Location Estimation Using Crowdsourced Spatial Relations

GEORGIOS SKOUMAS, National Technical University of Athens
DIETER PFOSER, George Mason University
ANASTASIOS KYRILLIDIS, University of Texas at Austin
TIMOS SELLIS, Swinburne University of Technology

The "crowd" has become a very important geospatial data provider. Specifically, nonexpert users have been providing a wealth of quantitative geospatial data (e.g., geotagged tweets or photos, online). With spatial reasoning being a basic form of human cognition, textual narratives expressing user travel experiences (e.g., travel blogs) would provide an even bigger source of geospatial data. Narratives typically contain qualitative geospatial data in the form of objects and spatial relations (e.g., "St. John's church is to the *North of* the Acropolis museum." The scope of this work is (i) to extract these spatial relations from textual narratives, (ii) to quantify (model) them, and (iii) to reason about object locations based only on the quantified spatial relations. We use information extraction methods to identify toponyms and spatial relations, and we formulate a quantitative approach based on distance and orientation features to represent the latter. Probability density functions (PDFs) for spatial relations are determined by means of a greedy expectation maximization (EM)-based algorithm. These PDFs are then used to estimate unknown object locations. Experiments using a text corpus harvested from travel blog sites establish the considerable location estimation accuracy of the proposed approach on synthetic and real-world scenarios.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Location estimation, spatial relations, crowdsourced geospatial data

ACM Reference Format:
Georgios Skoumas, Dieter Pfoser, Anastasios Kyrillidis, and Timos Sellis. 2016. Location estimation using crowdsourced spatial relations. ACM Trans. Spatial Algorithms Syst. 2, 2, Article 5 (June 2016), 23 pages.
DOI: http://dx.doi.org/10.1145/2894745

1. INTRODUCTION

Off-the-shelf geospatial information services are typically based on quantitative, coordinate-based data: maps are generated to answer geospatial questions such as "Where is the Monastiraki Metro Station (Athens)" based on their accurate descriptive

The research leading to these results received funding from the EU FP7 project GEOSTREAM (http://geocontentstream.eu/geostream/), grant agreement FP7-SME-2012-315631 and NGA NURI grant HM02101410004.
Authors' addresses: G. Skoumas, School of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytechniou 9, Politechnioupoli Zographou, 15780 Athens, Greece; email: gskoumas@dblab.ece.ntua.gr; D. Pfoser, Department of Geography and Geoinformation Science, George Mason University, 4400 University Drive, MS 6C3 Fairfax, VA 22032, USA; email: dpfoser@gmu.edu; A. Kyrillidis, Department of Computer Science, University of Texas at Austin, TX 78712, USA; email: anastasios@utexas.edu; T. Sellis, School of Software and Electrical Engineering, Swinburne University of Technology, PO Box 218 Hawthorn, Victoria 3122, Australia; email: tsellis@swin.edu.au.

DOI: http://dx.doi.org/10.1145/2894745

information. For instance, if such information appears in the precompiled database, this implies that we know—within some accuracy—the coordinates of the Monastiraki Metro Station. Such geospatial information services have been shown to be extremely useful tools across many disciplines (see Krisp [2013] and Schiller and Voisard [2004]) ranging from natural research management to transportation planning and from public information services to land use mapping. One of the reasons for the success of such services is the perceived accuracy of the information provided. It is generated by communities of mappers who contribute and maintain geospatial datasets. The resulting services provide precise answers to geographic queries based on collected quantitative information. However, the generation, processing, and preservation of such quantitative data is costly and time consuming. Although technology has helped to facilitate such geospatial data collection (e.g., all smart phones are equipped with GPS positioning sensors), curating quantitative data requires constant supervision and control to maintain a quality of service.

User-generated content has benefited many scientific disciplines (see Heipke [2010], Pfoser [2011], Sui et al. [2012], and Arsanjani et al. [2015]) by providing a wealth of new data sources. When generating geospatial data, most users are much more comfortable authoring qualitative geospatial data, as people typically do not use coordinates to describe their spatial experiences (trips, etc.) but rely on qualitative concepts in the form of toponyms (landmarks) and spatial relationships (near, next, north of, etc.). Thus, exploiting qualitative geospatial data (i.e., what "*North of* the Acropolis museum" means in terms of real coordinates) is a challenging user-generated content case.

In this article, we consider supervised learning methods for quantifying qualitative spatial relations, such as "North of," "near to," or "next to," to solve the following problem.

PROBLEM: *Given a set of objects P_V with a priori known coordinates in space, a set of objects P_U whose exact positions are unknown, and a set of predefined spatial relationships R between P_U and P_V objects, find probabilistic estimates for the positions of P_U objects in space.*

To better motivate the problem, consider the following narrative: "The Acropolis Pita place is *next to* the Monastiraki Metro Station." One of the challenges here in spatially understanding the scenario is the uncertainty associated with this statement. The spatial expression ("next to") might be interpreted differently by the various users. For example, it is apparent that the relation "next to" does not imply any orientation ("west of," "east of," etc.). Nevertheless, given a predefined grid of points over and around the Monastiraki Metro Station, we desire to associate probabilities to each location on the grid as candidate positions of the Acropolis Pita place. The probabilities assigned are drawn according to probabilistic models, trained and learned using user-generated texts including the "next to" relation in a region relatively close to the point of interest (POI). As such, we want to quantify what people imply when they say "next to." Being able to do so might allow us to actually discover the Acropolis Pita place (Figure 1 provides a toy example explanation). Eventually, by collecting more observations that mention the Acropolis Pita place using qualitative spatial relations, we will be able to refine the unknown location and thus locate places that otherwise could not be geocoded.

From the preceding discussion, it is obvious that the problem at hand involves high uncertainty, especially when the geospatial information source is user generated. Moreover, the transition from textual data (travel blogs) to location estimation of unknown POIs based on crowdsourced spatial relations is not at all straightforward. We therefore use a text mining preprocessing step. We employ natural language processing (NLP) tools and algorithms to extract spatial entities (POIs) and spatial relations between them (see Section 2). Following a probabilistic approach, we quantify the extracted spatial relations as PDFs. We first learn spatial relation models between known

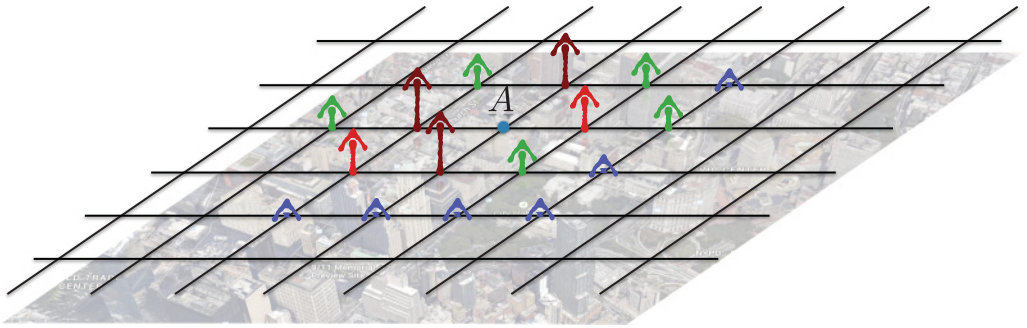


Fig. 1. Quantifying a spatial relation. Point A corresponds to a known reference POI (Monastiraki Metro Station), and the arrows on the predefined grid correspond to probabilities of spots being the "Acropolis Pita place *next to*" point A. The higher the arrow, the higher the probability that the location of interest (Acropolis Pita place) lies at the corresponding grid point.

POIs (i.e., focused over a predefined region of interest such as Athens) and given POI pairs with known locations linked together by a specific spatial relation, we train the corresponding spatial relation PDF comprised of distance and orientation (see Section 3). Here a greedy expectation maximization (EM)-based algorithm is used. The trained probabilistic models can then be used for location estimation tasks.

Given a specific spatial relation instance of the form (P_u, R_o, P_v) and by employing the trained model for spatial relation R_o, we can associate probabilities with each point on the discretized space. These probabilities are then used to estimate unknown POI locations. The more observations we have with respect to an unknown location, the more precise will be the estimate of the unknown POI's location. Actual location estimation experiments using textual narratives from travel blogs establish the validity and quality of the proposed approach.

Our contributions can be summarized as follows:

(i) We quantify qualitative spatial relations using a probabilistic path as presented in our previous work (see Skoumas et al. [2013]).
(ii) We propose a grid-based algorithm that performs location estimation based on the aforementioned probabilistic models.
(iii) We evaluate our location prediction algorithm with extended experiments on both synthetic and real-world location prediction scenarios.

2. TEXTUAL NARRATIVES AND QUALITATIVE SPATIAL RELATIONS

The resource in this work is textual narrative, and this section describes the basic preprocessing steps needed to extract qualitative spatial data in the form of relations from it. In terms of content sources, we focus on travel blogs as a potentially rich geospatial data source. This selection is based on the fact that people tend to describe their experiences in relation to their location, which results in "spatial narratives." To gather such data, we use classical Web crawling techniques [Drymonas and Pfoser 2010] and compile a database consisting of 250,000 texts obtained from 20 travel blogs.

Obtaining qualitative spatial relations from texts involves the detection of (i) spatial objects (i.e., POIs or toponyms) and (ii) spatial relations linking the POIs. Our approach involves geoparsing (i.e., the detection of candidate phrases) and geocoding (i.e., linking parts of the phrase/toponym to actual coordinate information).

2.1. Textual Narratives, POIs, and Spatial Relations

The extraction of qualitative geospatial data from texts requires the utilization of efficient NLP tools to automatically extract and map phrases to spatial relations. In the

past, the extraction of semantic relations between entities in texts has been developed in Bunescu and Mooney [2006], Fader et al. [2011], Akbik and Broß 2009, Mesquita et al. [2013], and Zelenko et al. [2002], whereas extraction of spatial relations (mostly topological relations) from texts is analyzed in Kordjamshidi et al. [2011], Yuan [2011], Loglisci et al. [2012], and Zhang et al. [2011], and recently in Wallgrün et al. [2014]. Although the preceding works are related to the current effort and problem setting, we implemented a specialized spatial relation extractor that addresses the problem of noisy crowdsourced data. The implementation is based on the Natural Language Processing Toolkit (NLTK) [Loper and Bird 2002], a popular Python natural language processing library.

Using NLTK, we managed to initially extract approximately 500,000 POIs[1] from our travel blog corpus. For the geocoding of the POIs, we rely on the GeoNames[2] geographical gazetteer, which has global coverage and contains more than 10 million places. This procedure associates (whenever possible) geographic coordinates with POIs found in the travel blogs using string matching based on the Levenshtein string distance metric [Hirschberg 1997]. Using the GeoNames gazetteer, we managed to geocode about 480,000 out of the 500,000 extracted POIs.[3]

Having detected and geocoded sets of POIs, the next step is to extract spatial relations. ReVerb [Fader et al. 2011] and EXEMPLAR [Mesquita et al. 2013] are available software tools for the extraction of semantic (not necessarily spatial) relations between identified entities in texts. However, it was not clear that these methods would perform well at the task of spatial relation extraction, which motivated us to develop a task-specific supervised spatial relation extraction method based on NLTK [Loper and Bird 2002] components and predefined strings and tag sequences. Specifically, we have a manually annotated dataset (see Skoumas et al. [2013])—that is, 12,000 sentences with two or more POIs, which result in about 1,500 "clean" spatial relation instances between POIs. We use 80% of the clean spatial relation instances with NLTK to extract the part-of-speech tag sequences, which contain POIs and spatial relations. Each extracted tag sequence becomes a "rule," which we use to extract spatial relations. This is a very common information extraction technique discussed in Chapter 7 of Loper and Bird [2002]. It turns out that the use of both tag sequences and string matching reduces the number of false positives considerably.[4]

---

[1]The POI extraction procedure contains three steps. Initially, we use NLTK's word tokenizer to tokenize each sentence. NLTK's maximum entropy part-of-speech tagger is used with the Penn Treebank tag set trained over the Penn Treebank corpus to obtain part-of-speech tags for each token of each sentence. Finally, we use NLTK's maximum entropy chunker for the named entity (POIs) recognition task. Named entities are definite noun phrases that refer to specific types of individuals, such as organizations, persons, and dates. In our case, the chunker will *only* recognize POIs that are mentioned via a proper name (e.g., "Monastiraki Metro Station"). The chunker is trained on the ACE corpus, and we extract locations, organizations, facilities, and geopolitical entities (GPEs) as they are described in Chapter 7 of Loper and Bird [2002].

[2]http://www.geonames.org/.

[3]Today, there are many available geotaggers (Google Geocoding API, Yahoo Geocoding API). Unfortunately, all of them have a small free geocoding quota, which becomes a serious problem when one needs to geocode a large number of POIs. A workaround is the use of gazetteers and custom solutions. This is why we also used the GeoNames gazetteer. Another problem that arises in geocoding is named entity disambiguation, typically characterized by more than one POI with the same name in a gazetteer. Disambiguation is simpler in our case, as we only consider sentences that contain two or more POIs, and we geocode a POI only if it is included in an extracted triplet of the form (POI_1, *Relation*, POI_2). If two or more candidates exist for each POI of a triplet, we use the minimum kilometric distance feature [Pouliquen et al. 2006]—that is, we keep the candidates that are closest in space and within a limit of 20km (city scale).

[4]As an example, consider the following phrase. "*Deutsche Bank* invested 10 million dollars *in Brazil*." Here, a simple string matching solution would extract a triplet of the form (*Deutsche Bank*, *in*, *Brazil*), which is a false positive. In our approach, the use of predefined tag sequences avoids this kind of mistake. On the other hand, for the phrase "*Deutsche Bank* invested 10 million dollars *in Rio de Janeiro*, which is *in Brazil*,"

Table I. Precision and Recall for Three Different Spatial Relation Extraction Approaches

Method	Precision	Recall
EXEMPLAR [Mesquita et al. 2013]	**0.71**	0.40
ReVerb [Fader et al. 2011]	0.15	0.43
NLTK [Loper and Bird 2002]	0.60	**0.82**

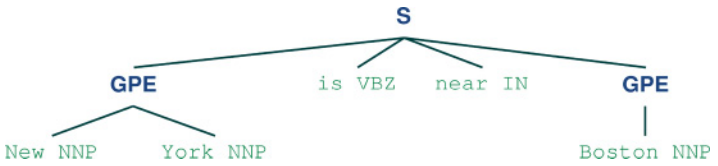


Fig. 2. Example of a POS-tagged word sequence with embedded named entities.

The quality of the various algorithms is summarized in Table I. Having the three algorithms to extract the relationships from the remaining 300 (20%) manually annotated sentences shows that our NLTK-based method and EXEMPLAR perform better than ReVerb in terms of precision and recall. Although our NLTK approach seems to have slightly lower precision than EXEMPLAR, it achieves a higher recall (i.e., cases for which the fraction of extracted relations relevant to the query is larger).[5]

2.2. RELEX Algorithm

The resulting relation extraction (RELEX) algorithm (Algorithm 1) integrates POI and relation extraction as described earlier in a single method. Initially, the raw text document is segmented into sentences (step 3). Each sentence is further subdivided (tokenized) into words and tagged for part of speech (steps 5 and 6). Named entities (POIs) are identified (step 7). We typically look for relations between specified types of named entities, which in NLTK are organizations, locations, facilities, and geopolitical entities (GPEs). In case there are two or more named entities in a sentence, we check if any of the predefined tag sequences apply (step 12). If so, we check if a spatial relation instance from our relation catalog exists (string matching) (step 14). Should there be a match, we record the extracted triplet (steps 15 through 18). Thus, a result comprises a triplet O of the form (P_u, R_o, P_v), in which P_u and P_v are named entities of the required types and R_o is the observed spatial relation that relates to P_u and P_v.

A relation extraction example is shown in Figure 2. Here, the sentence "Boston is near New York" is analyzed as explained, and two named entities are identified as GPEs.

We first check if the tag sequence "GPE (set of NNPs)—present verbal form (VBZ)—preposition/subordinating conjunction (IN)—GPE (set of NNPs)" exists in our set of predefined spatial relation tag sequences. Performing string matching on the intermediate chunks ("near") results in the triplet (*New York*, *Near*, *Boston*).

Algorithm 1 extracted approximately 500,000 triplets from our travel blog corpus consisting of 250,000 texts. Figure 3 shows a small sample of a spatial relationship graph, a spatial graph in which nodes represent POIs and edges label spatial relationships existing between them. The graph visualizes a sample of the spatial relationship data collected for New York City. In this work, we extracted spatial relation data for

our NLTK-based algorithm would extract a triplet of the form (*Rio de Janeiro*, *in*, *Brazil*), which is a true positive.

[5]At this point, we highlight the fact that EXEMPLAR and ReVerb are not trained specifically for this task, which is the cause of their lower performance. Our NLTK-based method performs slightly better, as it has been trained to extract spatial relations only extracted from a noisy crowdsourced dataset.

ALGORITHM 1: RELEX—Spatial Relation Extraction

Input: A database of texts T, a set of tag sequences A, a set of spatial relation strings R
Output: A set of triplets O of the form (P_u, R_o, P_v), where $P_u \neq P_v$ and $R_o \in R$

```
1  begin
2      foreach text t ∈ T do
3          Extract sentences from t into set S
4          foreach sentence s ∈ S do
5              Token s using NLTK
6              PosTag s using NLTK
7              Identify named entities using NLTK
8              if two or more named entities in s then
9                  Extract POI pairs in P
10                 foreach p ∈ P do
11                     p_A ← Extract tag sequence of p
12                     if p_A ∈ A then
13                         p_R ← Extract string pattern of p
14                         if p_R ∈ R then
15                             P_u ← p(1)
16                             P_v ← p(2)
17                             R_o ← p_R
18                             O.PushTriplet(P_u, R_o, P_v)
19                         end
20                     end
21                 end
22             end
23         end
24     end
25     return O
26 end
```

four different cities: London, New York, Paris, and Beijing. All four city datasets will be used in the experimental evaluation of our location estimation approach in Section 5.

3. SPATIAL RELATION MODELING

To increase the usefulness of qualitative spatial data, it needs to be quantified—that is, translating expressions such as "near" to actual distances. Our proposal is to use probabilistic modeling for this task. This approach includes the selection and extraction of respective features (distance and direction), as well as the methods to train and optimize the probabilistic model.

3.1. Feature Extraction

We model a spatial relation between two POIs, P_u, P_v, in terms of distance and orientation features. Assuming a projected (Cartesian) coordinate system, the distance is computed as the Euclidean metric between the two respective coordinates, whereas the orientation is established as the counterclockwise rotation of the x-axis, centered at P_v, to point P_u. For a concise and consistent mathematical formalization, consider that for each instance of each relation, we create a two-dimensional *spatial feature vector* $X = (X_d, X_o)^\intercal$, where X_d denotes the distance and X_o denotes the orientation between P_u and P_v. Several instances of a spatial relation lead to a set of two-dimensional spatial feature vectors, which we denote as $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$. Each spatial feature vector set will be used to train one probabilistic model for each spatial relation.

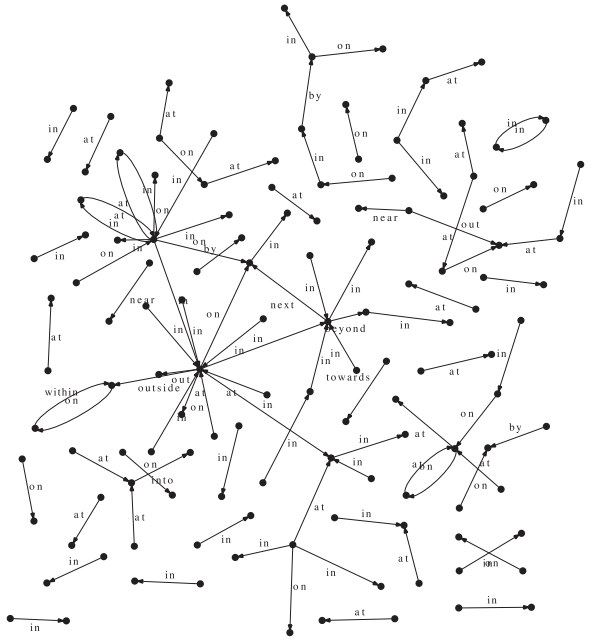


Fig. 3. Small sample of a spatial relationship graph for New York City.

An example of the feature extraction procedure is illustrated in Figure 4, where four instances of spatial relation *Near* are used to create the respective set of spatial feature vectors $\mathcal{X}_{near} = \{[X_{d1}, X_{o1}]^\intercal, [X_{d2}, X_{o2}]^\intercal, [X_{d3}, X_{o3}]^\intercal, [X_{d4}, X_{o4}]^\intercal\}$. In this scenario, $P_V = \{A, D, E, G\}$ is the set of reference points and $P_U = \{B, C, F, H\}$ is the set of points described based on the reference points.

3.2. Probabilistic Modeling

The next modeling step is the mapping of a feature vector representing spatial relations to preselected probability density functions (PDFs). Recent research on the quantitative representation of spatial knowledge has been conducted in relation to situational awareness systems, robotics, and image processing. Modeling uncertain spatial information for situational awareness systems is discussed in Kalashnikov et al. [2006] and Ma et al. [2009]. The authors propose a Bayesian probabilistic approach to model and represent uncertain event locations described by human reporters in the form of free text. Estimation of uncertain spatial relationships in robotics is addressed in Smith et al. [1990]. A probabilistic algorithm for the estimation of distributions over geographic locations is proposed in Hays and Efros [2008], where a data-driven scene matching approach is used to estimate geographic information based on images. Finally, image similarity based on quantitative spatial relationship modeling is addressed in Wang and Makedon [2003]. Based on the nature of the features extracted as explained in the previous section, we follow the modeling approach that we outlined in Skoumas et al. [2013].

Specifically, given the training data (e.g., a set of spatial feature vectors $\mathcal{X}$ for each spatial relation), we estimate a continuous density distribution using a Gaussian mixture model (GMM) for each of these relations. The intuition is that people use and understand spatial relation phrases differently. This results in multicomponent distributions of the features. Figure 5 illustrates a representative PDF example of distance

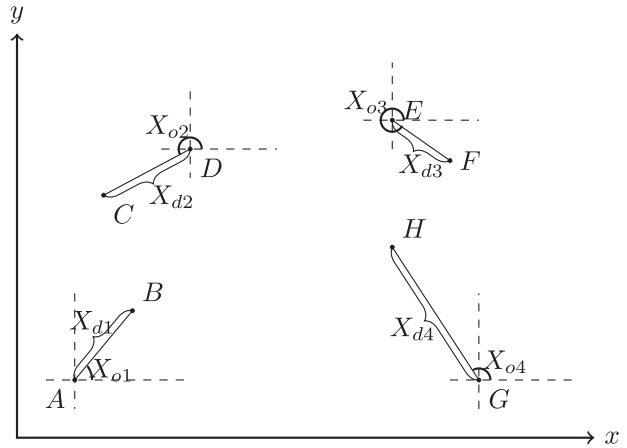


Fig. 4. Distance and orientation feature extraction procedure. In this case, the respective triplets are (*B*, *Near*, *A*), (*C*, *Near*, *D*), (*F*, *Near*, *E*), and (*H*, *Near*, *G*).

and orientation features for the spatial relation *South*, which further strengthens our intuition.

Moreover, in Li and Barron [1999], it is shown that for any heterogeneous multidimensional data that originates from an arbitrary PDF $p(\cdot)$, there exists a sequence of finite mixtures $p_k(x) = \sum_{i=1}^{k} w_i g(x; \theta_i)$ that achieves Kullback-Leibler (KL) divergence

$$D(p||p_k) - D(p||g_p) \leq \mathcal{O}(1/k)$$

for any $g_p = \int g(x; \theta) P(d\theta)$—that is, one can achieve a good approximation with rate $\mathcal{O}(1/k)$ by using a k-component mixture of $g(x; \cdot)$. Furthermore, this bound is achievable by employing a greedy training schema [Li and Barron 1999].

Other approaches to estimate continuous densities for spatial relations could be nonparametric. A well-known example of such an approach is kernel density estimation (KDE). In Section 5 of this work, we show that the GMM-based approach outperforms KDE in terms of location estimation accuracy.

In general, a GMM is a weighted sum of M-component Gaussian densities as $p(x|\lambda) = \sum_{i=1}^{M} w_i g(x; \mu_i, \Sigma_i)$, where x is a d-dimensional data vector (in our case $d = 2$), w_i are the mixture weights, and $g(x; \mu_i, \Sigma_i)$ is a Gaussian density function with mean vector $\mu_i \in \mathbb{R}^d$ and covariance matrix $\Sigma_i \in \mathbb{R}^{d \times d}$. To fully characterize f, one requires the mean vectors, the covariance matrices, and the mixture weights. These parameters are collectively represented in $\lambda = \{w_i, \mu_i, \Sigma_i\}$ for $i = 1, \ldots, M$.

In our setting, each spatial relation is modeled as a two-dimensional GMM, trained on each relation's spatial feature vector set. We assert that distance and orientation features are informative enough to model spatial relationships in a Cartesian context. For the parameter estimation of each Gaussian component of each GMM, we use EM (see Dempster et al. [1977]). EM enables us to update the parameters of a given M-component mixture with respect to a feature vector set $\mathcal{X} = \{X_1, \ldots, X_n\}$ with $1 \leq j \leq n$ and all $X_j \in \mathbb{R}^d$ such that the log-likelihood $\mathcal{L} = \sum_{j=1}^{n} \log(p(X_j|\lambda))$ increases with each re-estimation step (i.e., EM re-estimates model parameters λ until $\mathcal{L}$ convergence).

Finally, a main issue in probabilistic modeling with mixtures is that a predefined number of components is neither a dynamic nor an efficient and robust approach. The optimal number of components should thus be decided based on each dataset. In this work, we employ the greedy learning approach detailed in Skoumas et al. [2013] to

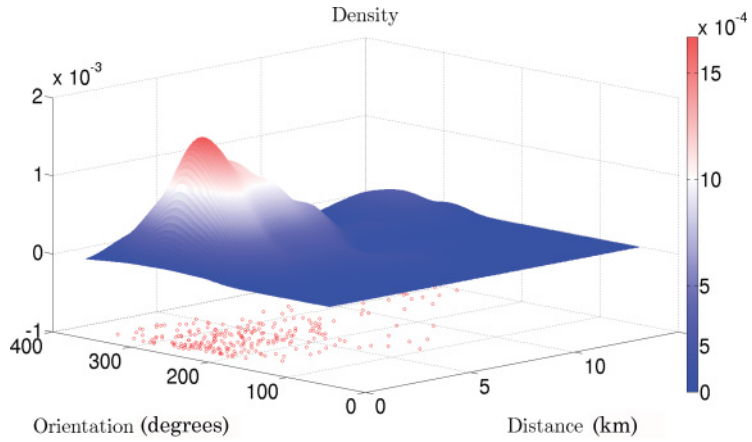


Fig. 5. Distance and orientation feature for the spatial relation *South* with the respective PDF.

dynamically estimate an optimized number of components in a GMM (see also Verbeek et al. [2003]).

At this point, we highlight that the proposed spatial relation modeling scheme is distribution independent. With the necessary tweaks, one can transform the schema to use mixtures of any distribution type. Such a selection is user defined. However, its application would be much harder in practice. We use mixtures of PDFs, as it was shown in Li and Barron [1999] that densities of heterogeneous and noisy data, such as our case of crowdsourced data, can be approximated by a sequence of finite mixtures. We particularly use GMMs due to their simplicity and their generally low classification errors (see Bishop [2006] and Duda et al. [2001]). The results that we obtain, and thoroughly analyze in Section 5, encourage their use in practice. Thus, GMMs provide a challenging baseline for potentially better mixture models to be explored in future work.

To visualize the actual models and the respective probabilities that they assign to partitioned space, Figure 6 depicts three instances of spatial relations, with the center of the grid denoting a reference (landmark) point. The examples have been derived from the New York, London, and Beijing datasets.

4. LOCATION ESTIMATION WITH SPATIAL RELATIONSHIP FUSION

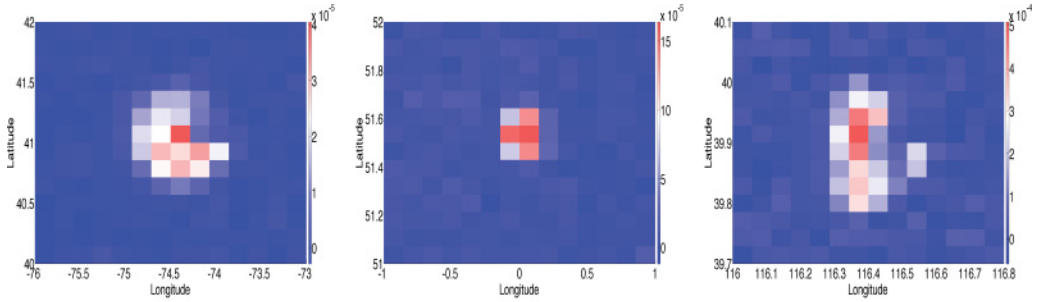
Location estimation in the context of this work refers to reasoning about object locations based on their spatial relations to known locations. Using the probabilistic models of spatial relations, we outline two approaches for solving this location estimation task.

4.1. Naive Method

Spatial relations can be considered as links connecting spatial objects. A simple approach one could follow to estimate the location of an unknown POI would be to calculate the mean location of the connected known POIs combined with a user-defined spatial extent, such as a 1km radius around the mean location. However, this would contain quite high uncertainty and would be close to a random selection, as it does not consider the semantics of the link (qualitative spatial relations). Experimentation will compare this naive approach to a grid-based method described in the following section.

4.2. Grid-Based Approach

Spatial relations essentially are observations that correlate objects in space. Having quantified them by employing probabilistic models allows us to reason about locations



(a) *Near* the grid center (b) *At* the grid center (c) *West* of grid center

Fig. 6. Spatial extension of spatial relationships—probabilities of specific spatial relationships (*Near*, *At*, *West*) relating vertices to the center grid cell.

taking uncertainty into account. Our goal now is to show how such probabilistic models can be employed in location estimation scenarios. Unknown locations can be estimated by fusing spatial relationship observations to known POIs (landmarks).

The qualitative location estimation (QLEST) algorithm (Algorithm 2) details our location estimation method for an unknown POI P_u and a given set of triplets $\mathcal{T}$ of the form (P_u, R_o, P_v) about P_u. Here, $P_v \in P_V$ is a landmark POI in a set of landmarks P_V. QLEST is a grid-based approach, where given a predefined grid of points over and around a landmark POI P_v, we desire to associate probabilities to each vertex on the grid as candidate positions of the unknown POI P_u. The probabilities assigned are drawn according to probabilistic models, trained and learned as described in Section 3.

ALGORITHM 2: QLEST—Qualitative Location Estimation

Input: A set of trained GMMs $\hat{\mathcal{G}}$, a bounding box $\mathcal{B}_\mathcal{B}$, grid dimensionality in $\mathcal{G}_\mathcal{D}$, an unknown POI $P_u$to locate, a set of landmark POIs P_V, and a set of triplets $\mathcal{T}$ of the form (P_u, R_o, P_v)
Output: Each region's (grid cell) likelihood $\mathcal{R}_\mathcal{L}$

begin
  $\mathcal{G}_\mathcal{V} \leftarrow$ Calculate grid vertices for $\mathcal{B}_\mathcal{B}$ based on $\mathcal{G}_\mathcal{D}$
  foreach $t \in \mathcal{T}$ **do**
    $\hat{\mathcal{G}}' \leftarrow$ Load GMM for spatial relation R_o of triplet tu
    foreach $gv \in \mathcal{G}_\mathcal{V}$ **do**
      $X \leftarrow [0, 0]$
      $X[1] \leftarrow$ Calculate distance between gv and P_v
      $X[2] \leftarrow$ Calculate orientation between gv and P_v
      $L_{gv} \leftarrow P(X|\hat{\mathcal{G}}')$
      $\mathcal{V}_\mathcal{L}$.PushLikelihood(L_{gv})
    end
  end
  $\mathcal{F}_{\mathcal{V}\mathcal{L}} \leftarrow$ Sum and normalize each vertex's likelihoods in $\mathcal{V}_\mathcal{L}$
  $\mathcal{R}_\mathcal{L} \leftarrow$ Calculate each region's likelihood using $\mathcal{F}_{\mathcal{V}\mathcal{L}}$
  return $\mathcal{R}_\mathcal{L}$
end

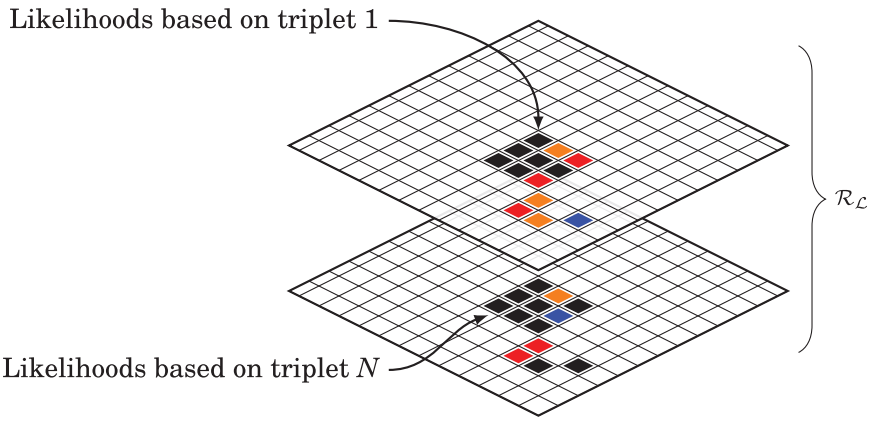


Fig. 7. Example of a QLEST run where grid likelihoods based on a set of N triplets $\mathcal{T}$ are fused.

Finally, all assigned probabilities are aggregated to define the overall probability of each grid cell.

The first step is to discretize space by partitioning it with respect to grid vertices (step 2). For example, for a grid dimensionality $\mathcal{G}_\mathcal{D} = 15$, we have $15 \times 15 = 225$ grid vertices and $14 \times 14 = 196$ grid cells (regions). Next, for each triplet t, we load the GMM relationship model $\hat{\mathcal{G}}'$, which corresponds to spatial relation R_o (step 4). Then we calculate the distance and orientation between each grid vertex gv and the respective landmark P_v of the loaded triplet and create a spatial feature vector X (steps 6 through 8). We calculate the probability of spatial vector X given the selected relationship model $\hat{\mathcal{G}}'$ and store it in matrix $\mathcal{V}_\mathcal{L}$ (steps 9 and 10). In this way, we assign likelihoods to each vertex of the grid for each given triplet t. Finally, all likelihoods per vertex are summed up and normalized in $\mathcal{F}_{\mathcal{V}\mathcal{L}}$ (step 11), and a probability is assigned to each grid cell (region) $\mathcal{R}_\mathcal{L}$ (step 12). The overall likelihood of a grid cell is calculated as the mean value of the likelihoods of its four vertices.

An example of a QLEST run is shown in Figure 7, where we estimate grid cell likelihoods for each triplet of a given set of N triplets. Finally, all grid cell likelihoods are fused to give the final region likelihoods $\mathcal{R}_\mathcal{L}$. The presented QLEST method, which fuses (spatial relationship) observations to estimate unknown point locations, will be used in the following experimental section in the context of synthetic and real-world location estimation scenarios.

5. LOCATION FUSION EXPERIMENTS

To assess the quality of the probabilistic spatial modeling and location estimation approach, we perform an extensive experimentation using synthetic and real-world location estimation scenarios. All text processing has been implemented in Python, whereas the relationship modeling and location estimation methods were implemented in Matlab.

5.1. Location Estimation for Synthetic Scenarios

We first assess location estimation using a synthetic data scenario. To generate POIs and spatial relations that connect them, we follow an approach similar to that of Algorithm 2—that is, we discretize space by partitioning it with respect to grid vertices $\mathcal{G}_\mathcal{V}$, which will be used as known POIs, and we then generate a random point representing an unknown POI P_u. For each grid vertex $gv \in \mathcal{G}_\mathcal{V}$, we calculate the distance

and orientation from P_u and create the spatial feature vector X. Finally, we pick the spatial relationship GMM $\hat{\mathcal{G}}'$ that maximizes the likelihood of X. Under a mathematical formalization, this means that $\hat{\mathcal{G}}'$ is picked as $\hat{\mathcal{G}}' \leftarrow \arg\max_{g \in \hat{\mathcal{G}}} P(X|g, gv)$, where $\hat{G}$ is again a set of trained spatial relation GMMs. Thus, for each random point P_u, we generate a set of triplets of the form (P_u, R_o, P_v), with P_v always being a grid vertex (the number of triplets is equal with the number of vertices). Following this procedure, we generate 1,000 location prediction scenarios for each of our four city datasets.

To provide some baseline results, we define a one-component baseline (BSL) model, which is a GMM model $p(x|\lambda) = \sum_{i=1}^{M} w_i g(x; \mu_i, \Sigma_i)$ with $M = 1$ (simple Gaussian distribution). We will models—the BSL model, the KDE model, and our optimized (optimized number M of Gaussian components) GMM model (OPT)—trained as described in Section 3.2.

Figure 8 illustrates the approach by means of four (very challenging) examples in the Beijing area. The red stars in the first column (Figure 8(a), (c), (e), (g)) illustrate the random points that were generated. The second column (Figure 8(b), (d), (f), (h)) show the assigned probabilities of each region after a full run of Algorithm 2. We observe that our approach assigns the highest probability to the grid cell where the random point was generated for the first two cases (Figure 8(a) through (d)), and the prediction is also accurate (i.e., the are high probabilities around the generated points) for the other two cases (Figure 8(e) through (h)).

Additionally, for the 1,000 generated points, we consider the cases in which the randomly generated point's region is among the top-k predicted regions with $k = \{1, 5, 10, 20\}$, respectively. The prediction accuracy results are shown in Figure 9. The results show the superior performance of the OPT model against both BSL and KDE models, with the KDE model performing slightly better than the BSL model. Additionally, Tables II and III show the actual prediction accuracy improvement of the OPT model with respect to the BSL model and the KDE model, respectively. In some cases (indicated in bold), the prediction accuracy improvement is equal to or greater than 30%.

Finally, based on the qualitative spatial relation rules between points described in Papadias and Sellis [1994], we measure the percentage of selected models $\hat{\mathcal{G}}'$ that are qualitatively correct—that is, they reveal a true spatial relation between a vertex and a random point. Figure 10 shows the percentage of qualitatively correct models $\hat{\mathcal{G}}'$. The percentage of OPT models is considerably higher than that of KDE and BSL models (with the KDE model performing slightly better than the BSL model in most cases). Table IV shows this improvement in relative terms. In some cases (indicated in bold), the qualitative accuracy improvement is more than 10%. The inherited noise in crowdsourced data is the reason for the error that is evident in some cases. Specifically, some of the models are similar (i.e., they return similar probabilities), although they express qualitatively different spatial relations.

5.2. Location Estimation for Real-World Scenarios

The ultimate goal of this work is to train probabilistic models so as to provide location estimates for cases, such as finding the the Acropolis Pita place, based on relative qualitative spatial relationships discovered in textual narratives. This is a potentially important method, as it provides a solution to the geocoding problem that exists for user-contributed data on the Web—that is, there are myriad POIs and locations whose coordinates do not exist in any gazetteer.

In addressing this challenge, we present extensive location estimation experiments on 3,000 real-world scenarios extracted from all four datasets as they were presented in Section 5.1. (about 800 real scenarios per region). We extract 3,000 POIs (considered

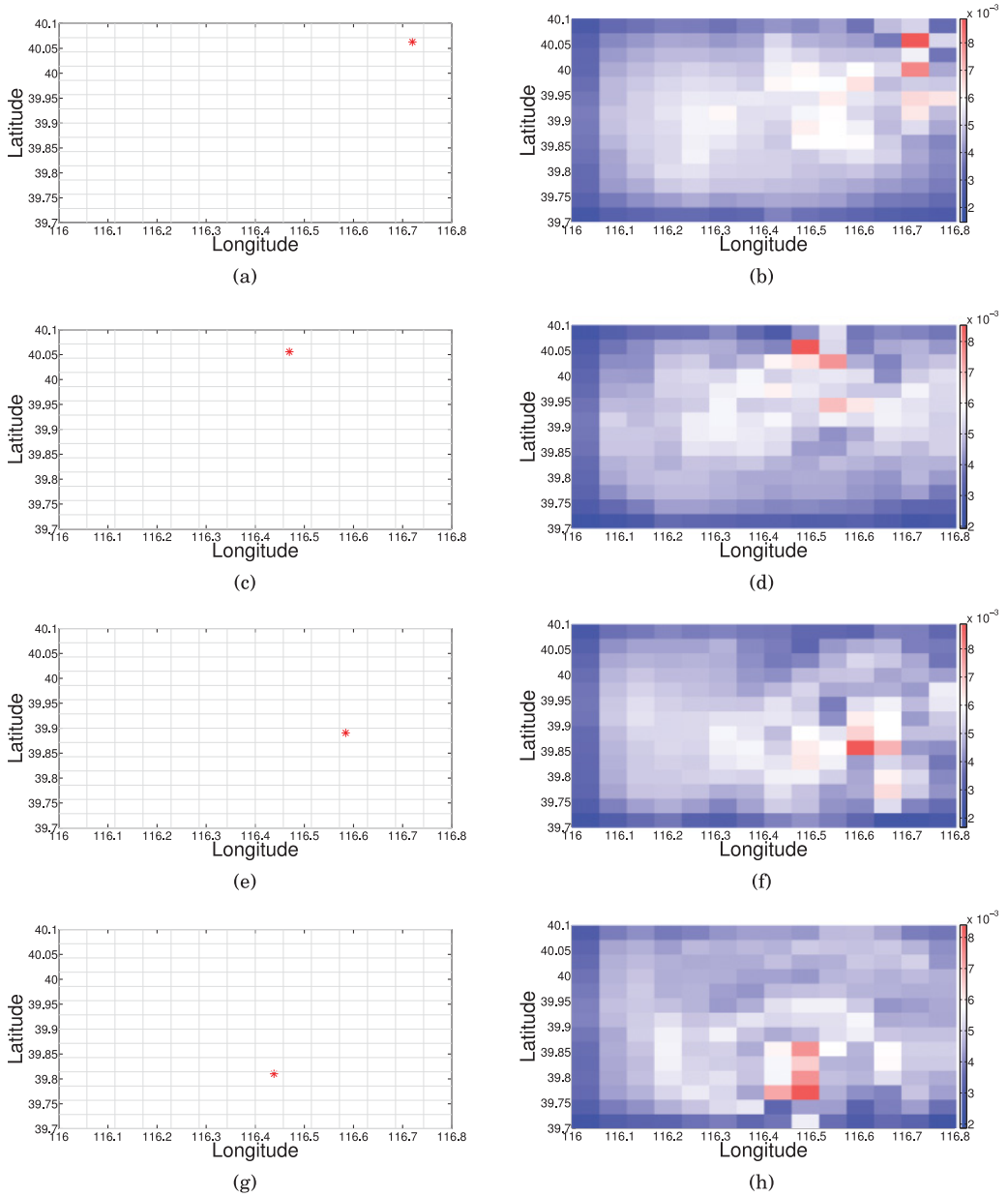


Fig. 8. Synthetic location estimation scenarios in the Beijing area. The first column ((a), (c), (e), and (g)) shows the generated point with a red star. The second column ((b), (d), (f), and (h)) shows the probability of each region after a full run of Algorithm 2 using heatmap colors.

as unknown) whose locations are given in (spatial) relation to known POIs. Note that these cases have *not* been used in the training phase. The experiments will also show the impact of the number of components per spatial relationship model on the quality of the location estimation outcome (e.g., the performance improvement when using the OPT instead of the BSL or KDE models). The spatial relation fusion procedure in the

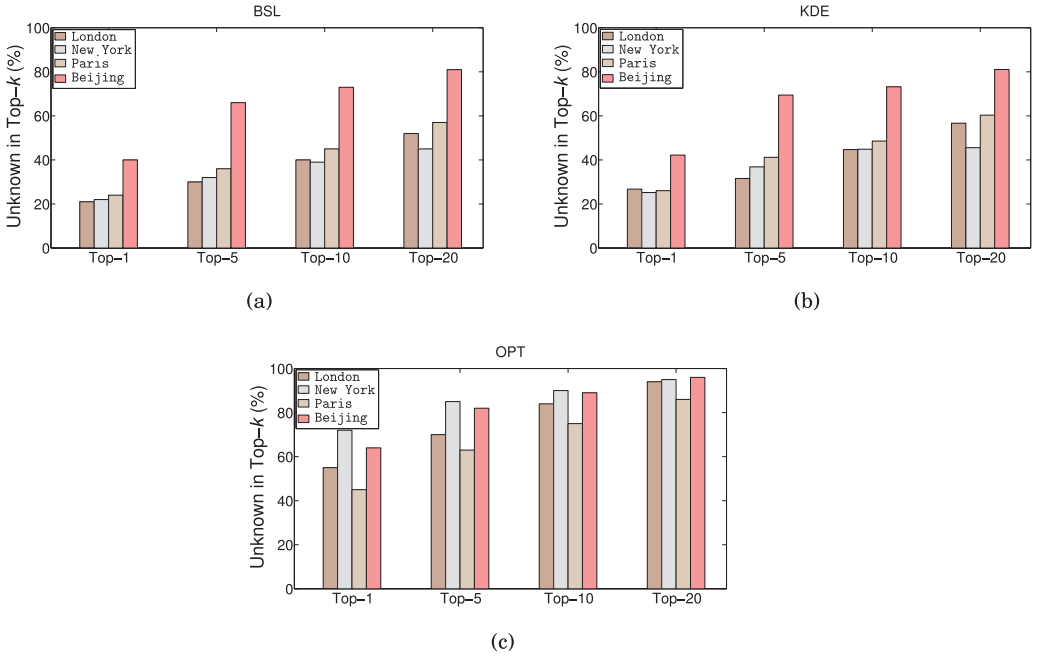


Fig. 9. Location prediction accuracy: prediction accuracy of the BSL model (a), the KDE model (b), and the OPT model (c) for k values 1, 5, 10, and 20, respectively.

Table II. Prediction Accuracy Improvement of the Optimized Model (OPT) Compared to the Baseline Model (BSL)

	OPT Improvement per Top-k Case			
Dataset	$k = 1$	$k = 5$	$k = 10$	$k = 20$
London	+**34**%	+**40**%	+**44**%	+**42**%
New York	+**50**%	+**53**%	+**51**%	+**50**%
Paris	+21%	+27%	+**30**%	+29%
Beijing	+24%	+16%	+16%	+15%

Table III. Prediction Accuracy Improvement of the Optimized Model (OPT) Compared to the Kernel Density Estimate Model (KDE)

	OPT Improvement per Top-k Case			
Dataset	$k = 1$	$k = 5$	$k = 10$	$k = 20$
London	+28%	+**38**%	+**39**%	+**37**%
New York	+**46**%	+**48**%	+**45**%	+**49**%
Paris	+18%	+21%	+**36**%	+25%
Beijing	+21%	+13%	+15%	+15%

real-world scenarios is the same as that presented in Algorithm 2. It only differs in that the reference landmarks are actual POIs (not grid vertices) extracted from textual narratives, and that we use the observed, in text, spatial relation model instead of the selected model $\hat{\mathcal{G}}'$.

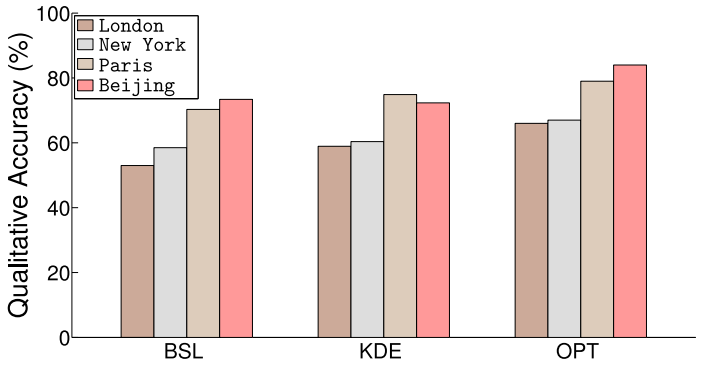


Fig. 10. Percentage of qualitatively correct spatial relation models for the BSL, KDE, and OPT models.

Table IV. Qualitative Accuracy Improvement When Using the Optimized Model (OPT) Compared to the Baseline Model (BSL) and the Kernel Density Estimate Model (KDE)

	OPT Improvement	
Dataset	BSL	KDE
London	+8%	+6%
New York	+**11**%	+9%
Paris	+**13**%	+7%
Beijing	+9%	+4%

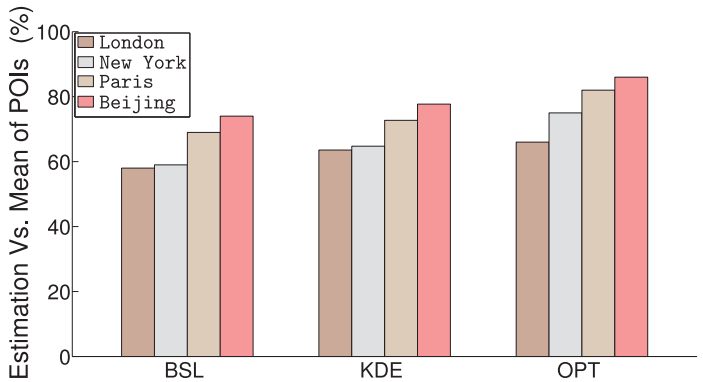


Fig. 11. Location prediction accuracy. Percentage of real scenarios—center of spatial probability density closer than the mean location of referenced POIs.

5.2.1. Naive Method. The first experiment estimates the location of an unknown POI using BSL, KDE, and OPT models and compares the result to the mean location of all referenced POIs (see Section 4.1). The results for all four datasets are given in Figure 11. The one-component BSL model provides better location estimation (the highest probability predicted point is closer to the unknown POI than the mean location) when compared to the mean location. It is closer to the actual POI's location in 56%, 57%, 70%, and 80% of the cases for the datasets of London, New York, Paris, and Beijing, respectively. The KDE model improves prediction accuracy by 3% to 5% in all cases—that is, 4%, 5%, 3%, and 4% for each respective dataset. Finally, the OPT model improves the results further by 12%, 20%, 16%, and 10% for each respective dataset.

Table V. Comparing the Naive Method to Models—Average Relative Error in Percentage (0% Means Same Computed Position)

Dataset	BSL	KDE	OPT
London	29%	28%	27%
New York	26%	24%	22%
Paris	19%	18%	16%
Beijing	12%	11%	11%

Table VI. Prediction Accuracy in Terms of Estimated Distance from the Unknown POI Location—BSL Versus OPT Models

	Dataset							
	London		New York		Paris		Beijing	
Distance	*BSL*	*OPT*	*BSL*	*OPT*	*BSL*	*OPT*	*BSL*	*OPT*
0–2km	8%	**11%**	14%	**24%**	10%	**12%**	21%	**29%**
2–4km	36%	33%	43%	33%	30%	**33%**	22%	**35%**
4–6km	30%	**31%**	21%	**30%**	37%	35%	39%	21%
6–8km	17%	16%	15%	9%	16%	15%	13%	11%
>8km	9%	9%	7%	4%	7%	5%	5%	4%

Table VII. Prediction Accuracy in Terms of Estimated Distance from the Unknown POI Location—KDE Versus OPT Models

	Dataset							
	London		New York		Paris		Beijing	
Distance	*KDE*	*OPT*	*KDE*	*OPT*	*KDE*	*OPT*	*KDE*	*OPT*
0–2km	9%	**11%**	16%	**24%**	10%	**12%**	23%	**29%**
2–4km	31%	**33%**	41%	33%	31%	**33%**	27%	**35%**
4–6km	33%	31%	24%	**30%**	36%	35%	33%	21%
6–8km	18%	16%	13%	9%	17%	15%	11%	11%
>8km	9%	9%	6%	4%	6%	5%	6%	4%

There are cases in which the mean location is closer to the unknown POI location than the predicted location of the BSL, KDE, and OPT models. Table V shows the actual average distance gap in a percentage for each dataset. The results show that although not outperforming the mean location in a small number of cases, the predicted location is still close to the computed mean location. The results show that for scenarios with good data coverage, such as Beijing (many spatial relationships extracted from texts), the spatial probabilities almost always (greater than 90%) outperform a naive method, and even if they do not, they produce similar results.

5.2.2. Location Estimates. Having established the validity of our approach, we want to measure the distance of the estimated to the actual POI location. Tables VI and VII illustrate the percentage of location estimation scenarios belonging to the predefined distance buckets of 0 to 2km, 2 to 4km, 4 to 6km, 6 to 8km, and greater than 8km. The tables show the location estimate errors for the model pairs BSL and OPT, and KDE and OPT for all four city datasets. To interpret the results, we can observe that the more results fall into the shorter distance buckets (smaller error), the better are the estimates for the specific model case. Again, an improvement of the result quality can be observed when contrasting the BSL and OPT models. The latter accumulates more estimates in the shorter distance buckets. A similar observation can be made when contrasting the KDE and OPT models, with the KDE model performing again slightly better than the BSL model in most cases. Specifically, for the case of London and New York, there is an increase (indicated in bold) in the first (0 to 2km) and third (4 to

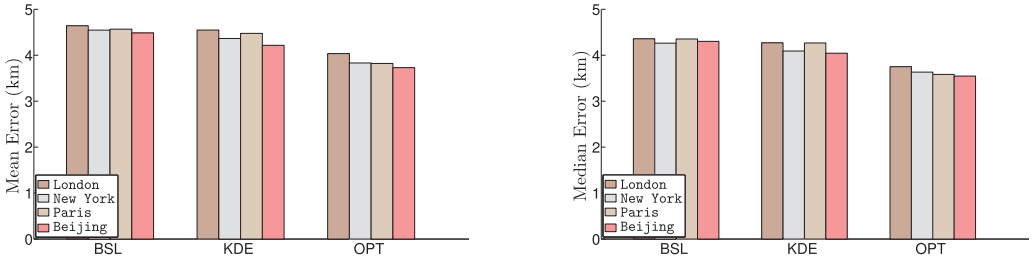


Fig. 12. Estimation error in kilometers: mean (a) and median (b) error.

Table VIII. Distance Between the Center of the Spatial Probability Distribution and the Unknown POI

	Percentage of Relations Considered		
Dataset	10%	50%	100%
London	15.3km	7.9km	7.7km
New York	16.2km	11.9km	11.1km
Beijing	14.4km	8.6km	**1.2**km
Paris	8.7km	**1.6**km	**0.8**km

6km) bucket percentages in Table VI, and in the first (0 to 2km) and second (2 to 4km) bucket percentages in Table VII, whereas for the Paris and Beijing datasets, there is an increase in the first (0 to 2km) and second (2 to 4km) distance buckets in both tables. This means that by using the OPT method, we obtain more precise location estimates than by using BSL or KDE models.

Assuming a perfect method, all results would be in the first bucket. The case of Beijing, with many available spatial relations, comes close to this ambition, as 29% and 35% (64% total) of the estimates are within 2km and 4km of the actual location, respectively. Moreover, Figure 12(a) shows the mean and Figure 12(b) the median error expressed in kilometers for BSL, KDE, and OPT models for all four datasets, respectively. In most cases, the mean and median error is close to 4km (4.1km $<$ *MeanError* $<$ 4.6km, 4km $<$ *MedianError* $<$ 4.3km) for BSL and KDE models with small variations. The OPT model outperforms, again, both BSL and KDE models by achieving mean and median errors that are smaller than 4km (3.9km $<$ *MeanError* $<$ 4.1km, 3.6km $<$ *MedianError* $<$ 3.9km) in almost all cases.

5.2.3. Case Studies. To illustrate the impact of the number of observations on the result quality, we visualize four concrete location estimation scenarios (one for each dataset) by progressively increasing the number of observations (spatial relationships) in each case. Figure 13 illustrates the aforementioned scenarios. Figure 13(a) through (c) illustrate an unknown POI (red star) in the greater London area whose position is described in relation to known POIs (black stars) using a total number of 15 spatial relations. Figure 13(a) shows the contours of the spatial probability distribution when we use 50% (randomly selected) of the observations, whereas Figure 13(b) shows the final distribution considering all spatial relations. Finally, Figure 13(c) is a close-up of Figure 13(b) with a Google Maps basemap overlay. In a similar fashion, Figure 13 shows the results for New York, Beijing, and Paris, with a total number of 20, 70, and 200 spatial relations being used in each case, respectively. These results demonstrate the considerable prediction accuracy. Especially the estimates for Beijing (see Figure 13(g) through (i) and Paris (see Figure 13(j) through (l)) clearly pinpoint the

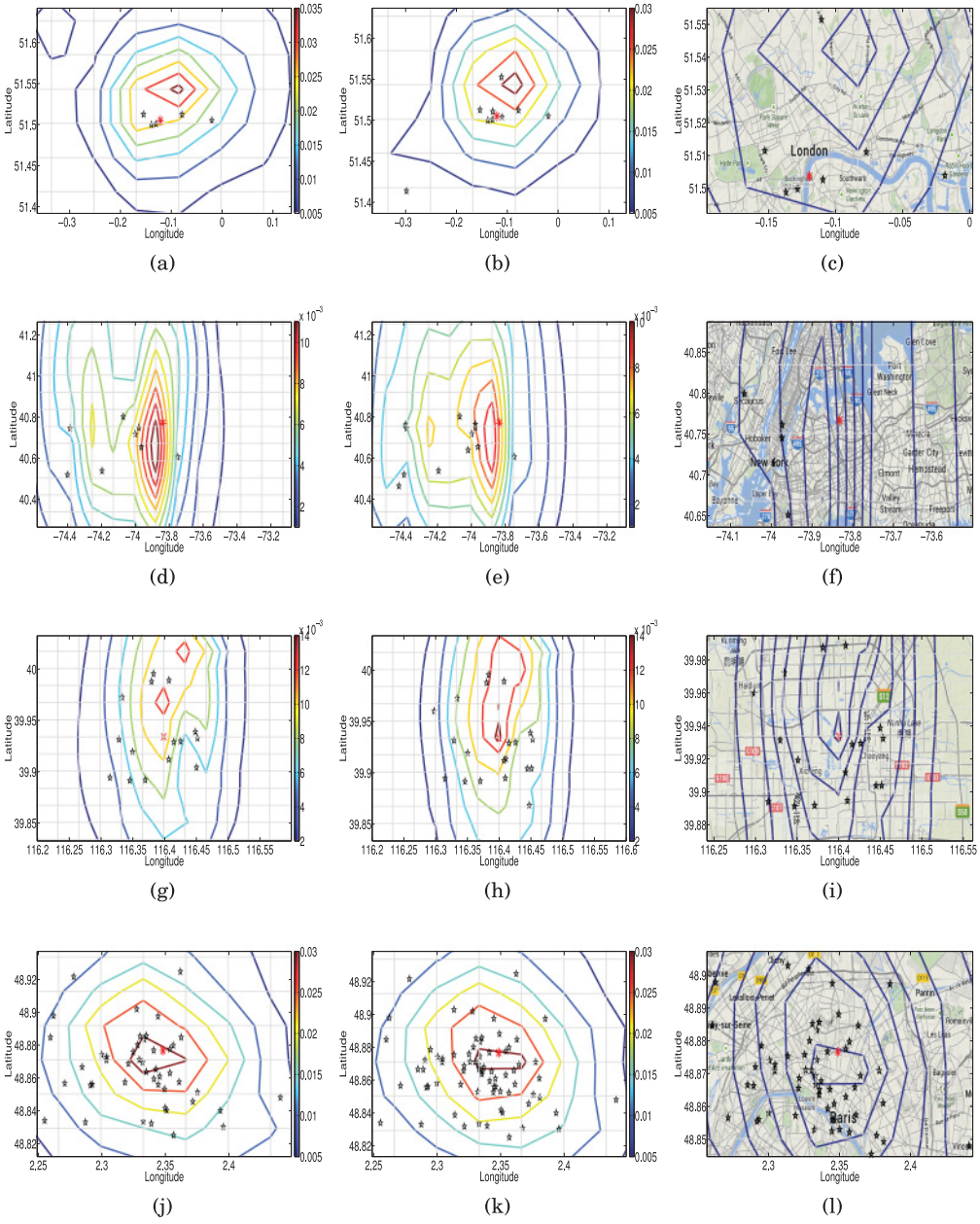


Fig. 13. Real-world location estimation scenarios. Rows 1 through 4 are scenarios for London, New York, Beijing, and Paris. Columns 1 through 3 shows results for 50%, 100%, and 100% (close-up on Google Maps) of the observations (discovered relations) considered in the estimation.

unknown POI location. What is further encouraging is that even for the cases of London (see Figure 13(a) through 9c)) and New York (see Figure 13(d) through (f)), for which the number of relations is considerably smaller, the proposed approach works reasonably well.

As expected, the prediction accuracy increases with the number of observations considered. This is confirmed by the mass of the probability moving closer to the unknown POI location when increasing the number of observations from a randomly selected 50% (first column in Figure 13) to 100% (second column in Figure 13). This effect is observed for all four cases. Table VIII shows the distances between the centers of the spatial probability distributions (OPT model result) and the actual POI locations as we increase the percentage of spatial relations considered in our estimate.

The results show that as we increase the number of relations, we achieve more accurate estimates. The improvement is considerable for all cases, with Beijing and Paris benefitting most and achieving errors of less than 2km (indicated in bold in Table VIII). Moreover, although the estimation quality (accuracy as well as precision) increases with the number of observations, even in the case of a small number of available observations we can rely on the crowd as a data source for location estimation.

Overall, we can conclude that the proposed OPT modeling using GMMs optimized by the greedy EM algorithm presented in Section 3.2 can efficiently handle the uncertainty introduced by user-contributed qualitative geospatial data. In combination with information extraction techniques and with our location estimation algorithm presented in Section 4.2, it provides us with the nontrivial means of textual narrative-based location estimation.

6. RELATED WORK

Work relevant to this article includes (i) location estimation of multimedia data and social network (mostly Twitter) users and (ii) location estimation of unknown points of interest based on spatial relations from textual data.

6.1. Location Estimation of Multimedia Data and Social Network Users

The work of Friedland et al. [2010] is one of the first attempts for multimodal location estimation on videos where visual, acoustic, and textual information is combined to declare where a video was recorded. Furthermore, Choi et al. [2013] extend this work, where the authors study human performance as a baseline for location estimation for three different combinations of modalities (audio only, audio + video, audio + video + textual metadata) and compare it to the automatic method's performance in Friedland et al. [2010]; the study demonstrates cases when humans could effectively identify an audio cue for estimating the video's location when the automatic method failed. Kelm et al. [2013] combine the data from the visual and textual modalities with external geographical knowledge bases by building a hierarchical model that combines data-driven and semantic methods to group visual and textual features together within geographical regions. As a result, the proposed method successfully located 40% of the videos in the MediaEval 2010 Placing Task test set within a radius of 100m.

From a different perspective, Cheng et al. [2010] consider geolocation prediction from Twitter data. In particular, the authors propose a probability framework to estimate the city-level location of a Twitter user based on tweet content. According to their results, about half of the Twitter users can be placed within 100 miles of their true locations. Following this line of work, Chang et al. [2012] propose to model the spatial usage of a word as a GMM, which is an approach that is also followed in our work. Content-based machine learning techniques for Twitter user localization are presented in Jaiswal et al. [2013], whereas Han et al. [2012, 2014] study the location estimation

problem that is based on the automatic identification of location indicative words—that is, words that implicitly or explicitly encode an association with a particular location.

Backstrom et al. [2010] have similarly found spatial scaling among online interactions: they show how this association appears so strong and important that it can be safely exploited to infer where Facebook users are only from the location of their friends.

McGee et al. [2013] extend the method [Backstrom et al. 2010], citing the difficulties of adapting from Facebook to Twitter: (i) user-provided data is significantly less precise in Twitter, (ii) the geographic scale of the study moves from only within the United States to a global scope, and (iii) social relationships in Twitter serve multiple roles, beyond signifying friendships. As such, McGee et al. [2013] seek to classify a user's Twitter relationships according to the probability that they serve as strong predictors of that user's location. For ground truth, users with at least three GPS posts are selected, and the median latitude and longitude values are selected as their location. The naive method discussed in our work is similar to this approach. In Section 5, we showed that the proposed modeling and grid-based location estimation algorithm outperforms mean location solutions.

Kong et al. [2014] propose several extensions to the Backstrom et al. [2010] model based on strategies for weighting which of a user's friends are likely to be most predictive of their location. Given a user, their friends are weighted according to a social tightness coefficient, which is computed as the cosine similarity of the two users' friends.

Finally, some topic models that take into account geographical lexical variation have been proposed in Eisenstein et al. [2010] and Hong et al. [2012]. These are very interesting approaches that connect words within high-level topics with specific geographic regions.

We highlight that although the preceding approaches study a similar problem to ours, they do not handle scenarios where the observations contain POIs with positions not stored in a geographical database (they assume that all extracted POIs are known). In stark contrast, our approach can further improve upon the works presented earlier by providing probabilistic location estimates for POIs observed for the first time.

6.2. Location Estimation of Unknown POIs

After the present article was written, we became aware of independent recent works on location estimation of POIs that are not stored in any geographical database. Specifically, the Moncla et al. [2014a, 2014b] propose an unsupervised geocoding algorithm that employs clustering techniques to estimate a spatial footprint of toponyms not found in gazetteers. The authors evaluate their approach with a corpus of real hiking descriptions in three different languages. However, there is an important difference with our setting: the authors assume that there is no uncertainty included in human descriptions, which is a rather strong assumption for real applications. In particular, they consider the hiking descriptions as a priori 100% correct, and they provide a heuristic solution based on predefined patterns and categories of spatial relationships. We believe that our approach—that is, probabilistic modeling of spatial relationships combined with our grid-based location estimation algorithm—is an important complementary feature in such scenarios that takes the uncertainty contained in user-generated texts into account and further improves other proposed approaches for location prediction of unknown POIs.

7. CONCLUSIONS

The increase in available user-generated data provides us with a unique opportunity to generate rich geospatial datasets. With textual narrative being the most popular form of human expression on the Internet, this work provides a method that effectively translates text into geospatial datasets. Our specific contribution is detecting spatial

relationships in textual narratives and using them to estimate the position of unknown object locations. This is also a first step toward solving the geocoding problem for "colloquial" locations generated by user-generated content. We introduce specific preprocessing techniques for extracting spatial relations from textual narratives and use a novel quantitative approach based on training probabilistic models for the representation of spatial relations. Combining these models and interpreting them as observations allows us to reason about unknown object locations. The proposed approach provides an optimized spatial relation modeling technique combined with a grid-based location estimation algorithm that achieves high-quality location estimation results as evidenced by a range of real-world datasets. Here, our probabilistic approach is robust with respect to handling the uncertainties that characterize geospatial observations derived from crowdsourced textual data. The results show that colloquial location estimation facilitated by crowdsourced geospatial narratives is a feasible approach.

Directions for future work include the optimization of the NLP techniques used for the automatic extraction of POIs and spatial relationship information from texts. Furthermore, we will investigate the implementation of global prediction models, which could complement geocoding methods in our increasingly non-Cartesian world. In addition, this will enable us to evaluate additional probabilistic and deterministic modeling techniques and to develop more efficient text-to-map applications.

REFERENCES

Alan Akbik and Jürgen Broß. 2009. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In *Proceedings of the Workshop on Semantic Search (SemSearch'09)*. 6–15.

J. J. Arsanjani, A. Zipf, P. Mooney, and M. Helbich. 2015. *OpenStreetMap in GIScience: Experiences, Research, and Applications*. Springer.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. 61–70.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics Series. Springer.

Razvan Bunescu and Raymond J. Mooney. 2006. Subsequence kernels for relation extraction. In *Proceedings of the 2006 Conference on Advances in Neural Information Processing Systems (NIPS'06)*. 171–178.

H. W. Chang, D. Lee, M. Eltaher, and J. Lee. 2012. @Phillies tweeting from philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'12)*. 111–118.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. 759–768.

Jaeyoung Choi, Howard Lei, Venkatesan Ekambaram, Pascal Kelm, Luke Gottlieb, Thomas Sikora, Kannan Ramchandran, and Gerald Friedland. 2013. Human vs machine: Establishing a human baseline for multimodal location estimation. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13)*. 867–876.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1, 1–38.

Euthymios Drymonas and Dieter Pfoser. 2010. Geospatial route extraction from texts. In *Proceedings of the 1st International Workshop on Data Mining for Geoinformatics (DMGI'10)*. 29–37.

R. Duda, P. Hart, and D. Stork. 2001. *Pattern Classification* (2nd ed.). John Wiley & Sons.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*. 1277–1287.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. 1535–1545.

Gerald Friedland, Oriol Vinyals, and Trevor Darrell. 2010. Multimodal location estimation. In *Proceedings of the 18th ACM International Conference on Multimedia (MM'10)*. 1245–1252.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of 24th International Conference on Computational Linguistics (COLING'12)*. 1045–1062.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49, 1, 451–500.

James Hays and Alexei A. Efros. 2008. im2gps: Estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. 1–8.

Christian Heipke. 2010. Crowdsourcing geospatial data. *Journal of Photogrammetry and Remote Sensing* 65, 6, 550–557.

Daniel S. Hirschberg. 1997. Serial computations of Levenshtein distances. In *Pattern Matching Algorithms*, A. Apostolico and Z. Galil (Eds.). Oxford University Press, 123–141.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. 769–778.

A. Jaiswal, Wei Peng, and Tong Sun. 2013. Predicting time-sensitive user locations from social media. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'13)*. 870–877.

Dmitri V. Kalashnikov, Yiming Ma, Sharad Mehrotra, Ramaswamy Hariharan, and Carter Butts. 2006. Modeling and querying uncertain spatial information for situational awareness applications. In *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems (GIS'06)*. 131–138.

Pascal Kelm, Sebastian Schmiedeke, Jaeyoung Choi, Gerald Friedland, Venkatesan Nallampatti Ekambaram, Kannan Ramchandran, and Thomas Sikora. 2013. A novel fusion method for integrating multiple modalities and knowledge for multimodal location estimation. In *Proceedings of the 2nd ACM International Workshop on Geotagging and Its Applications in Multimedia (GeoMM'13)*. 7–12.

Longbo Kong, Zhi Liu, and Yan Huang. 2014. SPOT: Locating social media users based on social network context. *Proceedings of the VLDB Endowment* 7, 13, 1681–1684.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing* 8, 3, 4:1–4:36.

Jukka M. Krisp. 2013. *Progress in Location-Based Services*. Springer.

Jonathan Q. Li and Andrew R. Barron. 1999. Mixture density estimation. In *Advances in Neural Information Processing Systems 12*, S. A. Solia, T. K. Leen, and K.-R. Muller (Eds.). Morgan Kaufmann, San Mateo, CA, 279–285.

Corrado Loglisci, Dino Ienco, Mathieu Roche, Maguelonne Teisseire, and Donato Malerba. 2012. An unsupervised framework for topological relations extraction from geographic documents. In *Database and Expert Systems Applications*. Lecture Notes in Computer Science, Vol. 7447. Springer, 48–55.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics—Volume 1 (ETMTNLP'02)*. 63–70.

Y. Ma, D. V. Kalashnikov, and S. Mehrotra. 2009. Towards managing uncertain spatial information for situational awareness applications. *IEEE Transactions on Knowledge and Data Engineering* 20, 10, 1408–1423.

Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM'13)*. 459–468.

Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. 2013. Effectiveness and efficiency of open relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*. 447–457.

Ludovic Moncla, Mauro Gaio, and Sébastien Mustière. 2014a. Automatic itinerary reconstruction from texts. In *Geographic Information Science*. Lecture Notes in Computer Science, Vol. 8728. Springer, 253–267.

Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014b. Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking description corpus. In *Proceedings of the 22nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL'14)*. 253–267.

Dimitris Papadias and Timos Sellis. 1994. Qualitative representation of spatial knowledge in two-dimensional space. *VLDB Journal* 3, 4, 479–516.

Dieter Pfoser. 2011. On user-generated geocontent. In *Proceedings of the 12th Symposium on Spatial and Temporal Databases (SSTD'11)*. 458–461.

Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Flavio Fluart, Wajdi Zaghouani, Anna Widiger, Ann Charlotte Forslund, and Clive Best. 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. 53–58.

Jochen Schiller and Agnès Voisard. 2004. *Location Based Services*. Morgan Kaufmann.

Georgios Skoumas, Dieter Pfoser, and Anastasios Kyrillidis. 2013. On quantifying qualitative geospatial data: A probabilistic approach. In *Proceedings of the 2nd ACM International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD'13)*. 71–78.

Randall Smith, Matthew Self, and Peter Cheeseman. 1990. Estimating uncertain spatial relationships in robotics. In *Autonomous Robot Vehicles*. Springer, 167–193.

Daniel Sui, Sarah Elwood, and Michael Goodchild. 2012. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer.

J. J. Verbeek, N. Vlassis, and B. Kröse. 2003. Efficient greedy learning of Gaussian mixture models. *Neural Computation* 15, 469–485.

Jan Oliver Wallgrün, Alexander Klippel, and Timothy Baldwin. 2014. Building a corpus of spatial relational expressions extracted from Web documents. In *Proceedings of the 8th Workshop on Geographic Information Retrieval (GIR'14)*. 6:1–6:8.

Yuhang Wang and Fillia Makedon. 2003. R-histogram: Quantitative representation of spatial relations for similarity-based image retrieval. In *Proceedings of the 11th ACM International Conference on Multimedia (MULTIMEDIA'03)*. 323–326.

Yecheng Yuan. 2011. Extracting spatial relations from document for geographic information retrieval. In *Proceedings of 19th International Conference on Geoinformatics*. 1–5.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing—Volume 10 (EMNLP'02)*. 71–78.

Xueying Zhang, Chunju Zhang, Chaoli Du, and Shaonan Zhu. 2011. SVM based extraction of spatial relations in text. In *Proceedings of the IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM'11)*. 529–533.

Received March 2015; revised December 2015; accepted February 2016