


# A framework for annotating OpenStreetMap objects using geo-tagged tweets

Xin Chen<sup>1</sup>  · Hoang Vo<sup>1</sup> · Yu Wang<sup>1</sup> · Fusheng Wang<sup>1</sup>

Received: 21 March 2017 / Revised: 30 May 2018 / Accepted: 8 June 2018 /  
Published online: 20 June 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

**Abstract** Recent years have witnessed an explosion of geospatial data, especially in the form of Volunteered Geographic Information (VGI). As a prominent example, OpenStreetMap (OSM) creates a free editable map of the world from a large number of contributors. On the other hand, social media platforms such as Twitter or Instagram supply dynamic social feeds at population level. As much of such data is geo-tagged, there is a high potential on integrating social media with OSM to enrich OSM with semantic annotations, which will complement existing objective description oriented annotations to provide a broader range of annotations. In this paper, we propose a comprehensive framework on integrating social media data and VGI data to derive knowledge about geographical objects, specifically, top relevant annotations from tweets for objects in OSM. We first integrate geo-tagged tweets with OSM data with scalable spatial queries running on MapReduce. We propose a frequency based method for annotating boundary based geographic objects (a polygon), and a probability based method for annotating point based geographic objects (Latitude and Longitude), with consideration of noise. We evaluate our methods using a large geo-tagged tweets corpus and representative geographic objects from OSM, which demonstrates promising results through ground-truth comparison and case studies. We are able to produce up to 80% correct names for geographical objects and discover implicitly relevant information, such as popular exhibitions of a museum, the nicknames or visitors' impression to a tourism attraction.

---

✉ Xin Chen  
xin.chen.1@stonybrook.edu

Hoang Vo  
hvvo@cs.stonybrook.edu

Yu Wang  
yuwang4@cs.stonybrook.edu

Fusheng Wang  
fusheng.wang@stonybrook.edu

<sup>1</sup> Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

**Keywords** Volunteered Geographic Information · Social media · OpenStreetMap · Twitter · Semantic annotation

## 1 Introduction

Large scale geo-crowdsourcing or peer-production Volunteered Geographical Information (VGI) [11], such as OpenStreetMap (OSM)<sup>1</sup> and Wikimapia,<sup>2</sup> has created high potential for establishing reliable sources of geographical information. As a prominent example, OSM accelerates the generation of massive geospatial information from community users and currently has more than 3.7 billion geographical objects. OSM is not only used by end users, but also adopted by companies to support map applications, location recommendation, sports watches, real estate search engine, and many other geospatial services.

OSM aims to provide two types of information about geographical objects: 1) geographical boundaries such as points, lines and regions and 2) annotations or tags. A tag consists of a ‘Key’ and a ‘Value’ to describe the objects. Example objects include building footprints, business places, or tourist attractions. The keys provide a broad class of features (for example, building or amenity) while the values detail the specific features, for example, “building=retail”, “amenity=school”.

However, many of the objects from OSM, in particular, places of interest, have limited annotations. For example, the existing tags in OSM focus on describing general geographical attributes of the real world, not including more detailed information or user reviews, like the nicknames or the impression of visitors to an famous tourism attraction, or data that isn’t current, like temporal exhibitions of a museum and popular events hosting at the places.

On the other hand, much of the social media data is associated with geo-locations. A recent study<sup>3</sup> shows that 1.0% of tweets are geotagged in some way, and 87% of geotagged tweets contain exact coordinates (longitude, latitude). This shows a major increase of geotagged tweets from 0.23% shown in a study in 2010.<sup>4</sup> Such geo-tagged tweets, if combined, could provide rich information that can be potentially associated with other geospatial data sources. For example, the work in [25] uses geo-tagged tweets as external contextual data to annotate mobile users. One natural question is, can we use such geo-tagged social media to support semantic annotations for geographical objects such as churches, museums, and tourism attractions?

In this paper, we propose to enrich OSM objects with semantic annotations by integrating and analyzing geo-tagged social media data, in particular, geo-tagged tweets. This will complement OSM’s objective description oriented annotations to provide a broader range of annotations. Thus, it could significantly improve the value of OSM to support geospatial services. Figure 1 illustrates the process for annotating OSM with a list of relevant words generated from geo-tagged tweets. For example, the tower bridge in London is annotated with the general name (TowerBridge), one popular exhibition (GlassWalkWay or GlassFloor), and many other nearby places. We propose a comprehensive framework on extracting relevant annotations (popular exhibitions of the place, place names, or place

<sup>1</sup>Openstreetmap. [www.openstreetmap.org](http://www.openstreetmap.org).

<sup>2</sup>Wikimapia API. [wikimapia.org/api](http://wikimapia.org/api).

<sup>3</sup><https://www.quora.com/What-percentage-of-tweets-are-geotagged-What-percentage-of-geotagged-tweets-are-ascribed-to-a-venue>

<sup>4</sup><http://thenextweb.com/2010/01/15/twitter-geofail-023-tweets-geotagged>

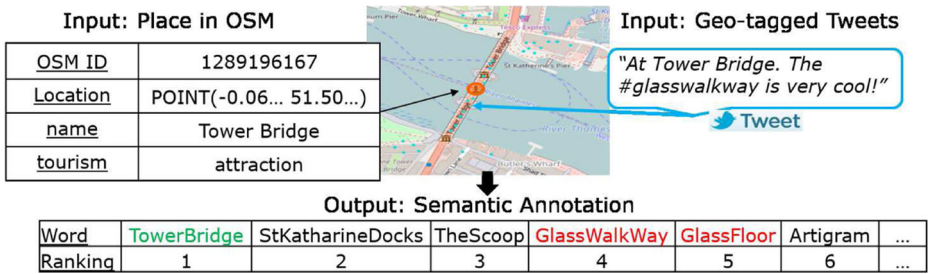


Fig. 1 Examples that use geo-tagged tweet to annotate geographical objects in OSM

nicknames) on top of non-relevant words (names of nearby places) from tweets for places in OSM. We formalize the problem as to find a ranking function that could rank relevant social signals (e.g., words in tweets) on top of non-relevant ones, and measure the likelihood that an annotation candidate is relevant to a given geographic object. Different from traditional information retrieval problems, a new spatial context is introduced into the problem. Thus, our approach will capture both the *relevance* and the *locality* of annotation candidates given a targeted location. As described next, major challenges exist for such spatial semantic annotation problem.

One immediate challenge is integrating large scales of spatial data, including both OSM data and tweets. Capturing local signals for a given location requires spatial data integration across all relevant geospatial objects. For example, we need to search the whole social media corpus for retrieving nearby tweets for a given tourism attraction. However, both VGI and social media platforms produce data at very large scales. OSM has more than 3.7 billion geographical objects and the number is increasing continuously on a daily basis. Moreover, spatial queries, which are essential to support spatial data integration, are highly compute-intensive due to the multi-dimensional nature.

To integrate massive spatial data, we take a MapReduce based approach which partitions the space (which is heavily skewed) into tiles and parallelizes spatial matching queries through MapReduce. This is especially effective to support heavy duty geometric computation during the query.

Another challenge is the difficulty with the estimation of spatial locality due to the diversity of geographical object representations. In OSM, many objects are represented with boundaries, for example, polygons. However, many objects only have a simple point based representation due to limited information or due to the small extent of the objects. For example, less than half of churches in OSM have boundaries.<sup>5</sup>

We propose two alternative methods that handle the two types of spatial objects: frequency based methods for objects with a clear boundary, and probability based method for objects with a point based representation. For the frequency based method, we consider all the tweets contained in the boundary of an object for the analysis. For probability based method, we estimate the probability of a nearby tweet for annotation contribution with respect to the distance between the tweet and the object. Kernel Density Estimation (KDE) model is used for this method.

<sup>5</sup><http://taginfo.openstreetmap.org/>

Another major challenge is the noisy feature of social media data. Social media comprise a broad range of topics. Social media contain large amount of informal languages and personal trivial words from interpersonal chatting or news retweeting, which requires carefully tuned methods to extract meaningful semantic information.

We provide multiple approaches to reduce or remove the effect of noises from signals. For frequency based methods, we provide multiple ways to weigh the relevance of terms for objects, including document corpus, tweet collections, and user collections. For probability based method with KDE, we provide an adaptive approach to minimize the noise effect by tuning the kernel bandwidth inversely with the word density.

While it is difficult to provide ground-truth to evaluate semantic annotations, we propose two alternative approaches to validate our work. We first validate the explicitly relevant annotations with names of places, for which the ground truth is available, and then propose to validate our methods with case studies, with manual evaluation of the relevance of annotation words.

In summary, our work has three major contributions. First, we study and formalize an important problem in geo-social media analytics: integrating social media data and VGI data to derive knowledge about geographical objects. Second, we propose a comprehensive framework on annotating OSM objects using geo-tagged tweets, including a frequency based method and a probability based method. Third, we evaluate our methods on a large geo-tagged tweets corpus and representative geographic objects from OSM, which demonstrates promising results through ground-truth comparison and case studies.

## 2 Related work

Geospatial services provide location based information to consumers, businesses, and governments. This industry is increasing dramatically with the high availability of cost-effective location sensing devices such as smart phones and GPSs. Businesses can rely on geospatial services for improved operational efficiency, targeted marketing and smarter decision making. Consumers can benefit from geospatial services for directions and searching places of interest.

**Geospatial data** Recent years have witnessed an explosion of geospatial data, which provides promising alternative data sources to support geospatial services. While commercial map platforms such as Google Map and Here Map provide APIs for retrieving points of interests, there are major restrictions for public use. Location-based social networks (LBSN) such as Yelp and FourSquare provide constrained access to their place repositories, which are themselves limited. CityGrid and Certain vertical recommendation sites such as TripAdvisor also contain business locations and related customer reviews or tips, which however contains very limited types of objects.

**Geospatial analysis with OSM** While the data consumption of OSM mainly comes from map rendering, geocoding, and smart routing, its analytical value has yet to be explored. The previous OSM data analytical work mainly focuses on the measurement of content bias [18] or predictive analysis such as fine-grained population estimation [5]. In this work, we integrate OSM data with geo-tagged social media for semantic annotation. Recently, Wu et al. [25] use geo-tagged tweets to annotate Twitter users. Work from Sengstock et al. in [21] extracts latent geographic features from Flickr tags, which is for general

geographic knowledge discovery. Coffey et al. [8] use probabilistic topic modelling for semantic enrichment of mobility data recorded in terms of trip counts with Twitter data.

**Geosocial networking** Previous studies that bring together social media users and geographic objects mainly rely on check-in data from Location-based social networks (LBSNs). Karamshuk et al. in [13] utilize user mobility and popularity of places in LBSNs for the problem of optimal retail store placement. Li et al. [16] study the common characteristics of popular venues with check-ins from Foursquare. Georgiev et al. use LBSNs to analyze event patterns [9] and the impact of the Olympic Games on local retailers [10].

**Geospatial analysis with social media** Previous studies have used geo-tagged social media to support data analytics for neighborhood characteristics [20], event detection [14], geolocation [12], or spatio-temporal data mining in particular application scenarios. Most prior works analyze geo-tagged social media within geographic granularity up to street level [15]. For example, Quercia et al. [19] use Flickr and Foursquare to examine the safety of streets. Thomee et al. [23] uncover the colloquial boundaries of locally characterizing regions. In our work, we explore geo-tagged tweets with fine-grained geographic context and extract semantic annotations for individual places of interests.

### 3 Overview

#### 3.1 Problem definition

Our goal is to use geo-tagged social media data to annotate geographical objects. We first define our problem as follows. Table 1 summarizes the notation used in the paper.

##### 3.1.1 Geographic objects

Peer-production VGI platforms, such as OSM or Wikipedia, contain a large amount of geographic objects. In our problem setting, we consider two common representations of

**Table 1** Summary of notation

Notation	Meaning
$p$	a point based geographic object
$b$	a boundary based geographic object
$l = (x, y)$	the location with (latitude, longitude)
$u$	a social media user
$d$	a document generated by social media users
$w$	a word (unigram) contained in the document
$s$	the score that measures the relevance of $w$ w.r.t. the geographic object
$N$	the number of objects
$\delta$	a constant Euclidean distance
$G$	a two dimensional Gaussian kernel function
$h$	a smoothing parameter called the bandwidth
$C$	a $2 \times 2$ covariance matrix.

geographic objects: 1) point based geographic objects and 2) boundary based geographic objects. Point based geographic objects are a set of points  $\mathbf{P} = \{p_1, p_2, \dots, p_{N_P}\}$  where each object  $p_i = [id_{p_i}, l_{p_i}]$  is represented as a single point in space with an object ID  $id_{p_i}$  and the location (latitude, longitude):  $l_{p_i} = (x_{l_p}, y_{l_p})$ . Boundary based geographic objects are a set of polygons  $\mathbf{B} = \{b_1, b_2, \dots, b_{N_B}\}$  where each object  $b_i = [id_b, L_b]$  is represented with an object ID and a closed polygon consisting of an ordered list of points that delineates the boundary of the object in space. The boundary  $L_b = \{l_1, l_2, \dots, l_N\}$  consists of latitude and longitude based point  $l_i = (x_{l_i}, y_{l_i})$ .

### 3.1.2 Geo-tagged social media

The geo-tagged social media signals can be represented as a set of documents  $\mathbf{D} = \{d_1, d_2, \dots, d_{N_D}\}$ . Each document  $d_j$  for  $j = 1, 2, \dots, N_D$  consists a tuple  $\langle id_{d_j}, id_u, l_{d_j}, W_{d_j} \rangle$  where  $id_{d_j}$  and  $id_u$  denotes the ID of the document and the ID of the user who generates this content. The document location  $l_{d_j}$  is a single point in the space represented by latitude and longitude based position  $(x_{d_j}, y_{d_j})$ .  $\mathbf{W}_{d_j} = \{w_1, w_2, \dots, w_{N_W}\}$  indicates a set of associated features extracted from the document  $d_j$ . While social media provide a wide range of signals such as image, video, their associated tags or metadata, for our study, we focus on unigrams from tweet content.

### 3.1.3 The semantic annotation problem

Given a collection of geographic objects  $\mathbf{P}$  or  $\mathbf{B}$  and a spatial-sensitive social media corpus  $\mathbf{D}$ , our goal is to integrate geographic information in  $\mathbf{P}$  and  $\mathbf{B}$  and social contextual information in  $\mathbf{D}$ . With integrated geospatial data, this work focuses on extracting semantic annotations from social signals w.r.t. fine-grained geographic objects either in the form of a point  $\mathbf{P}$  or a polygon  $\mathbf{B}$ . For example, restaurants or coffee shops are typically represented as points, and building footprints, parking lots, or pitches are represented as polygons. With a variety of scales, churches, on the other hand, have no dominant form for spatial representations. The semantic annotations for a targeted geographic object is a set of relevant words,  $\mathbf{A} = \{(w_1, s_1), (w_2, s_2), \dots, (w_{N_A}, s_{N_A})\}$ , where  $s_i$  ( $i = 1, 2, \dots, N_A$ ) is a score that measures the relevance of  $w_i$  w.r.t. the geographic object.

The semantic annotation problem can be defined as to find a ranking function  $\mathbf{f}(\mathbf{p}_i, \mathbf{w}_j)$  (or  $\mathbf{f}(\mathbf{b}_i, \mathbf{w}_j)$ ) for a word  $w_i \in V_D$  w.r.t. a given geographic object  $p_i$  or  $b_i$ , where  $\mathbf{V}_D = \{w_1, w_2, \dots, w_N\}$  refers to the vocabulary that includes all annotation candidates generated from the social media corpus  $D$ . Analogous to a typical information retrieval task, our goal is to provide satisfactory ranking function to rank relevant annotation keywords on top of non-relevant ones.

## 3.2 Overview of methods

The key challenge of the spatial semantic annotation problem is how to measure the likelihood that a word  $w_j$  is relevant to the given geographic object. A unique constraint for our problem is that the annotation candidates and the annotating targets possess a spatial context, whereas a traditional information retrieval problem ranks relevant documents to a query. Thus, our goal is to propose an applicable model that should capture both *the relevance* and *the locality* of annotation words w.r.t. the targeted locations of geographic objects.

### 3.2.1 Spatial data integration

The spatial relevance of tweets will be largely affected by the proximity to the geographical objects, the extent and representation of the geographical objects. For a boundary based geographic object, intuitively, tweets contained in the boundary will likely have a higher “signal to noise ratio” than those outside of the boundary. For point based geographic objects, similarly, nearby tweets within a distance should have higher relevance than those outside, and the relevance could be affected by the distance.

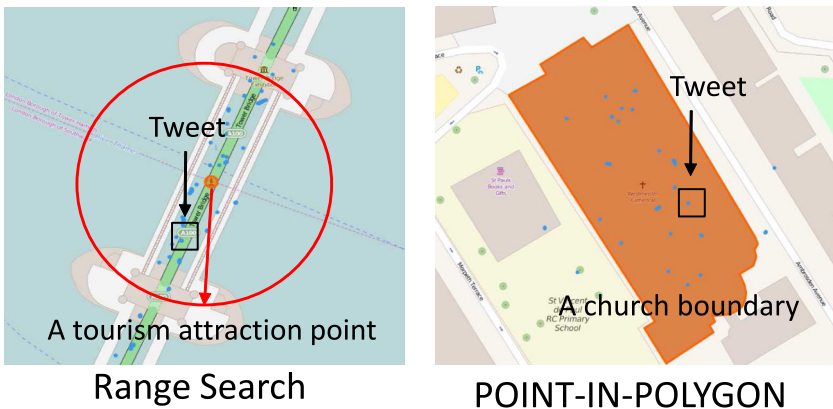
To capture the spatial locality of words, we propose to use spatial queries to cross-match tweets with geographical objects, and filter tweets based on spatial proximity – only tweets close to the geographical objects will be used. Due to the massive volume of geospatial objects from OSM, the vast number of tweets, and the high computational complexity associated spatial queries (for example, containment), such spatial queries will be very expensive.

To support scalable spatial queries, we first perform skew-aware space partitioning to generate balanced tiles, and then run spatial queries for each tile in parallel through MapReduce by invoking an on-demand spatial query engine. We then normalize query results for objects across boundaries. We extend our current work Hadoop-GIS [1, 4, 24] to support the queries needed for such data integration. The two query scenarios are illustrated in Fig. 2.

### 3.2.2 Frequency based semantic annotation

Once the nearby social signals are aggregated for each object, we propose two alternative methods to find the ranking functions that can produce relevant words within the refined annotation candidates: frequency based methods and probability based method.

Boundary represented geospatial objects normally have a larger extent compared to point based objects. Intuitively, a term that occurs frequently within a place may be a relevant annotation. We can count the occurrences of nearby words based on the Term Frequency (TF) w.r.t. the targeted location.



**Fig. 2** Two types of data integration between geographic objects and geo-tagged social media (1) Range Search and (2) POINT-IN-POLYGON

To reduce noisy terms, we can improve frequency based method by smoothing it with a weighting factor, using Inverse Document Frequency (IDF). IDF can measure how much information the word provides by checking whether the word is common or rare across all documents. So even though tweets has limited lengths, IDF should still give smaller weights to very commonly occurring words. Since multiple occurrences of a term from distinct tweets or users tend to contribute more than those from a single tweet, we then further propose collective tweet weighting and collective user weighting.

### 3.2.3 Probability based semantic annotation

For point based object representation, one issue is that aggregating nearby words requires a distance threshold. How to choose such threshold is challenging as different place categories may have different scales of neighborhoods. An inappropriate threshold, in this case, would result in high frequency words from irrelevant tweets.

To address these limitations, we propose to use probability based method which models the relevancy versus the distance. We take the Kernel Density Estimation (KDE) based methods for the ranking problem. KDE has been previously used for modeling human location [17] and generating semantic annotations for mobility data [25]. This work focuses on modeling geo-tagged words with KDE for annotating point based geographic objects. Other than frequency based methods, KDE models the spatial density of the word occurrences and then weights differently for words with different distances. The estimated spatial density can be controlled over a bandwidth parameter  $h$ . We can analyze and set the parameter  $h$  with respect to different types of annotation words or different place categories.

## 4 Spatial data integration

Integrating tweets with OSM will require two types of spatial queries as shown in 2. 1) Containment based query or point in polygon query: for each boundary based geospatial object, find all tweets contained in the boundary; and 2) Range search: for each point based geospatial object, find all nearby tweets within distance  $d$ . The later can be performed by generating a buffered circle with radius  $d$  and a containment query. We extend our previous work Hadoop-GIS, a MapReduce based spatial query system, to support the queries.

We propose to provide spatial data integration through MapReduce based spatial queries at large scale. MapReduce based systems have emerged as a scalable and cost effective solution for massively parallel data processing. However, most of these MapReduce based systems either lack spatial query processing capabilities or have limited spatial query support. While the MapReduce model fits nicely with large scale problems through key-based partitioning, spatial queries and analytics are intrinsically complex and difficult to fit into the model due to its multi-dimensional nature [3].

To support large scale spatial queries on these datasets, the following steps are performed: spatial partitioning; tile based spatial query processing with MapReduce; and result normalization or duplicate removal for boundary-crossing objects. The overall workflow is shown in Fig. 3.

The space of OSM is first partitioned into balanced tiles [2, 24] based on Sort-Tile-Recursive (STR) algorithm, which tries to order and pack spatial objects for bulk loading to generate an R-Tree for all the OSM objects. Note that OSM objects are first computed to generate the minimal bounding rectangles (MBR). The MBRs of the parent nodes of leaves will become natural partition boundaries.



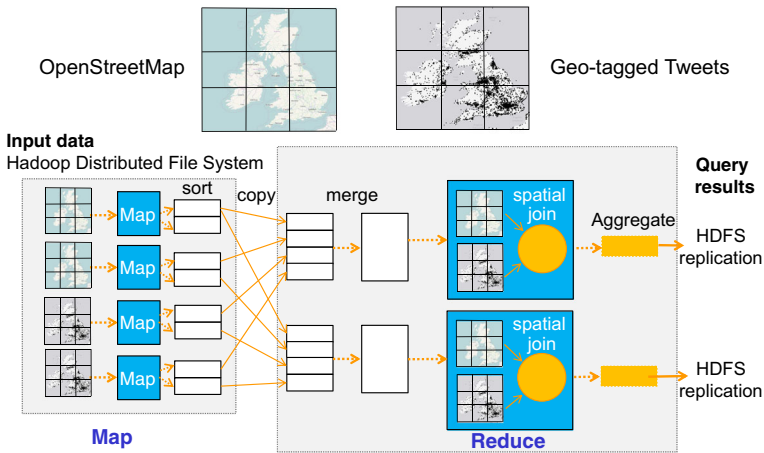


Fig. 3 The workflow of MapReduce based spatial data integration of geo-tagged tweets and OSM objects

Once the partition boundaries of OSM data are generated, MapReduce is used to match tweets for containing OSM objects. First, for each partition represented with an MBR, all containing OSM objects and tweets in each partition are identified through a map function by comparing the boundaries. Then a reducer function is started to match tweets to containing OSM objects for each tile. An R\*-Tree for tweets and an R\*-Tree for OSM objects are built in memory on-the-fly for each tile, respectively. Based on the two R\*-Trees for the two datasets, a spatial join algorithm is invoked to find the containment relationships between OSM objects and tweets through traversing the R\*-Trees [7]. Following this, a geometric computation on containment relationship is further checked if a tweet is contained not only in the MBR of an OSM object, but also in the polygon of the OSM object.

After all the matching is done, another MapReduce job is performed to identify all duplicated objects. As an object on the boundary of tiles will be assigned multiple times during the partitioning, there will be duplicated results. A sort is performed which removes all the duplicated objects in the result.

Note that the overhead of on-demand indexing is a very small fraction of the overall cost, but it significantly reduces the search space and provides very efficient queries. The computational geometry for containment relationship is heavy duty and takes a large portion of the total time, which is actually effectively parallelized through MapReduce.

### 5 Frequency based semantic annotation

We start with simple term frequency (*TF*) based approach to evaluate the term relevance for annotations based on frequencies of occurrences, and then refine it with document corpus based weighting (*TF-IDF*) to reduce weights for terms across multiple documents. As multiple occurrences of a term from distinct tweets contribute more than those from a single tweet, we further propose to consider collective tweet weighting (*TF-per-tweet-IDF*). Last, since distinct users tend to provide more independent opinions, we introduce collective user weighting (*TF-per-user-IDF*).

### 5.1 Term frequency based weighting

Given a geographic object, one intuitive semantic annotation method is to rank the nearby social media signals according to the frequencies of occurrence, i.e., term frequency (**TF**). Formally, given a geographic boundary  $b_i$  or a geographic point  $p_i$ , a containment based query  $contain(b_i, D)$  and a range-within-distance query  $range(p_i, D, \delta)$  aggregate all the social media documents located within the boundary of  $b_i$  as  $D_{b_i}$  and documents within the range of a distance  $\delta$  to  $p_i$  as  $D_{p_i}$ . The TF based ranking function  $TF(b_i, w_j)$  and  $TF(p_i, w_j)$  then measures the relevance of a word  $w_j$  in Eq. 1, where  $W_{b_i}$  and  $W_{p_i}$  indicate the set of associated features extracted from the document  $D_{b_i}$  and  $D_{p_i}$  respectively.

$$\begin{aligned}
 TF(b_i, w_j) &= |\{w_j \in W_{b_i} : l_{w_j} \text{ in } L_{b_i}\}| \\
 TF(p_i, w_j, \delta) &= |\{w_j \in W_{p_i} : dist(l_{p_i}, l_{w_j}) < \delta\}|,
 \end{aligned}
 \tag{1}$$

The TF based ranking function does not distinguish common words, stop words, or expression words, such as “im”, “start”, and “time”, which overwhelm important terms with richer semantics. To filter such non-relevant words and boost the ranking of more important words, we use the algorithm of term frequency-inverse document frequency (TF-IDF) to smooth the direct term frequencies.

### 5.2 Document corpus based weighting

Given a large collection of documents, **TF-IDF** is often used to represent the relative importance or uniqueness of a term to a specified document. Intuitively, TF-IDF based method gives a low weight to a word that is frequent in one document but also appears across many other documents. In our application scenario, tweets are usually short in length but accumulate exponentially as regard to the total number of documents, which provide a rich data source for smoothing the term frequencies.

Given a geo-tagged social media corpus  $D$  and a geographic object  $b_i$  or  $p_i$ , the TF-IDF based ranking function  $TFIDF(b_i, w_j, D)$  and  $TFIDF(p_i, w_j, D)$  measure the relevance of a word  $w_j$  in Eq. 2, where  $W_{b_i}$  and  $W_{p_i}$  indicate the set of associated features extracted from the document  $D_{b_i}$  and  $D_{p_i}$  respectively.

$$\begin{aligned}
 TFIDF(b_i, w_j, D) &= TF(b_i, w_j) * IDF(D, w_j) \\
 IDF(D, w_j) &= \log \left[ \frac{N_D}{1 + |\{d_k \in D : w_j \in W_{d_k}\}|} \right] \\
 TFIDF(p_i, w_j, \delta, D) &= TF(p_i, w_j, \delta) * IDF(D, w_j) \\
 IDF(D, w_j) &= \log \left[ \frac{N_D}{1 + |\{d_k \in D : w_j \in W_{p_k}\}|} \right],
 \end{aligned}
 \tag{2}$$

### 5.3 Collective tweet weighting

The collective signals from overall social media context have been effectively utilized to smooth a term frequency through the weight of inverse document frequency. Our spatial data integration framework, on the other hand, generates a local context through the aggregated nearby documents, which contain additional knowledge for the relevance of a term w.r.t. targeted objects. For example, a term mentioned in multiple tweets in the local context should imply higher relevance than a term with multiple occurrences coming from a single tweet.

We propose collective tweet weighting (**TF-per-tweet-IDF**) to smooth the direct term frequency by counting term occurrences per tweet, i.e., multiple mentions in a single tweet count only once. Formally, the collective weighting method is defined in Eq. 3, where  $D_{b_i}$  is aggregated documents located within the boundary of  $b_i$ , and  $D_{p_i}$  is aggregated documents within the range of a distance  $\delta$  to  $p_i$ .

$$\begin{aligned}
 TF_{tweet}(b_i, w_j) &= |\{d_k \in D_{b_i} : w_j \in W_{d_k}\}| \\
 TF_{tweet}(p_i, w_j, \delta) &= |\{d_k \in D_{p_i} : w_j \in W_{d_k}\}| \\
 TF_{tweet}IDF(b_i, w_j, D) &= TF_{tweet}(b_i, w_j) * IDF(D, w_j) \\
 TF_{tweet}IDF(p_i, w_j, \delta, D) &= TF_{tweet}(p_i, w_j, \delta) * IDF(D, w_j),
 \end{aligned}
 \tag{3}$$

### 5.4 Collective user weighting

One added knowledge within the social media platform is the author information. Terms from the same user tend to be similar and different users tend to generate more independent contents. By identifying the original source of each term, we can distinguish terms coming from the same user or from diverse users.

We propose collective user weighting (**TF-per-user-IDF**): the multiple occurrences of a term from the same user will be counted only once and the frequency of a term from distinct users will be the count of distinct users. Formally, it is defined in Eq. 4, where  $U_{b_i}$  and  $U_{p_i}$  are the set of users who generate the social media documents in  $D_{b_i}$  and  $D_{p_i}$  respectively.  $W_{u_k}$  indicates the set of associated features extracted from the document of the user  $u_k \in U_{b_i}$  or  $u_k \in U_{p_i}$ .

$$\begin{aligned}
 TF_{user}(b_i, w_j) &= |\{u_k \in U_{b_i} : w_j \in W_{u_k}\}| \\
 TF_{user}(p_i, w_j, \delta) &= |\{u_k \in U_{p_i} : w_j \in W_{u_k}\}| \\
 TF_{user}IDF(b_i, w_j, D) &= TF_{user}(b_i, w_j) * IDF(D, w_j) \\
 TF_{user}IDF(p_i, w_j, \delta, D) &= TF_{user}(p_i, w_j, \delta) * IDF(D, w_j),
 \end{aligned}
 \tag{4}$$

**Discussion** The frequency based methods with weighting are based on the assumption that the spatial relevance of a tweet to an OSM object is certain, i.e., a tweet is clearly contained in the object boundary. Thus, such methods work better for large objects with boundary based representations.

## 6 Probability based semantic annotation

For point based object representation, to associate spatial relevance of a tweet to an OSM object, a circle based approximate buffer is created for spatial matching. Choosing the right threshold for the buffer is challenging as each type of places may have very different scale of neighborhood. For example, a coffee shop has a much smaller extent than a church. A popular landmark such as a tourism attraction around a coffee shop may generate many tweets which are irrelevant to the shop. A frequency based method is not working any more, as it treats all nearby words with same spatial relevance regardless of the distance.

For objects with only point based representations, or for objects with very small extents, the spatial relevance will be dependent on the distance between the tweet and the object. We propose a probability based method to model the probability of the relevance of a word to a geospatial object as a function of the distance. Kernel Density Estimation (KDE) is

a non-parametric method for estimating a density function from a random sample of data. Prior work has utilized KDE for modeling the spatial density of word occurrences, individual mobility data [25], and check-ins from LBDNs. Our work investigates KDE model for annotating geographic points with the spatial probability of word occurrences.

### 6.1 Kernel density estimation

As mentioned earlier, frequency based method for boundary objects without enough spatial extents leads to data sparsity problem and introduces more noise from nearby landmarks. The essence of KDE based model is to estimate a spatial density from word occurrences. The counts of word occurrences are then smoothed out with the spatial density over the continuous space.

Formally, let  $\mathbf{L}^{w_j} = \{l_1^{w_j}, l_2^{w_j}, \dots, l_N^{w_j}\}$  refer to all occurrences of a word  $w_j \in V_D$  where  $V_D$  is from a geo-tagged social media corpus  $D$ . Given a two-dimensional Gaussian kernel function  $\mathbf{G}$  and a fixed bandwidth  $\mathbf{h}$ , we propose a ranking function (**KDE-fixed**) for the word  $w_j$  w.r.t. a geographic point  $p_i$  as described in Eq. 5, where  $C_h$  refers to a  $2 \times 2$  covariance matrix.

$$\begin{aligned}
 KDE_{fixed}(p_i, L^{w_j}, G, h) &= \frac{1}{|L^{w_j}|} \sum_{k=1}^{N^{w_j}} G_h(l_k^{w_j}, l_{p_i}) \\
 G_h(l_k^{w_j}, l_{p_i}) &= \frac{1}{2\pi h} \exp\left[-\frac{1}{2} (l_k^{w_j}, l_{p_i})^T \mathbf{C}_h^{-1} (l_k^{w_j}, l_{p_i})\right] \\
 C_h &= \begin{bmatrix} h & 0 \\ 0 & h \end{bmatrix}, \tag{5}
 \end{aligned}$$

Similar to the idea of collective tweet weighting, the KDE-fixed method refer to the word occurrences per tweet, i.e., multiple mentions of a word in a single tweet count only once. To extend the KDE-fixed method with collective user weighting, we propose an alternative KDE-fixed method (**KDE-fixed-per-user**): the multiple occurrences of a word from the same user will be counted only once and the centroid point of these multiple occurrences will present the location of the word.

With the KDE-fixed method, each word occurrence contributes to the overall ranking score according to its distance to the targeted point, which provides a more accurate estimation about the relevance and omits the requirement of a boundary for encompassing nearby words. Previous work [17, 22] suggests that the choice of the bandwidth value  $h$  determines the shape of the resulting spatial density. While a smaller  $h$  produces a sharper peaked distribution around the locations of word occurrences, an inappropriately large bandwidth  $h$  would generate an oversmoothed estimation. In the experiment, we try to adjust  $h$  with different values for our datasets with different types of objects.

### 6.2 KDE with adaptive bandwidth

The above KDE-fixed method requires tuning of the bandwidth which is time consuming. Besides, the smoothing is homogeneous for all the words regardless of the difference of their spatial densities. For example, the name of an iconic symbol in the city tend to accumulate near the landmark address. The bandwidth in such situation should obviously be different with that around a sparsely populated area.

In order to prevent either overfitting or oversmoothing, we take another adaptive based approach (**KDE-adaptive**) where the bandwidth is set adaptively for the KDE based ranking function. Given a term  $w_j \in V_D$ , a customized bandwidth  $h$  would be generated according to the provided occurrence locations  $L^{w_j}$ . Inspired by Breiman et al. [6], we set the bandwidth  $h_{w_j}$  as the distance between the targeted geographic point  $p_i$  and its  $k$ -th nearest neighbor. The formal definition of KDE-adaptive is described in Eq. 6, where  $h_j$  refers to the Euclidean distance to the  $k$ -th nearest neighbor to  $l_{p_i}$ .

$$\begin{aligned}
 KDE_{adaptive}(p_i, L^{w_j}, G) &= \frac{1}{|L^{w_j}|} \sum_{n=1}^{N^{w_j}} G_{h_j}(l_n^{w_j}, l_{p_i}) \\
 C_{h_j} &= \begin{bmatrix} h_j & 0 \\ 0 & h_j \end{bmatrix},
 \end{aligned}
 \tag{6}$$

Similar to the KDE-fixed method, multiple mentions of a word in a single tweet count only once. For the alternative KDE-adaptive method with collective user weighting (**KDE-adaptive-per-user**), the multiple occurrences of a word from the same user count once and use the centroid point as its location.

In our problem setting, noisy signals such as stop words, expression words or spams accumulate across time could overwhelm the spatial semantics of our interest. By setting the bandwidth according to  $k$ -th nearest neighbor, the adaptive kernel approach tunes the bandwidth inversely with the word density. For the word with a low density, the distance of its  $k$ -th nearest neighbor to a given object is larger than the word with a dense occurrence, which results in a larger bandwidth to adapt the sparseness of the data. In our experiment, we evaluate different choices of  $k$  for the datasets.

## 7 Experimental evaluation

We evaluate the performance of frequency based method and probability based method to annotate multiple types of geospatial objects extracted from UK with tweets. We also compare the difference between frequency based method and probability based method for annotating point based geographical objects. We provide both ground-truth based comparison and case studies with manual evaluation.

### 7.1 Datasets

We downloaded the entire set of OSM data from Planet OSM<sup>6</sup> and filtered the data to generate a collection of representative places from UK. The places of interests are selected according to the tag information in OSM data, for example, railway stations, sports centres, tourism attractions, tourism museums, historic sites, cinemas and theatres, and places of worship (i.e., churches). The geo-tagged tweets corpus was collected for the period between Nov 1, 2014 and Sep 09, 2015 and contains 343,779,205 geo-tagged tweets in total. For simplicity, only English words from tweet contents are considered as annotation candidates in the experiments. The overall statistics of the datasets is summarized in Table 2.

<sup>6</sup><http://planet.openstreetmap.org/>

**Table 2** Description on datasets

Geospatial data sources		# of objects
OSM in UK area	Boundary objects	4,156,607
	Point objects	506,086
Geo-tagged Tweets 11/01/2014 - 09/09/2015		343,779,205

## 7.2 Experimental settings

**Name detection experiment** We designed a name detection experiment to assess our proposed semantic annotation methods. While it is difficult to provide ground-truth to evaluate semantic annotations, one special OSM tag, the name for a given place, is provided by most OSM objects and could serve as ground truth to evaluate our proposed annotation methods. We built a ground truth dataset by extracting a subset of places with their name tags contained in OSM data and appearing in the nearby tweet contents.

**Ground truth Dataset Generation** In detail, to build the ground truth dataset, our spatial data integration framework first cross-matches the whole geo-tagged tweets corpus with geographical objects in OSM. For boundary based objects, the integrated corpus includes all boundaries that contain at least one tweet. For point based objects, the integrated corpus includes all points with at least one tweet detected within their buffered circle ranges. In our experiment, the radius distance for the buffered circle is set up to 0.002 decimal degrees (worth up to 250 meters). We then filter out the ground truth corpus with the place names appearing in the nearby tweet contents. A set of representative place categories is used in the following experiments (Table 3).

**Evaluation metrics** Given a geographic object, the semantic annotation result is a sorted list of relevant words ordered by their ranking scores from a semantic annotation methods described in Sections 5 and 6. We validate whether the top K words with the highest ranking scores will contain the place name. Given a collection of boundary based objects or point based objects in the ground truth, the name detection accuracy is the percentage of places with their names contained in the top K annotations.

**Table 3** Statistics of ground truth datasets

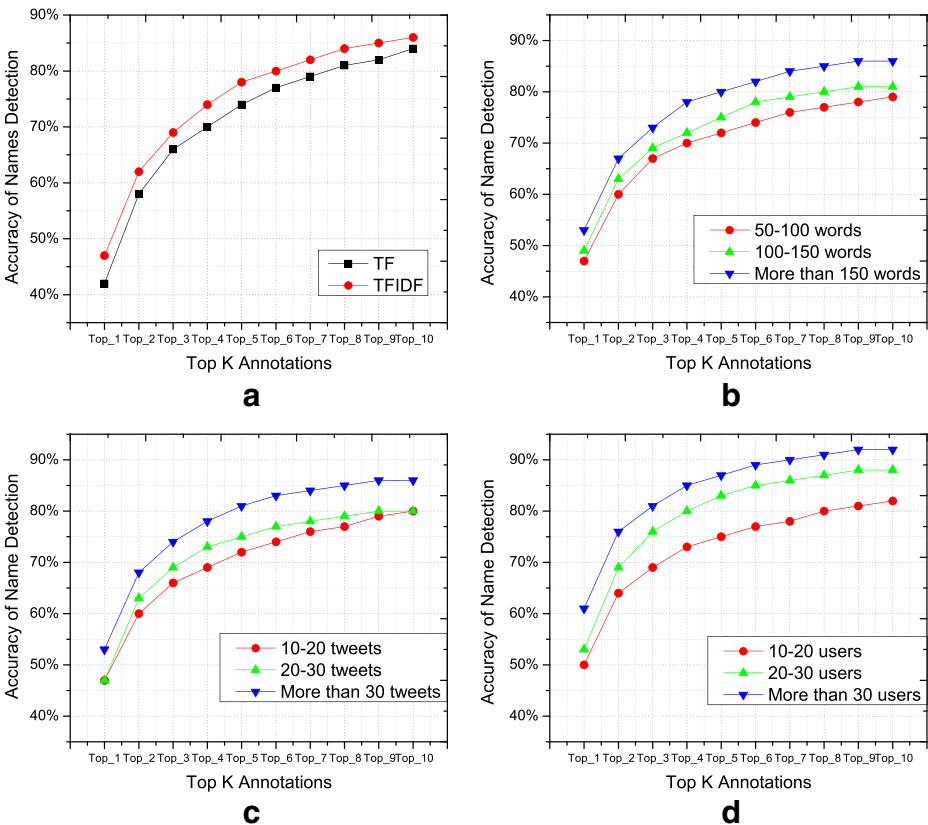
Geographic format	Place category	# of Objects
Boundary objects	Station	365
	Tourism	362
	Church	1,914
	Stadium	487
	Park	9,329
	Theatre	386
	Shop	447
Point objects	Station	2,467
	Church	1,317
	Tourism	749
	Sport	571

### 7.3 Evaluation of frequency based methods

We evaluate frequency based methods for boundary based objects with place name detection. Figure 4a illustrates the performance of TF and TF-IDF methods on all seven types of boundary based objects in the ground truth datasets. TF-IDF clearly outperforms TF for name detection accuracy. Such result indicates that collective signals from overall social media context can effectively smooth the direct term frequency through weighting in IDF.

In reality, some areas are more densely populated than others. We further examine the performance of TF-IDF method on places with different popularity. We then group all boundary based objects according to their contained user counts, tweet counts and word counts respectively. As shown in Fig. 4b–d, places that contain more signals tend to have a higher accuracy for name detection experiments, no matter how the places are grouped. With a closer examination, however, we find that grouping with user counts has a higher improvement than grouping with word counts and tweet counts. This implies that higher user appearance can supply richer information for annotations.

Based on the observation, we design two variants of TF-IDF method discussed in Section 5, i.e., TF-per-tweet-IDF, and TF-per-user-IDF, which incorporate local signals



**Fig. 4** a Name detection accuracy of TF and TF-IDF methods for top K results; b–d Name detection accuracy of TF-IDF method with boundary object grouping based on user count, tweet count and word count respectively

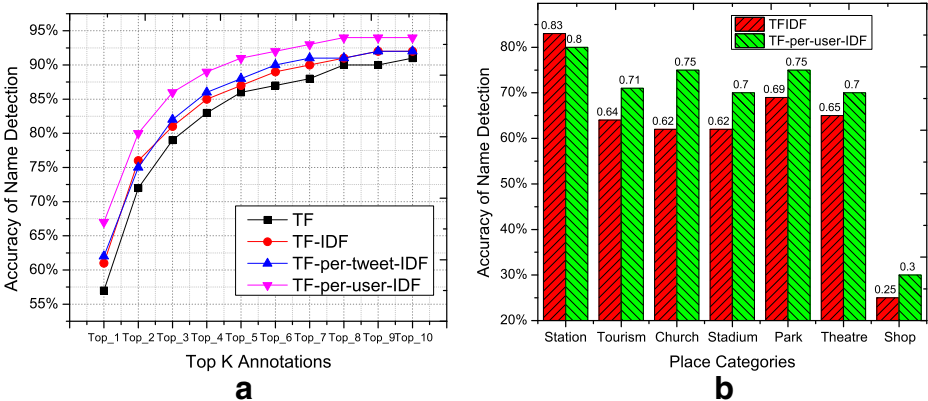
from aggregated nearby documents. We then use four frequency based methods to annotate names of the places containing tweets coming from at least 30 users. The results in Fig. 5a demonstrate the effectiveness of user information for enhancing annotations, and indicate that a larger number of distinct users mentioning the same keyword will provide stronger evidence for the relevance of the keyword to the corresponding places.

We also evaluate the performance for different place categories. Figure 5b shows the name detection accuracy using top 10 annotations for different place categories in our ground truth dataset. TF-per-user-IDF consistently outperforms TF-IDF across almost all categories. The only exception is railway stations, where TF-IDF performs better. Besides, the overall accuracy for railway stations is also much higher than other categories. This implies that geo-tagged tweets from stations contain more spatial dependent information and have less noise. On the other hand, the overall accuracy for shops has a very low accuracy, which implies its nearby social context has a much lower “signal to noise ratio”.

### 7.4 Evaluation of probability based methods

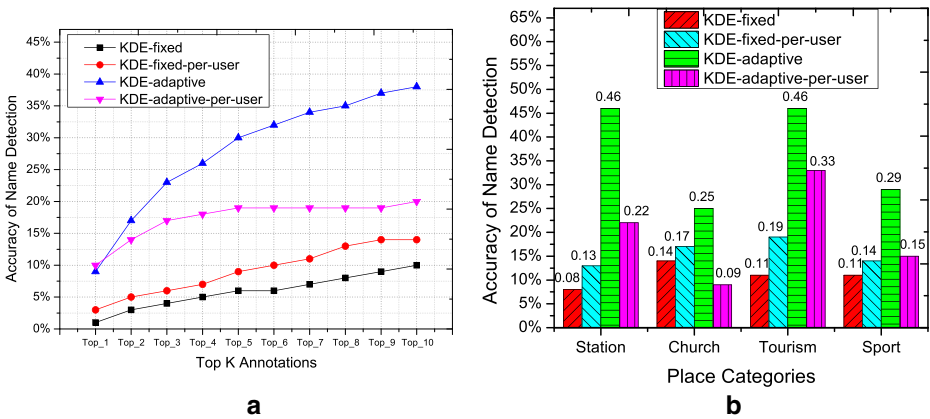
We evaluate probability based methods for annotating geographical points. We first compare different KDE based methods for annotating point objects. Figure 6a shows the name detection accuracy of all point objects combined. Figure 6b shows the accuracy with top 10 annotations for different types of places in our datasets. We experiment different parameters and compare the highest accuracies for both KDE-fixed based methods (with bandwidth value  $h$  set as 0.0001 decimal degree) and KDE-adaptive based methods (with the number of neighbors as 2). The adaptive bandwidth methods clearly outperform fixed bandwidth methods.

The probability based methods rely on the bandwidth parameter  $h$  for estimating the word density distribution. In order to prevent either overfitting or oversmoothing, smaller bandwidth values should be assigned to denser words and larger ones should fit to sparse words. To better understand the influence of bandwidth, we study the effect of bandwidth on accuracy. Figure 7 illustrates the accuracy trend with varying bandwidth for detecting names of churches. We experiment on churches because this category of places (in our dataset) includes both tourism hotspots with many dense words and local churches with only



**Fig. 5** a Name detection accuracy of frequency based methods; b Name detection accuracy of TF-IDF and TF-per-user-IDF for different place categories



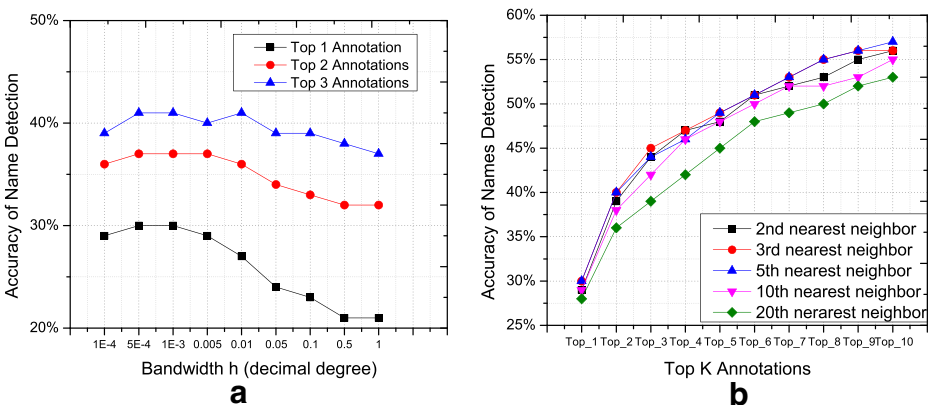


**Fig. 6** **a** Name detection accuracy of probability based methods; **b** Name detection accuracy of KDE-fixed-with-weighting and KDE-adaptive for different place categories

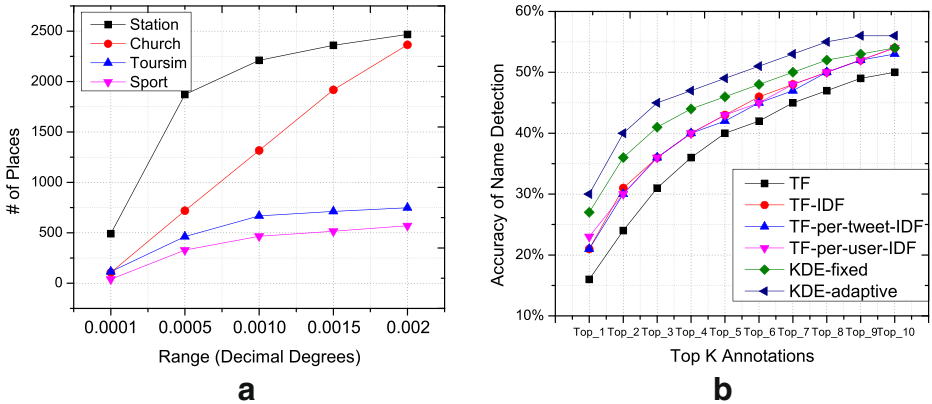
sparse words. We observe that KDE-fixed reaches the highest accuracy with  $h$  between 0.0005 and 0.01 decimal degrees (20 meters to 1 kilometer). For KDE-adaptive, we find that the accuracy is decreased when  $h$  is larger than the distance between the place and its 10th nearest neighbor.

### 7.5 Frequency based methods vs probability based methods

We compare frequency based methods and probability based methods for detecting the names of places for point based objects. To support point based objects with frequency based method, an approximate buffer with a distance threshold  $\delta$  is used to identify nearby tweets. The results in Fig. 8a show that, when the range search threshold  $\delta$  is decreased, the number of places with their names detected from nearby tweets is also decreased. This suggests that a smaller distance threshold  $\delta$  will lead to a loss of relevant information.



**Fig. 7** **a** Name detection accuracy with increasing bandwidth  $h$  for KDE-fixed; **b** Name detection accuracy for KDE-adaptive with  $h$  set to the distance between the place and its  $k$ -th nearest geo-tagged tweet



**Fig. 8** a Number of places with names detected from nearby tweets with varying range search threshold  $\delta$ ; b Name detection accuracy of different semantic annotation methods for point based objects

We compare the performance of frequency based methods versus probability based for identifying names of churches. As shown in Fig. 8b, both KDE-fixed (with  $h$  set to 0.01 decimal degree) and KDE-adaptive (with the number of neighbors as 3) outperform all frequency based methods.

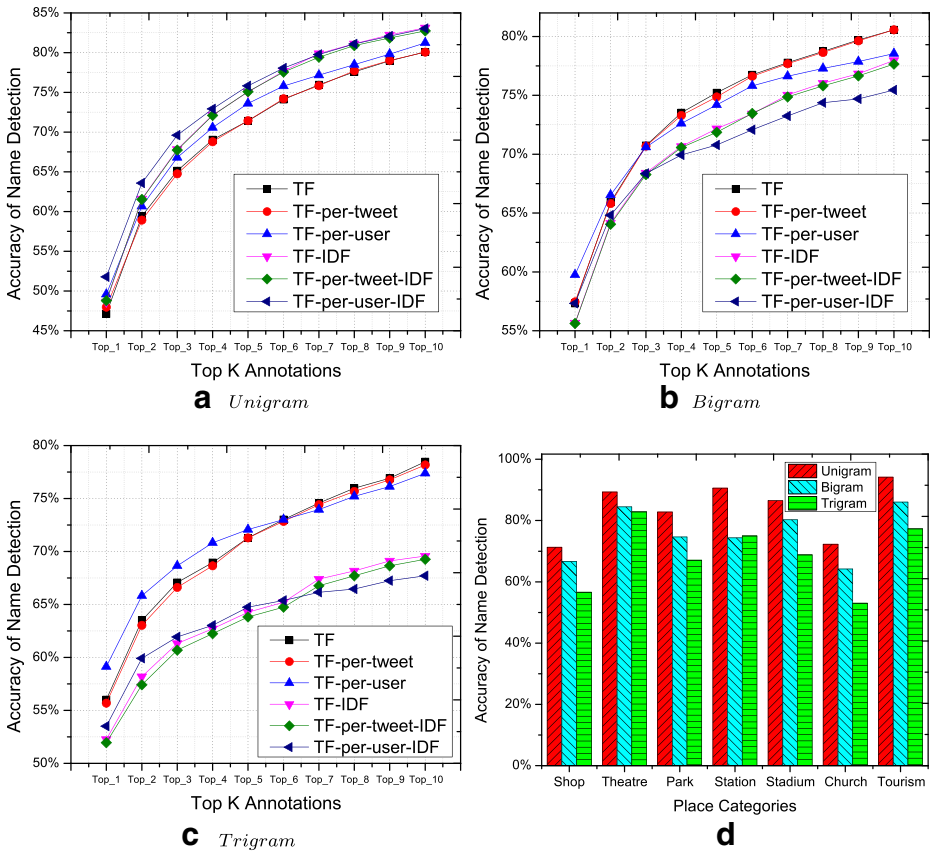
### 7.6 Case studies

We also perform two case studies to evaluate the annotation results with human interpretation, one for boundary based object (Imperial War Museum North) and the other for point based objects (Tower Bridge). We first classify semantic annotation results into three categories: explicitly relevant, implicitly relevant, and non-relevant. Explicitly relevant annotations are about major characteristics of an object, for example, the name and theme of a museum. Implicitly relevant annotations are more about derived information or minor information, for example, a collection in a museum. The case studies only generate semantic annotations from unigrams. However, bigrams and trigrams may contain more semantic information, so we perform additional experiments to compare unigrams, bigrams, and trigrams as shown in Fig. 9.

#### 7.6.1 Boundary object: imperial war museum North

For Imperial War Museum North, we compare top 20 annotations from four frequency based methods. The existing tags in OSM (Fig. 10a) mainly contain the name and place category. Example explicitly relevant annotations include “war”, “museum”, “imperial”, and “iwmn” (the abbreviation for the full name). Implicitly relevant annotations include “architecture”<sup>7</sup>, “wellingtonbomber” (hashtag for wellington bomber), “gunturret” (hashtag for gun turret), which are either the collections or the characteristics of Imperial War Museum. Non-relevant words include names of nearby places such as “univeristyofmanchester” (hashtag for University of Manchester) or the city name alone ‘manchester’ which is too broad as an

<sup>7</sup>The two original tweets for “architecture” are: #imperialwarmuseumnorth #manchester #salfordquays ... Impressive architecture #lovemanchester <https://t.co/eS4tJrkEgo> and The walls between art and engineering exist only in our minds #bridge #architecture #manchester <https://t.co/SVndM4ARck>



**Fig. 9** Name detection accuracy of 6 frequency based methods for **a** uni-gram, **b** bigram, and **c** trigram; **d** Name detection accuracy of TF-IDF method with boundary object grouping based on user count, tweet count and word count respectively

annotation. As shown in Fig. 10b, TF-per-user-IDF produces much more relevant annotation keywords (6 explicitly relevant and 4 implicitly relevant) than other frequency based methods.

### 7.6.2 Point object: tower bridge

For Tower Bridge, we compare top 20 annotations from two frequency based methods and two probability based methods. Explicitly relevant annotations include “walkway”, “glasswalkway” (the hashtag for glass walk way), “glassfloor”, which are either famous exhibition or a feature of Tower Bridge. The non-relevant words include common language or names of nearby businesses and landmarks. As shown in Fig. 11b, non-relevant words from frequency based methods contain more common language, and probability based methods generate names of nearby landmarks. KDE-adaptive method produces more relevant annotations than KDE-fixed method. The KDE-adaptive method in this case study detects one explicit relevant word as the top 1 result and 2 other implicitly relevant words among the top 6 results.

**a** A Boundary Object with its OSM Tags



OSM Tag Key	OSM Tag Value
alt_name	IWM North
building	yes
name	Imperial War Museum North
tourism	museum
wikipedia	en:Imperial War Museum North

**b** Top 20 Annotations by Frequency Based Methods

TF	TF-IDF	TF-per-tweet-IDF	TF-per-user-IDF
war <sup>1</sup>	imperial <sup>1</sup>	imperial <sup>1</sup>	imperial <sup>1</sup>
museum <sup>1</sup>	war <sup>1</sup>	war <sup>1</sup>	war <sup>1</sup>
imperial <sup>1</sup>	museum <sup>1</sup>	museum <sup>1</sup>	museum <sup>1</sup>
north <sup>1</sup>	north <sup>1</sup>	north <sup>1</sup>	north <sup>1</sup>
manchester	salford	salford	salford
salford	MrBaizen	Mrbaizen	manchester
greater	OnEuropeTour	Oneuropetour	SalfordQuays
MrBaizen	manchester	manchester	SalfordQuays
OnEuropeTour	SalfordQuays	SalfordQuays	greater
SalfordQuays	greater	greater	iwm <sup>1</sup>
posted	iwm <sup>1</sup>	iwm <sup>1</sup>	mediacity
im	travel	travel	posted
travel	mediacity	mediacity	im
photo	posted	posted	photo
food	Imperial	Imperial	umbrella
mediacity	WarMuseumManchester <sup>1</sup>	WarMuseumManchester <sup>1</sup>	quay
iwm	im	im	architecture <sup>2</sup>
quay	UniversityOfManchester	UniversityOfManchester	Wellington Bomber <sup>2</sup>
UniversityOfManchester	uom	uom	GunTurret <sup>2</sup>
Imperial	photo	photo	julandhur
WarMuseumManchester <sup>1</sup>	umbrella	umbrella	
ball	quay	quay	iwmn <sup>1</sup>

**c** Manual Evaluation for the Relevance of Annotation Words

Relevance	Annotation Words
<sup>1</sup> Explicitly relevant	'war', 'museum', 'imperial', 'north', 'iwm', 'ImperialWarMuseumManchester', 'iwmn'
<sup>2</sup> Implicitly relevant	'architecture', 'WellingtonBomber', 'GunTurret'
NOT relevant	All other words are not relevant, e.g., the city name 'manchester' (where the museum is located) is too general to be a relevant annotation.

**Fig. 10** Interpretation and evaluation of tweets based semantic annotations for Imperial War Museum North (boundary based geographical object)

**a A Point Object with its OSM Tags**



OSM Tag Key	OSM Tag Value
historic	monument
name	Tower Bridge
tourism	attraction
wikipedia	en: Tower Bridge

**b Top 20 Annotations by Both Frequency Based and Probability Based Methods**

TF	TF-per-user-IDF	KDE-fixed	KDE-adaptive
bridge <sup>1</sup>	tower <sup>1</sup>	LondonRiviera	TowerBridge <sup>1</sup>
tower <sup>1</sup>	bridge <sup>1</sup>	TowerOfLondon	StKatharineDocks
london	london	TowerBridge <sup>1</sup>	TheScoop
TowerBridge <sup>1</sup>	TowerBridge <sup>1</sup>	DesignMuseum	GlassWalkWay <sup>2</sup>
photo	greater	MoreLondon	GlassFloor <sup>2</sup>
greater	thames	TowerHill	artigram
im	im	brigde <sup>1</sup>	katharines
posted	posted	20fenChurchStreet	fowd
day	photo	TheGherkin	bermondsey
uk	uk	WalkieTalkie	wihs
thames	day	LondonBridge	LondonRiviera
england	WalkWay <sup>2</sup>	ShardView	MoreLondon
city	view	Bermondsey	SuperYacht
morning	england	TheShard	CityHall
view	tourist	CityHall	TheGherkin
night	londres	shard	LondonBridge
love	city	gherkin	CheeseGrater
time	morning	bflofaniko	Tamise
beautiful	river	CheeseGrater	WalkieTalkie
londres	travel	ThePetCoach	RenzoPiano

**c Manual Evaluation for the Relevance of Annotation Words**

Relevance	Annotation Words
<sup>1</sup> Explicitly relevant	'bridge', 'tower', 'TowerBridge'
<sup>2</sup> Implicitly relevant	'WalkWay', 'GlassWalkWay', 'GlassFloor'
NOT relevant	All other words are not relevant, e.g., 'TowerOfLondo' and 'LondonBridge' are two nearby landmarks with similar names to Tower Bridge.

**Fig. 11** Interpretation and evaluation of tweets based semantic annotations for London Bridge (point based geographical object)

**7.7 Textural feature comparison between unigram, bigram and trigram**

The above experiments focus on unigrams to annotate geographic objects. To evaluate whether our results are consistent for different types of ngram features, we extract the unigrams, bigrams, or trigrams from geo-tagged tweets in name detection experiments for boundary based places shown in Fig. 9. We use six frequency based methods to detect the

names of places that contain tweets with at least 20 distinct unigrams, bigrams or trigrams. The ngram (unigram, bigram, or trigram) names should also be mentioned in the boundaries of the places. In the end, the ground truth datasets contain 3,106 place objects for unigram experiments (Fig. 9a), 2,228 place objects for bigram experiments (Fig. 9b), and 641 place objects for trigram experiments (Fig. 9c).

The accuracies of unigram experiments are higher than bigram and trigram for all types of place categories (Fig. 9d). Such results may come from the limit of ground truth datasets. As we have to exclude the places with only one word name for bigram experiments and trigram experiments, the remaining place objects with at least one bigram or trigram name may contain more noise and are harder to rank the place names in the top 10 annotations.

In contrast to the results of unigram experiments (Fig. 9a), we also find that, among the six frequency based methods, the bigram and trigram experiments have highest accuracies for the methods without document corpus based weighting, i.e., TF, TF-per-tweet, and TF-per-user (Fig. 9b–c). In general, bigrams and trigrams may contain more semantic information. The results indicate that we should use more customized methods for different types of features to explore the rich semantics of tweet content and many other geo-tagged social media data.

## 8 Conclusion

Vast amounts of spatial big data are being increasingly generated through geocrowdsourcing (VGI) and active users (social media). Integrating multiple sources of spatial big data could provide new insights and create new forms of value. In this paper, we present integrated spatial data analytics to support geo-tagged tweets based annotation for OpenStreetMap objects. Our spatial data integration is built on a MapReduce based spatial query engine which makes it possible to quickly integrate large scale spatial data. We first propose frequency based methods optimized through various weighting schemes to annotate objects with clear boundaries, and then propose probability based methods based on KDE optimized with adaptive bandwidth to annotate objects with point based representations. Our experiments from ground-truth comparison and human interpretation of annotation results demonstrate promising results.

## References

1. Aji A, Sun X, Vo H, Liu Q, Lee R, Zhang X, Saltz J, Wang F (2013) Demonstration of hadoop-gis: a spatial data warehousing system over mapreduce. In: SIGSPATIAL/GIS
2. Aji A, Vo H, Wang F (2015) Effective spatial data partitioning for scalable query processing. coRR
3. Aji A, Wang F (2012) High performance spatial query processing for large scale scientific data. In: SIGMOD/PODS 2012 PhD symposium
4. Aji A, Wang F, Vo H, Lee R, Liu Q, Zhang X, Saltz J (2013) Hadoop-GIS: a high performance spatial data warehousing system over mapreduce. In: Proc VLDB Endow
5. Bast H, Storandt S, Weidner S (2015) Fine-grained population estimation. In: SIGSPATIAL/GIS
6. Breiman L, Meisel W, Purcell E (1977) Variable kernel estimates of multivariate densities. *Technometrics*
7. Brinkhoff T, Kriegel H-P, Seeger B (1996) Parallel processing of spatial joins using r-trees. In: ICDE
8. Coffey C, Pozdnoukhov A (2013) Temporal decomposition and semantic enrichment of mobility flows. In: SIGSPATIAL/GIS Workshop LBSN
9. Georgiev P, Noulas A, Mascolo C (2014) The call of the crowd: event participation in location-based social services. In: AAAI conference
10. Georgiev P, Noulas A, thrive C, Mascolo. (2014) Where businesses predicting the impact of the olympic games on local retailers through location-based services data. In: AAAI conference

11. Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal*
12. Jurgens D, McCorrison J, Xu YT, Ruths D (2015) Geolocation prediction in twitter using social networks: a critical analysis and review of current practice
13. Karamshuk D, Noulas A, Scellato S, Nicosia V, Mascolo C (2013) Geo-spotting: mining online location-based services for optimal retail store placement. In: *ACM SIGKDD, ACM*
14. Krumm J, Horvitz E (2015) Eyewitness: Identifying local events via space-time signals in twitter feeds. In: *SIGSPATIAL/GIS*
15. Lee R, Wakamiya S, Sumiya K (2013) Urban area characterization based on crowd behavioral lifelogs over twitter. *Personal and ubiquitous computing*
16. Li Y, Steiner M, Wang L, Zhang Z-L, Bao J (2013) Exploring venue popularity in foursquare. In: *INFOCOM, 2013 Proceedings IEEE*
17. Lichman M, Smyth P (2014) Modeling human location data with mixtures of kernel densities. In: *SIGKDD*
18. Quattrone G, Capra L, De Meo P (2015) There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In: *CSCW*
19. Quercia D, Aiello LM, Schifanella R, Davies A (2015) The digital life of walkable streets. In: *WWW*
20. Quercia D, Schifanella R, Aiello LM, McLean K (2015) Smelly maps: The digital life of urban smellscapes. *ICWSM*
21. Sengstock C, Gertz M (2012) Latent geographic feature extraction from social media. In: *SIGSPATIAL/GIS*
22. Silverman BW (1986) *Density estimation for statistics and data analysis*. Chapman & Hall, London
23. Thomee B, Rae A (2013) Uncovering locally characterizing regions within geotagged data. In: *WWW*
24. Vo H, Aji A, Wang F (2014) Sato: a spatial data partitioning framework for scalable query processing. In: *SIGSPATIAL/GIS*
25. Wu F, Li Z, Lee W-C, Wang H, Huang Z (2015) Semantic annotation of mobility data using social media. In: *WWW*



**Xin Chen** is a Ph.D. candidate at Department of Biomedical Informatics at Stony Brook University. He received his M.S. in Biomedical Engineering from Chinese Academy of Medical Sciences, China, and B.S. in Biomedical Engineering from Huazhong University of Science and Technology, China. His major research interests include machine learning, social media and biomedical informatics.

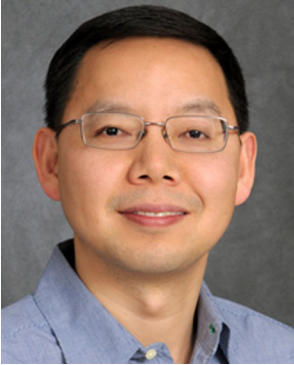


**Hoang Vo** is a Ph.D. candidate at Department of Computer Science at Stony Brook University. He received his B.S. in Information System from University of Wisconsin-La Crosse.



**Yu Wang** is a Ph.D. candidate at Department of Computer Science, Stony Brook University, USA. She received her B.Eng. in Computer Science from Wuhan University, China, in 2010. Her current research is about clinical data analytics with a particular focus on frequent sequence mining from large scale electronic healthcare records.





**Fusheng Wang** is an assistant professor at Department of Biomedical Informatics and Department of Computer Science at Stony Brook University. He received my Ph.D. in Computer Science from University of California, Los Angeles, and M.S. and B.S. in Engineering Physics from Tsinghua University, China. Prior to joining Stony Brook University, he was an assistant professor at Emory University. He was a research scientist at Siemens Corporate Research (Princeton, NJ) before joining Emory University. He is a senior member of the International Society for Optics and Photonics (SPIE). He received an NSF CAREER award in 2014.