Strategies for Geographical Scoping and Improving a Gazetteer

Sanket Kumar Singh
University of Alberta
sanketku@ualberta.ca

Davood Rafiei
University of Alberta
drafiei@ualberta.ca

ABSTRACT

Many applications that use geographical databases (a.k.a. gazetteers) rely on the accuracy of the information in the database. However, poor data quality is an issue when data is integrated from multiple sources with different quality constraints and sometimes with little information about the sources. One major consequence of this is that the geographical scope of a location and/or its position may not be known or may not be accurate.

In this paper, we study the problem of detecting the scope of locations in a geographical database and its applications in identifying inconsistencies and improving the quality of a gazetteer. We develop novel strategies, including probabilistic and geometric approaches, to accurately derive the geographical scope of places based on the spatial hierarchy of a gazetteer as well as other public information (such as area) that may be available. We show how the boundary information derived here can be useful in identifying inconsistencies, enhancing the location hierarchy and improving the applications that rely on gazetteers. Our experimental evaluation on two public-domain gazetteers reveals that the proposed approaches significantly outperform, in terms of the accuracy of the geographical bounding boxes, a baseline that is based on the parent-child relationship of a gazetteer. Among applications, we show that the boundary information derived here can move more than 20% of locations in a public gazetteer to better positions in the hierarchy and that the accuracy of those moves is over 90%.

CCS CONCEPTS

• **Information systems** → *Geographic information systems*; *Probabilistic retrieval models*; Global positioning systems;

KEYWORDS

Geographical scoping, gazetteer improvement, geotagging

ACM Reference Format:
Sanket Kumar Singh and Davood Rafiei. 2018. Strategies for Geographical Scoping and Improving a Gazetteer. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3178876.3186078

1 INTRODUCTION

Gazetteers are extensively used in many different domains and applications because of their wide coverage and detailed information about places. For example, an incoming tweet may have the GPS coordinates of the capturing device but to detect a populated place the tweet is coming from, one may use a gazetteer such as GeoNames [31] to map those coordinates to an actual location entity. There is an increasing number of different devices that record GPS coordinates (e.g. phones, cars, cameras, etc.), and mapping the GPS coordinates to a an administrative location or a populated place can be a useful service, for example, in dispatching services such as ambulance, police, etc. The literature also reports numerous domains where gazetteers are used, including toponym resolution in text [15], geotagging tweets [35], documents [6] and entities [33], etc. To support many of these applications, one needs to both effectively and efficiently join a gazetteer with other geo-coded data. GeoNames reports serving over 150 million web service requests per day (as of October 2017), and many of those services can benefit from more accurate information about places and their boundaries.

However, there are a few challenges that hinder progress in this area: (1) most public gazetteers either do not have bounding boxes for many of their locations (e.g. GeoNames) or their bounding boxes are not accurate (e.g. OSMNames[1], see Section 5.1 for details). In the absence of a bounding box, there is no direct way of checking if an entity falls inside or outside a region boundary[2], and applications have to implement their own ad-hoc solutions; (2) data in a gazetteer is prepared by public and is not necessarily accurate especially for less populated places [1]; (3) there are inconsistencies within gazetteers and in relationship with other sources (see Fig. 1).

Our approach to address those challenges is through maintaining bounding boxes for places. Attaching a bounding box to each place has a number of benefits, including more efficient support for reverse geo-coding queries and better monitoring and enforcement of consistency constraints in the form of relationships between bounding boxes. Since boundaries change due to growths, splits and mergers, maintaining bounding boxes is a continuous process.

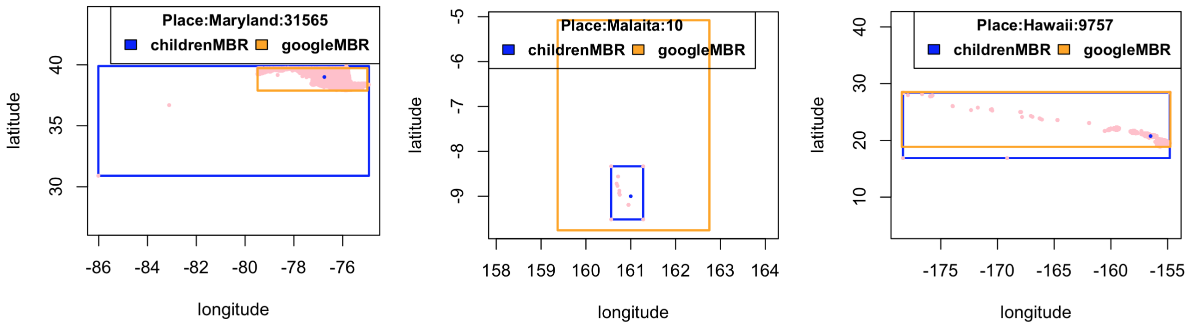
The problem to be studied in this paper is if a bounding box can be accurately constructed for each place based on incomplete and sometimes erroneous information that is available, and if those bounding boxes improve the quality of a gazetteer. We take, as a bounding box, the minimum bounding rectangle (MBR) that satisfy all stated constraints in a gazetteer including the parent-child relationships. Despite their imprecision in some cases, for example, compared to polygons, MBRs provide a simple abstraction that is more efficient for querying [3] and enforcing constraints [22]. Sometimes the stated constraints cannot all be satisfied when creating MBRs. We formalize the search for an MBR as a probabilistic optimization, which tries to find the most likely MBR by dropping the least likely constraints.

Our contributions can be summarized as: (1) We provide a systematic study of the problem of improving and enriching a gazetteer

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186078

[1]http://osmnames.org
[2]Reverse Geocoding API from Google and other search engines convert the coordinates of a point on the map to a human readable location or address, but the details of their proprietary solutions often are not made public.



(a) Presence of outliers

(b) Insufficient coverage

(c) Less accurate center

Figure 1: Examples of inconsistencies between a gazetteer and other sources. (a) MBR of the US state Maryland is overestimated due to outliers, (b) MBR of Malaita, a province in Solomon Islands, is underestimated due to less data coverage, and (c) the given center is far from the center of the MBR. (childrenMBR is the bounding box of the children, as listed in Geonames for the quoted place, and googleMBR is the MBR returned by Google Maps; the child count for each place is indicated, and the center and the children are shown with a blue and pink dots respectively.)

using bounding boxes of places; to the best of our knowledge, this is the first time such a study over millions of places is conducted. (2) We propose strategies for detecting and resolving inconsistencies in a gazetteer. (3) We evaluate our strategies and report their accuracy in detecting the boundaries of places and in improving the places hierarchy. (4) We report on the effectiveness of our bounding boxes in refining the places hierarchy and in augmenting the gazetteer with other data sources including YFCC100M [30].

The rest of this paper is organized as follows: Section 2 reviews the related work, and Section 3 presents our proposed strategies for constructing minimum bounding rectangles of places. Section 4 lists some application areas and settings. Our algorithms are evaluated in Section 5. Section 6 concludes the paper.

2 RELATED WORK

The literature related to our work can be grouped into (1) estimation of the spatial extent of geographic entities, (2) conflict resolution and data cleansing techniques, and (3) automatic gazetteer expansion and enrichment.

Estimation of the spatial extent of geographic entities The geographic boundary of a place can be estimated using geo-tagged entities such as photos [14], videos, and online documents. Chen et al. [8] develop a method to find the spatial extent of places with vague boundaries. They define geographical boundary of a place using the density of Flickr photos mapped to a region. The authors use Kernel Density Estimation to interpolate the boundary for regions where photos are sparse. Parker and Downs [19] cluster points before forming a minimum convex envelope to enclose each cluster. A drawback of both approaches is that if a place is widely spread and have disjoint regions (e.g. places containing islands such as Hawaii), then each region forms its own boundary instead of forming a boundary at a particular level such as country, province or district. For polygonal boundaries of places, Voronoi diagrams are used to approximate the extent of places from their centroids[2]. Also a notion of fuzzy MBR to model the spatial extent of a geographical location is introduced by Somodevilla et al. [28], though no evaluation on the quality or the accuracy is provided.

Conflict resolution and data cleaning Conflict resolution can be viewed as making decisions between different versions of data to determine a golden record. There is a large body of work in this area (e.g. see the survey by Bleiholder and Naumann [4] and the tutorial note by Dong and Naumann [10]). In a more recent work, Prokoshyna et al. [23] combine logical reasoning and a quantitative method to develop a data cleansing approach. The quantitative method in this work involves setting some constraints based on the statistical properties of attribute values and flagging an inconsistency if such constraints are violated. The authors propose a minimal-set repair algorithm to find attribute values that minimize a statistical distortion. Volha et al. [5] resolve conflicts in the context of dbpedia using a fusion function, which is learned from labeled data. The set of fusion functions are manually defined in advance by domain experts. As a generic cleaning technique, outlier detection may be used to filter objects which do not follow the general expected underlying distribution. Two commonly used techniques in spatial domains are Boxplot [13] and Bagplot [25], which are evaluated as part of our proposed heuristic approaches.

Automatic gazetteer expansion and enrichment Automatic gazetteer creation or enrichment involves adding new records or attributes, in the form of a new place or a missing feature of an existing place. Popescu et al. [20] create a gazetteer using diverse information sources and different algorithms for entity extraction, categorization, coordinate discovery and ranking. More recently, Oliveira et al. [18] attempt to enrich the GeoSEn [7] gazetteer using the geographical information gathered through crowd sourcing. The authors augment the spatial hierarchy of the gazetteer by adding places at what appears to be district and street granularity. These works can be seen as orthogonal to ours.

3 GEOGRAPHICAL SCOPING

Given a set of containment relationships and constraints for places in a gazetteer, we want to construct a bounding box for each place such that ideally all stated or known constraints are satisfied. We take as the bounding box of a place, any minimum bounding rectangle (MBR) that is parallel to the latitude and the longitude axes and satisfy the constraints. This does not necessarily give the most accurate bounding box especially if the true bounding box is not convex; however, compared to arbitrary polygons, MBRs are more efficient for checking containment relationships and constraints. A challenge in detecting the spatial extent of a place is that often there is not sufficient information about a place and the relationships can be conflicting or contradictory. For example, the recorded center for Hawaii in GeoNames, as shown in Figure 1(c), is quite far from the center of its MBR. Also, as shown in Figure 1(b), there are no child locations in the northern region of googleMBR and it is difficult to construct this MBR based on parent-child relationships alone.

3.1 Hierarchical Approach

One strategy for building an MBR is to enforce the containment relationships, i.e. each parent MBR must contain the MBRs of its children. Gazetteers are good at describing the containment relationships between places. For example, GeoNames places each location into an administrative level such as country or state and allows queries to retrieve the children within an administrative level. The bounds of the MBR of a place can be calculated in a bottom-up approach by taking the minimum of south-west coordinates and the maximum of north-east coordinates of all children of the place in the spatial hierarchy. A major problem with this strategy is that the number of children can vary greatly for different places. For example, more populated places tend to have more children than rural towns, water bodies, natural regions, etc. Another problem is the skewed distribution of the children and that the children are not always spread over the whole region boundary (as shown in Fig. 1(b) and 1(c)).

3.2 Geometric Approach

A major drawback of the hierarchical approach is that for places that have either very few or no children in the gazetteer, no good geographical extent can be obtained. For example, the area of an MBR for Malaita province, constructed from points in GeoNames (as shown in Fig. 1(b)), is approximately 19 times smaller than that of the bounding box obtained from Google Maps. Our geometric approach aims at addressing this problem.

Given the center point c and the area a of a place, one can construct an infinite number of rectangles, all centered at c with an area a; with no additional information, it is hard to predict which rectangle is more likely. That said, our next statement gives some evidence that maybe a square is a better choice.

Conjecture 3.1. *Let R be the set of all rectangles with a center point c and area a and $r \in R$. Assuming that all rectangles are equally likely, the expected area of overlap between R and r is maximized when r is a square.*

Our geometric approach constructs as the MBR of the place a square centered at C and with an area A. The bounds of the MBR can be obtained as follows: the latitudes of north-east (NE_{lat}) and south-west (SW_{lat}) points are obtained by shifting the latitude of the center in north and south by a factor F given as

$$\mathbf{NE_{lat}} = C_{lat} + F \quad \text{and} \quad \mathbf{SW_{lat}} = C_{lat} - F$$

where $F = (\sqrt{A})/(2 * L)$, L is the distance between two consecutive latitudes ($\approx$ 111 km) and C_{lat} and C_{long} are the latitude and longitude of center C. The longitudes of the endpoints are obtained similarly except that the distance between two consecutive longitudes shrinks as we move toward the poles. Hence we first obtain the distance between two longitudes at a given latitude D_{lat} before shifting the longitude of the center by F_{ne} and F_{sw}, defined as:

$$F_{ne} = (\sqrt{A})/(2 * D_{nelat}) \quad \text{and} \quad F_{sw} = (\sqrt{A})/(2 * D_{swlat})$$

where $D_{nelat} = L * cosine((NE_{lat} * \pi)/180)$ (see [29] for details), and $D_{swlat} = L * cosine((SW_{lat} * \pi)/180)$. Thus

$$\mathbf{NE_{long}} = C_{long} + F_{ne} \quad \text{and} \quad \mathbf{SW_{long}} = C_{long} - F_{sw}.$$

Note that the accuracy of the coordinates of the endpoints depends on the accuracy of L and that of the given center C.

The bounding box estimated using this approach is expected to be accurate when the child locations are distributed uniformly around the given center. However, this approach may not perform well if the child locations include outliers or the given center is away from the mass of child locations.

3.3 Probabilistic Approach

Geographic information in a gazetteer can be both incomplete and inaccurate, and this can lead to inconsistencies. Some of those inconsistencies may be detected using a rule-based method, but the main challenge which still remains is how to handle the uncertainty in spatial data. If each fact or statement in a gazetteer can be assigned a probability that it is true, then detecting a bounding box for a place can be treated as a constraint optimization problem.

3.3.1 **The Model.** Locations in a gazetteer are described by a latitude and a longitude; public gazetteers often do not provide much detail on how the coordinates are obtained and if the given coordinate is actually near the center point of the MBR of the place. To test this, we randomly selected 1000 places each from GeoNames and OSMNames. The coordinates of these places were checked against the center point of the bounding box obtained from a different source (in our case MBR obtained from Google Maps referred to as googleMBR). We found that only 63% of places in GeoNames and about 97% of the places in OSMNames had a coordinate within 10 km of the center obtained from googleMBR.

Modeling the center point: Let d_c denote the distance between a given center c of a place and its true MBR center. If we assume d_c follows the normal distribution with parameters μ and σ, then

$$Pr(d_c|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(d_c-\mu)^2 \big/ 2\sigma^2} \tag{1}$$

where μ and σ are respectively the mean and the standard deviation of d_c. Let P_{center} denote this probability for fixed values of μ and σ. The parameters of the distribution can be easily estimated from the data. In our random sample of 1000 places, μ and σ are 88.894 and 408.760 for GeoNames and 2.057 and 7.09 for OSMNames respectively.

Modeling children: Let q be the probability that an arbitrary location is placed under a correct parent in the gazetteer. The value of q can be estimated by checking for each place in a sample whether its children are assigned a correct parent node. In our sample of 1000 places, the value of q is calculated as 0.968 for GeoNames and 0.882

for OSMNames respectively. One may observe that the probability that an arbitrary location is placed under a correct parent is relatively high; hence based on this empirical result and without much additional knowledge of which places may be correct or incorrect children, an arbitrarily chosen child is more likely to be correct than incorrect hence better be included in the parent MBR. In other words, for a parent place with n children and an MBR that includes i ($i \leq n$) of its children, the probability that a random child location of the parent is enclosed in the MBR can be written as

$$P_{children} = \frac{i}{n}. \tag{2}$$

Putting it together: Assuming independence for the two events in Equations 1 and 2, we can put together the two probabilities into an objective function. Given a place with center c, MBR area A and children set S, we want to find a set $S' \subseteq S$ of children such that

$$\begin{aligned} &\underset{S' \subseteq S}{\text{argmax}} && (P_{children} \cdot P_{center}) \\ &\text{subject to} && area(MBR(S')) \leq A \end{aligned} \tag{3}$$

where area(MBR(S')) refers to the area of MBR formed from places in S'. Everything equal, the model selects an MBR with a center point closest to the given center c. Equation 3 provides a way to model the inclusion and exclusion of locations under an MBR and to estimate the center of the MBR with high certainty.

3.3.2 **Optimization.** Optimizing Eq. 3 can be computationally intensive since the number of possible MBRs is only bounded by the size of the power set of S. The problem may be broken down into two cases: (1) there is an MBR that includes all points and rectangles in S and the area of the MBR is not larger than A, (2) there is an MBR that includes all points and rectangles in S and has an area larger than A. For (1), the MBR that includes all points in S with an area not exceeding A will maximize $P_{children}$. As noted in our sample of 1000 places, there is more uncertainty in finding a correct center than including a correct child, hence one may maximize the term $P_{children}$ before maximizing P_{center}. This means the MBR that includes all points in S can simply be expanded (if needed), moving the center of the MBR to the given center and maximizing the objective function, without violating the area constraint.

Now consider the case where there is no MBR that includes all children with an area less than or equal to A. With the same reasoning, we may optimize $P_{children}$ before plugging in P_{center}.
Naive algorithm: A naive approach to perform this optimization is to enumerate all possible MBRs and select the one that maximizes the objective function in Eq. 3. With n data points, there are n possible choices for each side of an MBR; hence there are $O(n^4)$ MBRs to choose from. Among those MBRs, the algorithm selects the MBR that maximizes the objective function. This is an expensive process for large values of n. Our next algorithm prunes the search space without affecting the correctness of the result.
Improved algorithm: To further prune the search space, one may only consider MBRs where the constraint on the area is not violated. Our improved algorithm first finds an initial solution by dropping extreme points in each direction until the area constraint is met. The algorithm then tries to improve upon the initial solution while making sure the area constraint is satisfied. Let m be the number of points that are dropped to find the initial solution. This sets a limit

Algorithm 1: Find an optimal MBR (as per Equation 3) when child locations are all points

Input: (1) A – area of MBR of the place P,
(2) C – centre of the place P,
(3) S – set of n unique locations $\{p_1 \ldots p_n\}$, under P

bestMBR ⟵ Nil
maxProbability ⟵ 0
Drop furthest points (one at a time) from each side
At each drop, form an MBR for the remaining points. Stop when the area is ≤ A and let MBR at this point be M'
Let m be the number of points strictly outside M'
for *i = 0 to m* **do**
  for *j = 0 to m-i* **do**
    for *k = 0 to m-i-j* **do**
      for *l = 0 to m-i-j-k* **do**
        currentMBR ⟵ the MBR formed after dropping i, j, k, l points from north, east, west and south directions in S
        if *area(currentMBR) > A* **then**
          continue
        x ⟵ (i+j+k+l) `// # of excluded places`
        C' ⟵ centre of currentMBR
        Calculate P_{center} using C and C' in Eq. 1
        $P_{children}$ = (n-x)/n (as in Eq. 2)
        currProb ⟵ $P_{children} \cdot P_{center}$
        if *currProb > maxProbability* **then**
          maxProbability ⟵ currProb
          bestMBR ⟵ currentMBR
return bestMBR

on the number of points in each side that an MBR can pass through. There are $O(m^4)$ such MBRs. The number of wrong entries in a gazetteer is expected to be a small fraction, hence m is expected to be much smaller than n. In our experiments with GeoNames and OSMNames, the maximum respective value of m was 62 and 78 while that for n was 74765 and 708. In terms of the running time, the improved algorithm was faster by up to 6 orders of magnitude for $n < 100$, and the gap was getting bigger for places that had more children to a point where it was not possible to run the naive algorithm on an Intel Core i5 machine running at 2.7GHz with 8GB RAM. One can also do a binary search when selecting the last side of the MBR, reducing the complexity of the naive algorithm to $O(n^3 log(n))$ and that of the improved algorithm to $O(m^3 log(m))$. The full details of the improved algorithm are given in Algorithm 1.

A limitation of our probabilistic optimization model (also referred to as POM in our experiments) is that the optimization does not kick in unless the area of children MBR is greater than or equal to the known area of the place. This is addressed in our next approach.

3.4 Heuristic Approaches

The MBR of children (as discussed in Section 3.1) can vary dramatically in shape and size due to uneven distribution or lack of enough child locations under the parent. To construct an MBR for such places, heuristic approaches may be used. Our first heuristic is based on detecting outlier children.

Outliers removed Gazetteers sometimes have locations that are wrongly placed. Such placements may show as an outlier especially if the wrong child is quite far from other children listed under the same parent. Hence, an outlier detection method may be used to identify and remove such places before constructing an MBR. Two approaches that are used in geographical contexts are Boxplot [32] and Bagplot [25]. Since boxplot is a univariate method, it can be applied across the latitude and longitude dimensions independently. A point may be deemed an outlier if it is classified as an outlier in any one of the two dimensions. Bagplot is a bivariate extension of boxplot, which generates a convex hull with 50% of the points (called a 'bag') and an outer loop (known as 'fence'), which can vary in size depending on the number of points one wants to include. We expand the outer loop till the area of the MBR formed by the enclosed points is closest or equal to the given area of the place. All points outside the outer loop are excluded from the MBR.

Hybrid MBR An MBR of a place may be obtained using its center point and area; an MBR of a place may also be obtained based on the children listed. If we treat each MBR as a random variable which is 1 for points inside the MBR and 0 for points that fall outside, the region where the two MBRs overlap is where both random variables are taking the value of 1. The rectangle marked by the intersection of the two MBRs is expected to give a more reliable description of the boundary. However, the region of overlap can be much smaller than the actual MBR. We next discuss how this intersection region can be expanded such that its area matches the given area.

(1) Hybrid MBR with uniform enlargement (H-enlarge) Let l and w denote the length and the width of an MBR. One way to enlarge the MBR is to enlarge both l and w by a constant s. Given an area a, we want $(l + s)(w + s)$ to be close to a. In other words, the value of s can be obtained by solving the following quadratic equation:

$$s^2 + (l + w)s + (lw - a) = 0 \quad \text{where } s > 0.$$

The coordinates of the expanded MBR are obtained by shifting the latitude and longitude of north-east and south-west corners by $s/2$ degree in each direction (as in Section 3.2).

(2) Hybrid MBR with scaling (H-scale) The sides of the intersection region can be scaled by a factor s such that the area of the expanded MBR becomes a. Hence, the value of s can be obtained as

$$s = \sqrt{(a/lw)}.$$

It should be noted that outliers may be removed before performing any of the above expansions. Also, when there is no intersection between the two MBRs, the expansion may be applied to the MBR that is expected to be more accurate (our experiments use children MBR in those cases).

As an example, Figure 2 depicts the MBRs for Budapest, Hungary, obtained using different methods discussed in this section.

4 IMPROVING A GAZETTEER

Maintaining the spatial footprints of places in a gazetteer can both improve the quality of the database and offer benefits to other applications that use it. We study three such areas of improvements (see Sections 4.1, 4.2 and 5.2 for our experimental evaluation).

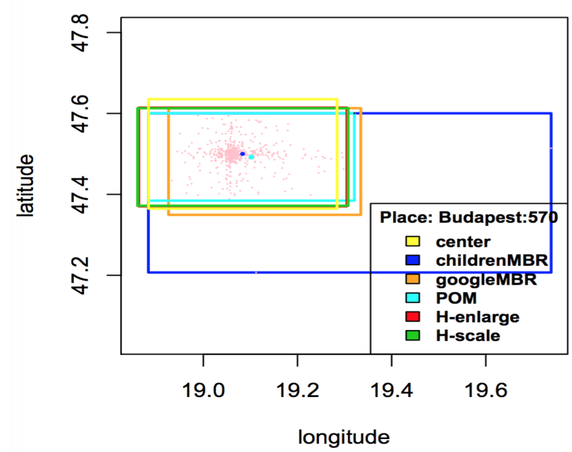


Figure 2: The MBRs for Budapest, Hungary, obtained using different methods. (The blue and cyan dots represent the given center and the center obtained from the POM approach respectively. The MBR for H-enlarge overlaps with that of H-scale, hence not visible, and the yellow MBR obtained from the geometric approach is expected to be square but it doesn't look like one because of the way endpoint coordinates are calculated on the spherical surface of earth and the varying distance between longitudes, as discussed in Section 3.2.)

4.1 Gazetteer Refinement

The hierarchy of a gazetteer may be refined based on the spatial footprints of places to improve its overall accuracy. Here are two such refinements that we have experimented with.

Change of parent The MBR of each parent node in a gazetteer is expected to contain the MBR of its children. A child node c listed under a parent node p may be considered for a possible parent change if this containment relationship is violated. In such cases, there can be multiple locations that can contain the MBR of c and they may or may not be true parent places. One approach to reduce the likelihood of selecting a wrong parent is through setting some constraints. For example, one may change the parent from p to p' if p' is the only place that can contain c.

Restructuring children Sometimes locations are not placed at the right level or granularity; ideally we want to place each child at the lowest level of the hierarchy. A restructuring can check for each place c in level l if its MBR is fully contained in the MBR of another place p at the same level. If one such relationship holds and p is the only node that has this relationship with c, then it is likely that c is a child of p and better be placed under p. This process may continue until no more move is possible.

4.2 Gazetteer Enrichment via Geotagging

A gazetteer may be enriched by including or integrating information about geographic locations, such as tags, tweets, news, photos, etc. Since many resources on the Web are not geotagged, a related question is if using the geographical extent of places can improve geotagging; we study this in the context of geotagging photos and videos from Flickr.

One approach for geotagging is to divide the earth surface into a grid of equal-sized cells [16, 26, 27] and to predict the most probable cell for a given photo or video. A problem with fixed-sized cells is that the spatial scope of a place can be distributed over several cells (when the cells are too small) or multiple geographical boundaries can be collapsed into one cell (when the cells are too large), affecting the accuracy. An alternative is to use MBRs to better maintain the locality relationships and place boundaries A geotagging using MBRs can be carried out in two steps as follows:

(1) MBR prediction Based on the hypothesis that different users inside the boundaries of a place may use similar tags to describe the place, one may first predict an MBR for a photo/video using the textual annotations attached. The relevance of a tag t to an MBR can be expressed in terms of the probability that a user inside the MBR m_j uses t to tag his/her photos, i.e.

$$p(t_i|M_j) = \frac{\text{\# of users who use tag } t_i \text{ in MBR } m_j}{\text{\# of users in mbr } m_j}$$

where M_j is the model of MBR m_j. To avoid zeroing the score in case a tag is not seen in m_j, a smoothing function may be used. Using Jelinek-Mercer smoothing [34], we have

$$p(t_i|m_j) = \alpha p(t_i|M_j) + (1-\alpha)p(t_i|M_{mbrs})$$

where α is the smoothing factor with a value in the range (0, 1) and $p(t_i|M_{mbrs})$ is the model for all MBRs, defined as

$$p(t_i|M_{mbrs}) = \frac{\text{\# of users who use tag } t_i \text{ over all MBRs}}{\text{\# of users over all MBRs}}.$$

In our experiment, we set the value of α at 0.8. In general, α can be set using a validation set, trying different values of the smoothing factor and selecting a value that gives the best MBR prediction over the validation set. Assuming independence between the tags inside an MBR, the relevance score of a test instance T with tags $t_1, \ldots, t_n$ is given as

$$p(T|m_j) = \prod_{i=1}^{n} p(t_i|m_j).$$

In our experiments, we use the log of $p(T|m_j)$ as our scoring function for numerical stability. One may note that the relevance score can be biased toward user-specific tags, which often do not carry any location information (e.g. person name). To avoid such ambiguity, we remove all tags which are used just by a single user. Furthermore, to allow a locality of the tags, the same user in different cells or MBRs is considered as a new user. Finally, the MBR with the maximum score is selected. In case no MBR is found (which will happen if none of the tags of the test instance is seen during training), one may predict the MBR which has the maximum number of photos or videos assigned.

(2) Coordinate estimation Given the MBR of a photo or a video, the actual coordinate within the MBR can be predicted based on the same technique that finds the coordinates within a grid cell [27].

4.3 Enforcing topological constraints

The accuracy and correctness of a gazetteer may be warranted through topological constraints, which can harness the relationships between MBRs such as containment, disjointness, overlap, etc. Such constraints may be classified into *hard constraints* (those which cannot be violated) and *soft constraints* (those which can be violated but the violations are rare). An example of the former, expressed in terms of MBRs, is "if the MBR of place A does not contain the MBR of place B, then location A is not part-of or doesn't contain location B."

Level	# of places	μ	σ	Threshold	
				$\alpha = 0.05$	$\alpha = 0.01$
ADM1	2000	0.2756	0.1451	0.514	0.613
ADM2	2000	0.1819	0.1655	0.454	0.566
ADM3	2000	0.1634	0.1462	0.403	0.503

Table 1: Expected area of overlap between MBRs at each level (ADM1 = province, ADM2 = district or large city, ADM3 = locality or small town) and the respective thresholds at each level of significance.

Soft constraints may be enforced, based on parameters such as the expected area of overlap between MBRs, whcih may be estimated from data. For example, Table 1 shows a few statistics for the normalized overlap area of MBRs at different administrative levels, each based on a random sample of 2000 places from GeoNames. While μ and σ are the mean and the standard deviation of the normalized overlap areas of places, the column threshold gives an upper bound on the mean normalized overlap area of a place, calculated using one-tailed critical values of z-score (Z = 1.645 and 2.326 corresponding to $\alpha = 0.05$ and 0.01 respectively) at different levels of significance and setting the sample size 'n' as one. Examples of soft constraints based on Table 1 are: (SC1) "for any location at ADM2 level, its average normalized overlap area with all other places cannot be much below or higher than the mean normalized overlap area (0.1819)," and (SC2) "if an update at ADM3 level makes the mean normalized overlap area of a place greater than 0.403, then the null hypothesis that the update follows the data distribution can be rejected at $\alpha = 0.05$."

5 EXPERIMENTAL EVALUATION

We evaluate the proposed approaches in terms of both the accuracy and the effectiveness of the bounding boxes that are constructed.

5.1 Accuracy of an MBR

The accuracy of a bounding box of a place may be measured against published data from authoritative sources such as government agencies and international organization bodies. One such official dataset that we are aware of, and is used in our experiments, is the US Census dataset[3] which provides cartographic boundary files for places in the US. An MBR of a place from the boundary data can be obtained by finding the maximum and the minimum of all the given coordinates. We are not aware of similar comprehensive list of boundary regions for places outside the US, but there are sources that provide data on a best-effort basis. As one such source, we use Google Reverse Geocoding API[4] to fetch the 'true' bounding box for a location. Google Maps has been used in similar context in the literature [12, 17]. In our experiments, we evaluate the accuracy of an MBR for places in the US using both the US Census data and Google Maps as baselines; for each place outside the US, the MBR obtained from Google Maps is used as a baseline.

[3]https://www.census.gov/geo/maps-data/data/kml/kml_state.html

[4]https://developers.google.com/maps/documentation/geocoding/intro

Dataset and Preprocessing Our evaluation is conducted using the two large public gazetteers, namely GeoNames and OSMNames. The features used in our evaluation are (1) the latitude and longitude of each place, which we assume as the center of the MBR of the place, and (2) the parent-child relationship expressed in the spatial hierarchy. After removing all locations with non-unique geo-coordinates, we put together the following datasets. From GeoNames, we extracted (1) the set of 50 USA states (Geo-50), (2) 540 random world locations each with at least one child (Geo-540), and (3) 140 random world places each with at least one child and with the area of its children MBR larger than the given area of the place (Geo-140). From OSMNames, we similarly extracted (1) 1500 random locations each with at least one child (OSM-1500) and (2) 160 random locations (referred to as OSM-160) with the same constraints as those for Geo-140. As the MBR area of a place, we used the area of the MBR returned by Google and from the US Census dataset. In our experiments, the radius of the earth is taken as 6371 km and the distance between consecutive longitudes at equator is 111.0 km.
Evaluation Measure The accuracy of a bounding box is measured in terms of its coverage of the actual area of the place and the fraction of child places it encloses. For a predicted MBR P and true MBR T, the following measures of accuracy are used:

- *Area Overlap Accuracy (AOA)* - The ratio of the area of the region covered by the intersection of P and T to the area of the region covered by the union of P and T. As an evaluation metric, AOA is used in other domains (e.g. image segmentation [11]).
- *False Negative for area overlap* (FN_{area}) - The ratio of the area of T not covered by P and the area of T.
- *False Positive for area overlap* (FP_{area}) - The ratio of the area of P not part of T and the area of P.
- *Point Overlap Accuracy (POA)* - The ratio of the number of points in the intersection region of P and T over the number of points in the union region.
- *False Negative for point overlap* (FN_{point}) - The ratio of the number of points in T which are not covered by P over the number of points covered by T.
- *False Positive for point overlap* (FP_{point}) - The ratio of the number of points in P which are not covered by T over the number of points in P.

Methods	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Children	68.90	29.94	**1.26**	**97.94**	2.05	**2.0**
Center	63.49	25.91	19.68	87.65	**2.02**	12.32
Children_woOutlier	**92.44**	**3.24**	4.61	97.33	2.03	2.63
H-enlarge	74.55	17.10	13.15	91.97	2.03	7.99
H-scale	75.78	16.26	13.83	92.65	2.03	7.31

Table 2: Evaluation result for Geo-50 dataset using US Census dataset as baseline in (%)

Overall accuracy: The results on Geo-50 (see Tables 2 and 3) show that children MBR without outliers (children_woOutlier) outperforms all other strategies and across both the baselines, namely Google Maps and US Census data. This is mainly due to a large number of child locations per US state (i.e. 40,210) in Geo-50. This also shows that gazetteers have a good coverage of developed places.

Methods	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Children	68.03	29.06	**3.25**	97.95	2.04	**2.0**
Center	64.58	23.88	20.83	77.64	**2.02**	12.33
Children_woOutlier	**90.80**	**2.49**	7.05	**97.34**	**2.02**	2.63
H-enlarge	76.24	14.32	13.85	91.97	**2.02**	8.00
H-scale	76.00	15.17	14.67	92.65	**2.02**	7.32

Table 3: Evaluation result for Geo-50 dataset using Google Maps as baseline in (%)

As shown in Table 4, with a large false positive rate for Geo-140 and OSM-160, when the area of children MBR is larger than expected, our probabilistic approach (POM) and hybrid approaches (H-enlarge and H-scale) perform very well (between 7% to 33% improvement over Children MBR), accurately excluding outliers and shrinking the area close to a given area. Our hybrid approaches also do better when the places do not contain many spatial points or there is too much uncertainty in data, as shown for Geo-540 and OSM-1500 where hybrid approaches perform the best. This is because our hybrid approaches construct their MBRs based on the overlap between center MBR and children MBR, which is expected to have less uncertainty.
Varying the MBR area and the number of child locations: To better understand how each method performs under different settings and if one method is better under a more specific setting, we varied the MBR area and the child count and studied the performance of different methods across different testsets. The success rate of each method is defined as the ratio of the number of instances for which the method gives the best accuracy over total number of instances in a given area range or child count range. The success rate of each method on different testsets is shown in Fig. 4 (a) and (b) for different ranges of MBR areas and in Fig. 4 (c) and (d) for different ranges of child counts.

It can be observed from Fig. 4 (a) that the probabilistic approach (POM) performs best for places with area greater than 7383.75 km^2, which includes mostly large cities, districts, provinces and small countries. These places also corresponds to the range of child count greater than 553 in which our probabilistic approach performs best, as shown in Fig. 4 (c). On the other hand, for majority of places with small area or less child count (see Fig. 4 (a), (b), (d)), conflation techniques such as center MBR and hybrid MBR seem to work well.
Enlargement vs Scaling: In this experiment, we study the effect of applying uniform scaling and uniform enlargement operations on the accuracy of hybridMBR. Figure 3 shows the number of instances in which a particular operation outperforms the other operation at different administrative levels on Geo-540. One can observe that the performance of H-enlarge is comparable to that of H-scale for places at higher administrative levels (e.g. province) while H-enlarge performs better for places at lower levels in the spatial hierarchy. This is because H-scale operation extends the intersection region without changing the shape of the intersection region greatly. In case of uniform enlargement, the intersection

Methods	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}	AOA	FP_{area}	FN_{area}	POA	FP_{point}	FN_{point}
Dataset	**Geo-540**						**OSM-1500**					
Children	34.20	6.27	61.68	94.49	5.50	**1.85**	26.27	10.85	66.04	90.20	9.79	**2.6**
Center	65.39	22.71	22.23	91.73	**4.29**	7.49	63.25	24.71	24.43	91.87	**4.56**	6.76
Children_woOutlier	34.80	**5.44**	61.72	**94.52**	5.45	1.86	25.97	**10.83**	66.37	90.20	9.78	2.61
H-enlarge	**71.86**	17.95	**17.85**	94.28	4.75	4.28	**64.29**	23.79	**23.63**	94.15	4.90	4.08
H-scale	63.74	24.37	25.04	94.49	4.73	4.06	59.43	28.16	28.02	**94.32**	4.91	3.87
Dataset	**Geo-140**						**OSM-160**					
Children	44.43	53.66	**7.02**	91.32	8.67	**0.71**	36.33	59.86	**18.85**	60.99	39.0	**3.75**
Center	63.78	24.09	23.66	84.87	6.79	12.94	53.39	**32.60**	32.34	76.47	**15.81**	14.99
POM	**78.16**	9.42	17.29	**92.89**	6.79	2.12	43.98	39.57	46.20	64.20	34.15	19.52
Children_woOutlier	67.12	**9.30**	27.95	87.10	**6.44**	7.56	39.76	43.79	30.82	64.51	33.96	5.43
H-enlarge	72.09	17.98	17.87	89.37	6.82	7.89	**55.33**	31.03	30.98	**77.41**	17.18	12.18
H-scale	72.16	17.96	17.85	89.98	6.85	7.23	55.04	31.41	31.36	77.40	17.27	11.99

Table 4: Evaluation result for different testsets in (%)

region is equally incremented in both horizontal and vertical directions, and this results in an MBR shape that is close to a square. This seems to work better for places at a coarser granularity since it covers the majority of area around the intersection region.

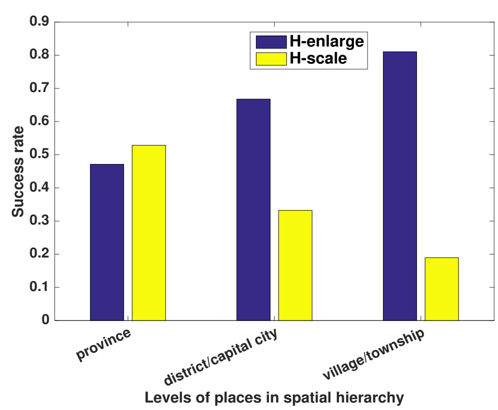


Figure 3: Success rate of hybrid strategies varying the level of places.

Comparison with MBRs from OSMNames: One benefit of using OSMNames for our evaluation is that it allow us to compare the bounding boxes of places in OSMNames with our MBRs, in reference to MBRs from a third party (i.e. google MBR in our case). The area overlap accuracy for places in OSM-160 using their bounding boxes from OSMNames gives an accuracy of 31.48% whereas our geometric approach achieves 55.33% on the same testset; that means our methods can construct better MBRs than those in OSMNames. On OSM-1500, the area overlap accuracy for the bounding boxes in OSMNames is 66.53%, compared to 64.29% accuracy of our hybridMBR. This shows that MBRs obtained from our methods are pretty close to those in OSMNames.

5.2 Effectiveness of an MBR

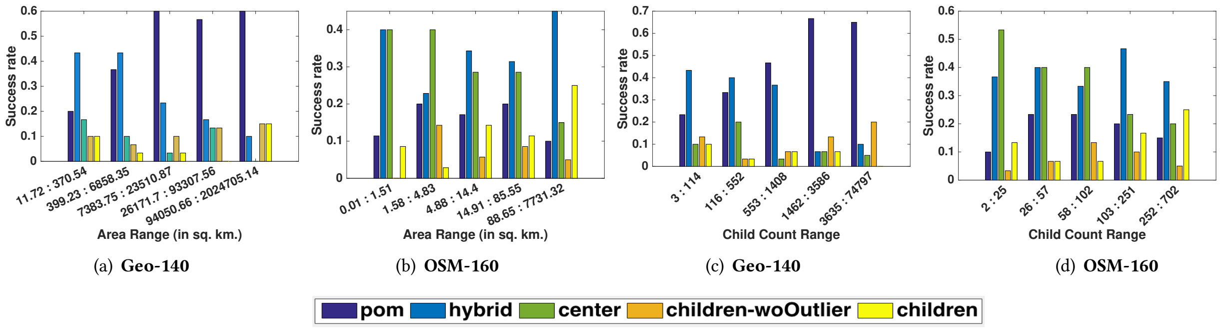
Effectiveness of an MBR can be measured in terms of its usability in some of the applications discussed in Section 4.

Gazetteer refinement The dataset used was GeoNames with its spatial hierarchy constructed using feature class, feature code and administrative level codes. To generate an MBR for each place, we extracted the area of places from a public domain site (in our case Wikipedia Infobox) and used it as input to the strategies discussed in Section 3. Since the area information was available for a limited set of places, and not many places did have children, an MBR was constructed using one of our methods in the given order: (1) the probabilistic model (POM) was used when the area information was available and the expected area of the MBR was less than the children MBR; (2) the geometric model (Center MBR) was used when the place had no children; (3) the hierarchical model (Children MBR) was used otherwise. We ended up generating MBRs for 93,274 locations which are also available online[5].

To evaluate the refined hierarchy, we obtained two samples of 100 places randomly selected from the list of places processed under *change of parent* and *restructuring children* (as discussed in Section 4.1). For former, we verified whether a place which was identified as inconsistent was actually inconsistent and for the latter, we verified whether a location was part of another location. This was done manually by looking into Wikipedia text and/or Google Maps.

The total number of places identified as wrongly placed under *change of parent* was 67,820 while the total number of places moved deeper in the hierarchy was 2,081,709 (roughly 20% of places in GeoNames). This shows that in the absence of a geographic scope, there are many places which are kept directly below the root level. Furthermore, our evaluation result shows that 91% of places in our sample (91/100) are moved correctly down the hierarchy. This provides a strong evidence in support of an accurate restructuring. For *change of parent*, the fraction of places which were actually inconsistent was 3/100; this empirical result shows that the MBRs are robust enough to support the movement of places deeper in the hierarchy (i.e. vertical movements) but inconsistent for moving the nodes across the hierarchy (i.e. horizontal movements). The children places identified as wrongly placed were often streams,

[5]https://goo.gl/WxcbMG



(a) **Geo-140** (b) **OSM-160** (c) **Geo-140** (d) **OSM-160**

Figure 4: Success rate of the methods varying the area of MBR of the places in (a) and (b) and varying child count in (c) and (d).

forests or places which lied near the geographical boundary of a parent. This is mainly due to the vague geographic scope of natural landscapes and mis-alignments of child MBRs, which does not allow an MBR to fall completely under a parent node.

Gazetteer enrichment via geotagging The dataset used for this experiment consisted of 4,631,717 photos/videos for training and 500,000 photos/videos for testing. This dataset which was provided in 'Placing task'[6] was extracted by organizers from the YFCC100M corpus (see [9] for details). Our preprocessing of this data included removing stopwords (as detected by Weka [7]) as well as tags that contained special characters and doing a stemming using Snowball [21] stemmer. As a baseline for comparison with our MBR-based approach, we also divided the surface of earth into cells of size 0.3 degree latitude and longitude. The training instances were then mapped to the MBRs (generated in Section 5.2 under gazetteer enrichment) and cells independently which gave 52,294 MBRs and 44,926 cells with at least one photo or video. The reason for generating cells of size 0.3 degree was to keep the number of cells relatively close to the number of MBRs; otherwise a method with fewer cells or MBRs is expected to perform better in MBR prediction with less chance of making an error. The ground truth (i.e. geo-coordinates for test photos and videos) was provided in the dataset.

The prediction accuracy is measured in terms of the Average Distance Error (ADE), calculated as the mean Haversine distance [24] between predicted and true coordinates over all photos and videos. We also measured the prediction accuracy for MBRs (and similarly for cells) defined as the ratio of the number of instances for which a predicted MBR (cell) is the same as the true MBR (cell) to the total number of instances in the testset.

Our experiment gives an ADE of 2561.114 km for MBRs compared to 3039.674 km for cells, with a prediction accuracy of 41.23% for MBRs and 32.37% for cells. This clearly shows that geo-tagging using MBRs is more accurate than a grid-based approach. The analysis of the results show two major reasons for a wrong prediction of MBR or cell, which leads to a large distance error; (1) There are several instances of photos or videos in the testset, which mention some general terms only (e.g. 'affect', 'ipad') and they do not carry any location-specific information. These tags can occur anywhere on the world map and it is difficult to predict a location for them. (2) There are cells and MBRs which are sparsely populated, i.e. they have very few users assigned (usually 1 or 2). As a result, even though there are several dense MBRs or cells (with the number of users of order 10^5) containing multiple tags of a test instance, a sparsely populated cell or MBR is predicted as best cell or MBR for the test instance. Also, such cases are seen more often for cells which are created randomly without any knowledge of geographical scope of the locations.

[6]http://www.multimediaeval.org/mediaeval2016/placing

[7]http://weka.sourceforge.net/doc.dev/weka/core/Stopwords.html

6 CONCLUSIONS AND FUTURE WORK

We studied different strategies for building the bounding boxes of places using the spatial hierarchy of a gazetteer and information such as the area of places, which are available in public domains. Our extensive evaluation on various datasets and settings show that an accurate bounding box can be constructed and that these bounding boxes can improve other applications that use a gazetteer. Also, while our POM-based approach works best for places at district, provinces or higher level, our geometric and heuristic approaches can be employed for places without enough coverage in a gazetteer.

Our work can be extended in a few directions, and we are exploring some of those directions. First, one can integrate the proposed strategies into a single model that is invariant to parameters such as the number of child locations, the distribution of points, etc. Second, the local features of a region such as landscape of places in a close proximity, population density, etc. can be used to improve the accuracy of boundary boxes. Third, a limitation of our probabilistic model is that it is not applicable when the area of a children MBR is smaller than a given area. Also our optimization in Section 3.3.2 makes use of the uncertainty of events that are observed on our gazetteers; relaxing those assumptions can make the optimization more challenging. Fourth, our heuristic approach, which removes the outliers, does not work well when the center of a place is away from its child locations; this is another area for further research. Finally, exploring other applications of a gazetteer enriched with MBRs is also a possible future direction.

ACKNOWLEDGMENTS

This research is supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Dirk Ahlers. 2013. Assessment of the accuracy of GeoNames gazetteer data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*. ACM, 74–81.

[2] Harith Alani, Christopher B Jones, and Douglas Tudhope. 2001. Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science* 15, 4 (2001), 287–306.

[3] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. 1990. The R*-tree: an efficient and robust access method for points and rectangles. In *ACM Sigmod Record*, Vol. 19. Acm, 322–331.

[4] Jens Bleiholder and Felix Naumann. 2009. Data fusion. *ACM Computing Surveys (CSUR)* 41, 1 (2009), 1.

[5] Volha Bryl and Christian Bizer. 2014. Learning conflict resolution strategies for cross-language wikipedia data fusion. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 1129–1134.

[6] Cláudio Elízio Calazans Campelo and Cláudio de Souza Baptista. 2008. Geographic scope modeling for web documents. In *Proceedings of the 2nd international workshop on Geographic information retrieval*. ACM, 11–18.

[7] Cláudio Campelo and Cláudio de Souza Baptista. 2009. A model for geographic knowledge extraction on web documents. *Advances in Conceptual Modeling-Challenging Perspectives* (2009), 317–326.

[8] Jiaoli Chen and Shih-Lung Shaw. 2016. Representing the Spatial Extent of Places Based on Flickr Photos with a Representativeness-Weighted Kernel Density Estimation. In *International Conference on Geographic Information Science*. Springer, 130–144.

[9] Jaeyoung Choi, Claudia Hauff, Olivier Van Laere, and Bart Thomee. 2016. The Placing Task at MediaEval 2016. *MediaEval 2016 Workshop* (Oct. 20-21 2016).

[10] Xin Luna Dong and Felix Naumann. 2009. Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment* 2, 2 (2009), 1654–1655.

[11] Jan Funke, Fred A Hamprecht, and Chong Zhang. 2015. Learning to segment: training hierarchical segmentation under a topological loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 268–275.

[12] Maurizio Gibin, Alex Singleton, Richard Milton, Pablo Mateos, and Paul Longley. 2008. An exploratory cartographic visualisation of London through the Google Maps API. *Applied Spatial Analysis and Policy* 1, 2 (2008), 85–97.

[13] Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 2 (2004), 85–126.

[14] Livia Hollenstein and Ross Purves. 2012. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science* 2010, 1 (2012), 21–48.

[15] Ehsan Kamalloo and Davood Rafiei. 2018. A coherent unsupervised model for toponym resolution. In *Proceedings of the Web (former WWW) Conference*. ACM. https://doi.org/10.1145/3178876.3186027

[16] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2015. Geotagging social media content with a refined language modelling approach. In *Pacific-Asia Workshop on Intelligence and Security Informatics*. Springer, 21–40.

[17] Weimo Liu, Md Farhadur Rahman, Saravanan Thirumuruganathan, Nan Zhang, and Gautam Das. 2015. Aggregate estimations over location based services. *Proceedings of the VLDB Endowment* 8, 12 (2015), 1334–1345.

[18] Maxwell Guimarães de Oliveira, Cláudio EC Campelo, Cláudio de Souza Baptista, and Michela Bertolotto. 2016. Gazetteer enrichment for addressing urban areas: a case study. *Journal of Location Based Services* 10, 2 (2016), 142–159.

[19] Jonathon K Parker and Joni A Downs. 2013. Footprint generation using fuzzy-neighborhood clustering. *Geoinformatica* 17, 2 (2013), 285–299.

[20] Adrian Popescu, Gregory Grefenstette, and Pierre Alain Moëllic. 2008. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 85–93.

[21] Martin F Porter. 2001. Snowball: A language for stemming algorithms. (2001).

[22] Rosanne Price, Nectaria Tryfona, and Christian S Jensen. 2001. Modeling topological constraints in spatial part-whole relationships. In *International Conference on Conceptual Modeling*. Springer, 27–40.

[23] Nataliya Prokoshyna, Jaroslaw Szlichta, Fei Chiang, Renée J Miller, and Divesh Srivastava. 2015. Combining quantitative and logical data cleaning. *Proceedings of the VLDB Endowment* 9, 4 (2015), 300–311.

[24] C Carl Robusto. 1957. The cosine-haversine formula. *The American Mathematical Monthly* 64, 1 (1957), 38–40.

[25] Peter J Rousseeuw, Ida Ruts, and John W Tukey. 1999. The bagplot: a bivariate boxplot. *The American Statistician* 53, 4 (1999), 382–387.

[26] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. 2009. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 484–491.

[27] Sanket Singh and Davood Rafiei. 2016. Geotagging Flickr Photos And Videos Using Language Models, In MediaEval 2016 Working Notes Proceedings. *Available from World Wide Web: http://slim-sig.irisa.fr/me16proc/*.

[28] María J Somodevilla and Fred E Petry. 2004. Fuzzy minimum bounding rectangles. In *Spatio-Temporal Databases*. Springer, 237–263.

[29] Kurt Stüwe. 2007. *Geodynamics of the lithosphere: An introduction*. Springer Science & Business Media.

[30] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[31] Mark Wick and Bernard Vatant. 2012. The geonames geographical database. *Available from World Wide Web: http://geonames. org* (2012).

[32] David F Williamson, Robert A Parker, and Juliette S Kendrick. 1989. The box plot: a simple visual method to interpret data. *Annals of internal medicine* 110, 11 (1989), 916–921.

[33] Jiangwei Yu and Davood Rafiei. 2016. Geotagging Named Entities in News and Online Documents. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 1321–1330.

[34] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* 22, 2 (2004), 179–214.

[35] Wei Zhang and Judith Gelernter. 2014. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science* 2014, 9 (2014), 37–70.