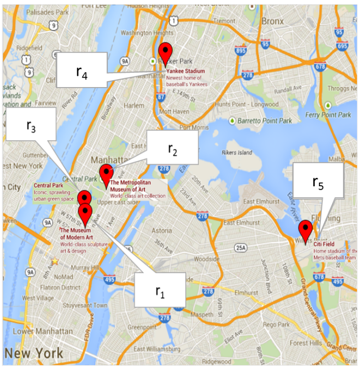
r_4: (40.8294, -73.9269), 5/15/13

Word	Relevance Score
yankee	102144
stadium	98309
game	16975
yankees	14301

r_5: (40.7564, -73.8460), 6/5/13

Word	Relevance Score
citi	40914
field	40416
honduras	13359
israel	12631

(b) Locations on the map


r4
r3
r2
r5
r1
Yankee Stadium
Newest home of baseball's Yankees
The Metropolitan Museum of Art
World-class art collection
Central Park
Iconic, sprawling urban green space
The Museum of Modern Art
World-class sculpture, art & design
Citi Field
Home stadium of the Mets baseball team
Bronx
Manhattan
Harlem
Washington Heights
Astoria
Woodside
Elmhurst
East Elmhurst
Rikers Island
Flushing
New York
Lower Manhattan

(c) Annotation by static POI information (method for comparison)

r_1 : (40.7583, -73.9854)

POI	Distance
Forever 21	26
Disney Store	26
Hodgson Russ LLP	30
00 Commercial Blinds	33
New York Society of Security analysts	35

r_2: (40.778, -73.962)

POI	Distance
995 Fifth Avenue	56
Frank E. Campbell – The Funeral Chapel	56
Thomas W. Loeb, MD	59
Olive & Bettes	70
Universal Funeral Chapel	70

r_3: (40.7684, -73.9745)

POI	Distance
Victorian Garden Amusement Park	66
Wollman Ice Skating Park	70
Central Park Carousel	174
Central Park Zoo	174
The Arsenal	282

r_5: (40.7564, -73.8460)

POI	Distance
Citi Field	9
Shea Stadium Home Run Apple	138
Mama's of Corona	151
Mets Plaza	170
NYC Parks & Recreation	183

*r_4: (40.8295, -73.9269): Annotation by POI is the same as r_5, r_6 in Figure 8.

(d) Tweets as ground truth:

r_1: katyperry @ Museum of Modern Art (MoMA) http://t.co/j9ql2faeRw
r_2: .emmyrossum #MetGala #PunkFashion @ The Metropolitan Museum of Art http://t.co/T11YQfm6KG
r_3: @minkakelly #foxupfront @ Wollman Park http://t.co/56dcTzx5YJA
r_4: @bearpascoee @4stillrunning @ruebenrandle #bleachercreatures #rollcall yankees #204 @ Yankee Staduim http://t.co/PIGxyatsF9
r_5: Houduras vs. Israel http://t.co/kUxOy8hozw

Figure 8: Case study for user #115.

fact that the United Center is the home arena of ice hockey team Chicago Blackhawks. In Figure 9(b), we show the actual POIs of each tweets, which are consistent with out annotation.

In summary, the above case studies show that our method can correctly annotate the semantics (including landmarks and events) to a user's location records. Further, our clustering results on the annotation documents well represent the user's interests in local events.

6. PROFILING USER AS AN APPLICATION

Given the annotation document $\mathcal{A}(r_i)$ for each location record r_i of a user U, we can discover the interests of the user by examining the similarities of all the annotation documents. For example, a New York City sports fan may have several annotation documents related to the sport events in New York City. Therefore, assuming there are K underlying interests for this user, the annotation documents should be partitioned into K groups so that each cluster represents one interest.

6.1 Profiling Method

A potential solution is to apply some clustering algorithms. In order to do this, a distance measure between two documents must be specified. In the literature, the cosine similarity is frequently used to measure the document similarity due to its length invariance [28]. However, in our problem, the absolute scores in a document are important in differentiating the event days and the normal days. For example, suppose we have annotation documents on two location records: $\mathcal{A}(r_1) = \{(nets, 150), (nba, 100), (others, 1)\}$ and $\mathcal{A}(r_2) = \{(nets, 10), (nba, 8), (others, 5)\}$, where record r_1 is on a game day. The cosine similarity will then consider these two documents as very similar: $Cos(\mathcal{A}(r_1), \mathcal{A}(r_2)) = 0.93$. Therefore, we propose to use the Euclidean distance as our similarity measure, which better preserves the original relevance scores. Indeed, the Euclidean distance between the two documents in the previous example is large: $Euc(\mathcal{A}(r_1), \mathcal{A}(r_2)) = 173.26$.

Various clustering methods, such as K-means and hierarchical clustering, can be applied to cluster the annotation documents based on the proposed similarity measure. In this paper, we simply use the K-means method.

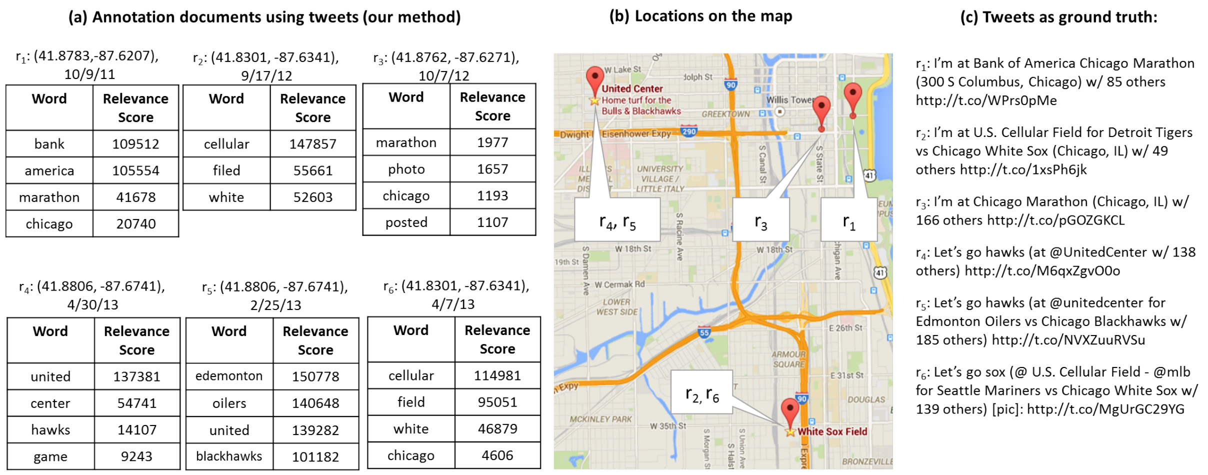
6.2 Case Studies

We run the user profiling method on user #901 and user #115 in Section 5.7. We use all the location records of a user and cluster all the annotation documents to profile the user's interests. There are 33 location records of user #901 and 250 records of user #115.

Case User #901. In Figure 10, we report the top-5 words (based on the sum of relevance scores) for each document cluster and the ground truth profile of this user. K is set as 5 in the K-means clustering algorithm. We can see that Cluster 1 and Cluster 4 both contain annotation documents that are related to Yankees' games, and the words "stadium", "yankee" and "bronx" also appear in the user's reference interest profile. Taking a closer look at Cluster 4, we see that it is about a special event, namely the Old-Timers' Day. This is a popular event held annually to celebrate the accomplishments of Yankees' former players. Our method is able to differentiate it from Yankees' regular games (Cluster 1). In addition, Cluster 2 corresponds to the user's interest in Rangers. The annotation document of record $\mathbf{r_2}$ forms a cluster (Cluster 3) by itself, as this NBA basketball game differs from other sports games that this user attended. Finally, we note that Cluster 5 has the largest number of documents (23 out of 33 total documents). This is because, in practice, most tweets are not related to specific events. The annotation documents for those records tend to include arbitrary words with low relevance scores and typically form one cluster.

Figure 10(b) shows the words from this user's tweets ranked by tf-idf score. Note that our annotation method only uses the tweets from the crowd (excluding the tweet from this target user). The sports related terms are frequently seen in this person's tweets, indicating he is a sports fan. Words describing routine behaviors of a user also have high tf-idf scores, such as "staten" and "island" (home of this user).

Case User #115. We use K-means with $K = 7$ in this case and report the top-5 words of each cluster in Figure 11. We only show 3 clusters in the figure as we exclude the clusters of background words. In Figure 11(a), we can see



(a) Annotation documents using tweets (our method)

r_1: (41.8783,-87.6207), 10/9/11

Word	Relevance Score
bank	109512
america	105554
marathon	41678
chicago	20740

r_2: (41.8301, -87.6341), 9/17/12

Word	Relevance Score
cellular	147857
filed	55661
white	52603

r_3: (41.8762, -87.6271), 10/7/12

Word	Relevance Score
marathon	1977
photo	1657
chicago	1193
posted	1107

r_4: (41.8806, -87.6741), 4/30/13

Word	Relevance Score
united	137381
center	54741
hawks	14107
game	9243

r_5: (41.8806, -87.6741), 2/25/13

Word	Relevance Score
edemonton	150778
oilers	140648
united	139282
blackhawks	101182

r_6: (41.8301, -87.6341), 4/7/13

Word	Relevance Score
cellular	114981
field	95051
white	46879
chicago	4606

(b) Locations on the map

(c) Tweets as ground truth:

r_1: I'm at Bank of America Chicago Marathon (300 S Columbus, Chicago) w/ 85 others http://t.co/WPrs0pMe

r_2: I'm at U.S. Cellular Field for Detroit Tigers vs Chicago White Sox (Chicago, IL) w/ 49 others http://t.co/1xsPh6jk

r_3: I'm at Chicago Marathon (Chicago, IL) w/ 166 others http://t.co/pGOZGKCL

r_4: Let's go hawks (at @UnitedCenter w/ 138 others) http://t.co/M6qxZgvO0o

r_5: Let's go hawks (at @unitedcenter for Edmonton Oilers vs Chicago Blackhawks w/ 185 others) http://t.co/NVXZuuRVSu

r_6: Let's go sox (@ U.S. Cellular Field - @mlb for Seattle Mariners vs Chicago White Sox w/ 139 others) [pic]: http://t.co/MgUrGC29YG

Figure 9: Case study for user #329.

Cluster	Top-5 words
1	yankee, stadium, yankees, bronx.
2	garden, madison, square, rangers, penguins.
3	barclays, center, brooklyn, nets, spurs.
4	yankee, stadium, yankees, old, oldtimersday.
5	public, island, plaza, staten, drinking.

(a)

Word	tf-idf
staten	13.43
stadium	13.08
rpx	12.22
hylan	11.18
island	11.16
drinking	7.57
rangers	7.44
yankee	6.32
bronx	5.58
plaza	5.22
madison	4.64
garden	4.81
photo	4.42
ale	4.15

(b)

Figure 10: Interest profiles for user #901. (a) Top words in each cluster of the annotation documents. (b) Words from tweets ranked by tf-idf scores.

Cluster	Top-5 words
1	art, metropolitan, museum, metgala, punkfashion.
2	yankee, mets, sox, stadium, orioles.
3	citi, field, mets, edcny, subwayseries.

(a)

Word	tf-idf
metgala	42.09
punkfashion	42.09
yankee	33.87
stadium	31
bagatelle	28.30
citi	27.19
field	25.41
metropolitan	24.57
marquee	22.73
bleacher	18.83
creatures	15.95
superstudio	15.68
art	14.37
museum	14.10

(b)

Figure 11: Interest profiles for user #115. (a) Top words in each cluster of the annotation documents. (b) Words from tweets ranked by tf-idf scores.

that Cluster 1 corresponds to this user's interest in the Met Gala event and Cluster 2 corresponds to this user's interest in Yankees' game. Meanwhile, Cluster 3 summarizes terms that are related to events held at the Citi Field, including the Electric Daisy Carnival festival and the baseball games. Our results are consistent with this user's interests inferred from his tweets as shown in Figure 11(b).

These two case studies show that our interest profiling using annotation documents can reveal the real interests of this user inferred from the tweets.

7. CONCLUSION

In this paper, we address a novel problem of annotating dynamic semantics to the mobility data using external contextual data. The proposed solution enables us to understand the purposes and interests of a user from his location history, which could benefit a wide range of applications in real world. To this end, we study different methods for annotation and have discussed their advantages and disadvantages. We show that KDE is the best model to capture both the locality and the relevance of words. The effectiveness of our method has been verified through quantitative evaluations and case studies.

There are a number of extensions that could be further explored. First, our current method does not explicitly differentiate between landmark words and event-related words. An event-related word tends to have high density at its center only when the event occurs, whereas a landmark word always has high density at its center. Considering such temporal characteristic may help us differentiate these two types of words. Second, our current method will generate the same annotation for different users, as long as their location records have the same timestamps and locations. Considering the personal location history may enable us to further refine the results. For example, we can promote words which appear in the annotation documents of multiple location records of a user. Third, the semantics at a uni-gram level may be shallow. We can apply multiword expression methods for more interpretable annotations. Finally, as an interesting direction for future work, we plan to explore other types of external contextual data to compensate for the data sparsity data. Potential data sources include environment data, news data, and crime data.

8. REFERENCES

[1] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, and A. Vaisman. A model for enriching trajectories with semantic geographical information. In *Proc. ACM GIS*, 2007.

[2] N. Andrienko and G. Andrienko. Designing visual analytics methods for massive collections of movement data. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 2007.

[3] D. Ashbrook and T. Starner. Using gps to learn significant locations and predict movement across multiple users. *UbiComp*, 2003.

[4] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proc. WWW*, 2008.

[5] J. Bithell. An application of density estimation to geographical epidemiology. *Statistics in medicine*, 9(6):691–701, 1990.

[6] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 1977.

[7] X. Cao, G. Cong, and C. S. Jensen. Mining significant semantic locations from gps data. *Proc. VLDB*, 2010.

[8] D. Chakrabarti and K. Punera. Event summarization using tweets. *ICWSM*, 11:66–73, 2011.

[9] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. ACM CIKM*, 2010.

[10] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. ACM KDD*, 2011.

[11] K. Dehnad. Density estimation for statistics and data analysis. *Technometrics*, 29(4):495–495, 1987.

[12] N. Donthu and R. T. Rust. Note-estimating geographic customer densities using kernel density estimation. *Marketing Science*, 8(2):191–203, 1989.

[13] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. In *Proc. PNAS*, 2009.

[14] G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 2004.

[15] B. Guc, M. May, Y. Saygin, and C. Körner. Semantic annotation of gps trajectories. In *Proc AGILE*, 2008.

[16] 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 *UrbComp*, 2013.

[17] D. Inouye and J. K. Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *PASSAT and SocialCom*. IEEE, 2011.

[18] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. In *Proc. ACM KDD*, 2010.

[19] L. Liao. *Location-based activity recognition.* PhD thesis, University of Washington, 2006.

[20] M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. In *Proc. SIGKDD*. ACM, 2014.

[21] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. Cheung. Mining, indexing, and querying historical spatiotemporal data. In *Proc. ACM KDD*, 2004.

[22] M. Mathioudakis, N. Bansal, and N. Koudas. Identifying, attributing and describing spatial bursts. In *Proc. VLDB*, 2010.

[23] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. ACL, 2004.

[24] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proc. SAC*, 2008.

[25] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Celebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, et al. Mead-a platform for multidocument multilingual text summarization. 2004.

[26] B. Sharifi, M.-A. Hutton, and J. Kalita. Summarizing microblogs automatically. In *Proc NAACL*, 2010.

[27] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Trans. IEEE TKDE*, 2008.

[28] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *AAAI Workshop for Web Search*, 2000.

[29] H. Takamura, H. Yokono, and M. Okumura. Summarizing a document stream. In *Advances in Information Retrieval*, pages 177–188. Springer, 2011.

[30] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 2007.

[31] K. Xie, K. Deng, and X. Zhou. From trajectories to activities: a spatio-temporal join approach. In *Proc. LBSN*, 2009.

[32] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semitri: a framework for semantic annotation of heterogeneous trajectories. In *Proc. EDBT*, 2011.

[33] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semantic trajectories: Mobility data computation and annotation. *ACM Trans. TIST*, 4(3):49, 2013.

[34] Z. Yan, N. Giatrakos, V. Katsikaros, N. Pelekis, and Y. Theodoridis. Setrastream: semantic-aware trajectory construction over streaming movement data. In *SSTD*. Springer, 2011.

[35] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proc. WWW*, 2011.

[36] J.-D. Zhang and C.-Y. Chow. igslr: personalized geo-social location recommendation: a kernel density estimation approach. In *Proc. SIGSPATIAL*. ACM, 2013.

[37] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding transportation modes based on gps data for web applications. *Trans. ACM TWEB*, 2010.

[38] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proc. WWW*, 2009.

[39] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering personally meaningful places: An interactive clustering approach. *TOIS*, 2007.