GeoHighlight: An Interactive Point-Recommendation Approach for spatial Data

Behrooz Omidvar-Tehrani[†], Plácido A. Souza Neto[‡]
[†]The Ohio State University, USA, [‡]Federal Institute of Rio Grande do Norte - IFRN, Brazil
[†]omidvar-tehrani.1@osu.edu, [‡]placido.neto@ifrn.edu.br

ABSTRACT

spatial data is becoming increasingly available in various domains such as transportation and social science. Discovering patterns and trends in this data provides improved insights for planning and decision making for smart city management, disaster management and other applications. However, exploratory analysis of such data is a challenge due to its huge size and diversity of spatial data. It is often unclear for the analyst *what to see next* during an analysis process, i.e., lack of guidance. To tackle this challenge, we formulate guidance as an optimization problem and develop GeoHighlight, an efficient interactive guidance approach for spatial data. At each step of an interactive process, k-most interesting geographical points become highlighted to guide the analyst through further steps. We illustrate the efficiency and usability of our framework in an extensive set of experiments.

1. INTRODUCTION

Nowadays, there exists huge amounts of spatial data in various fields of science, such as agriculture, transportation and social science. Analysis of such data is interesting as it is grounded on reality: each record represents a specific geographical location. Moreover, understanding patterns and trends provides analysis insights leading to improved user planning and decision making. Some instance applications of spatial data are smart city management, disaster management and autonomous transport.

Spatial data analysis is often performed in *exploratory context*: the analyst does not have a precise query in mind and she explores data in iterative steps in order to find potentially interesting results. Traditionally, an exploratory analysis scenario on spatial data is described as follows: the analyst visualizes a subset of data using a query in an off-the-shelf product (e.g., Tableau[1], Spotfire[2]). The result will be illustrated on a geographical map. Then she investigates on different parts of the visualization by zooming in/out and panning the map in order to discover patterns and trends of interest. The analyst may iterate on this process several times by issuing different queries and focusing on different aspects of data.

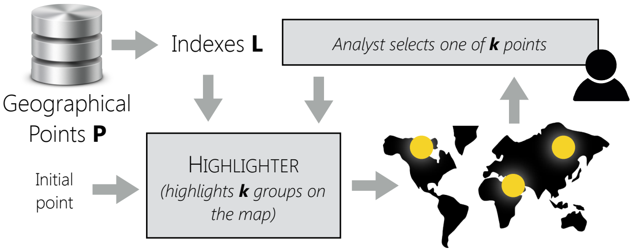
Geographical
Points **P**
Indexes **L**
Analyst selects one of ***k*** points
Initial
point
HIGHLIGHTER
(highlights ***k*** groups on the map)

Figure 1: GeoHighlight Framework

The focus of the literature in spatial data analysis is on *efficiency* of exploratory iterations: *"how can analysts navigate in spatial data fluidly?"* The common approach is to design pre-computed indexes which enable efficient retrieval of spatial data (e.g., [?]). However, there has been fewer attention to the *value* of spatial data. Despite the huge progress on efficiency front, an analyst may easily get lost in the plethora of geographical points as she doesn't know what to investigate next in an exploratory context. In other words, although iteration transitions can be performed efficiently, but the decision to form a transition remains unclear. The following example illustrates the challenge in practice.

Example 1. *Liam is planning a dinner date in New York. He is open to any restaurant and he wants to explore different options (i.e., exploratory analysis). He uses Yelp website*[3] *and asks for all restaurants in New York with a fair price. His query results in 49,000 restaurants. As he has no other preferences, an exhaustive investigation needs scanning each restaurant independently which is nearly infeasible. On the other hand, while he is looking at primary restaurants in the list, he shows interest in Indian restaurant menus. An ideal system can capture this feedback in order to short-list a small subset of restaurant that the analyst should consider as high priority.*

To overcome the challenge of value in exploratory analysis, visualization environments offer a complete tool-set to manipulate data (filter, aggregate, etc.). In practice, this

[1] *http://www.tableau.com*
[2] *http://spotfire.tibco.com*
[3] *http://www.yelp.com*

Proceedings of the VLDB Endowment, Vol. 10, No. 6
Copyright 2017 VLDB Endowment 2150-8097/17/02.

duplicates the problem: the analyst is left alone in a huge space of data and tools. The principled challenge for the analyst is *"what to see next"* in the exploratory context. A *guidance* mechanism is then necessary to point out potential future directions of analysis.

In this demo paper, we introduce GeoHighlight framework to highlight a subset of geographical points based on analyst feedback and facilitate the decision-making process by providing guidance on what the analyst is potentially interested in. The highlighted set should have high quality. Quality is formulated as optimization of two dimensions: *relevance* and *diversity*. First, a highlighted point should be relevant to historical choices of the analyst. Second, highlights should be diverse to unveil different aspects of analysis.

Typically in the literature, we exploit user explicit feedback for recommendation. For instance, we use preferences of the user on movies to predict what kind of movies she/he is interested to watch in the future and make a recommendation. However, we can also exploit implicit user feedback for guidance. Guidance differs from recommendation. We don't prune data in guidance. We only say based on what you have implicitly payed attention on, here is few highlights that you may need to look at. Implicit feedback can be captured in different ways like spare time in a dedicated page or gaze.

Example 2. *Tiffany is a data scientist and is tasked to design a* chain marketing *strategy for a Peking Duck product whose headquarters is in New York. She already knows that the product has success in the local area. So she analyzes Yelp data*[4] *(i.e., restaurant check-ins) to find out what other locations exhibit similar eating profiles as New York. She asks for k geographical points which have relevant eating profile to New York and are the most diverse. Given $k = 3$, Tiffany receives points from San Fransisco, Washington DC and Marlton, NJ. She selects Marlton due to its proximity to reduce transportation costs. Then she asks for other 3 best points for Marlton. She can then make the city-to-city chain marketing strategy.*

In this paper, we address the problem of guidance. Despite the great progress on spatial data analysis in recent years, we point out following challenges for guidance: *i. Genericness.* Considering the heterogeneous nature of spatial datasets, it is challenging to come up with a generic guidance approach which is independent of data type and distribution. *ii. Size.* The gigantic size of spatial datasets hinders its effective discovery. *iii. Efficiency.* Guidance should be done efficiently in consecutive steps, so that the train of the analyst's thoughts won't break. Despite progress in efficient spatial processing [14], sub-second interactivity is still missing.

There exist few instances of information-highlighting methods [9, 11, 13, 12]. However all these methods are *objective* and do not apply to the context of spatial guidance where user feedback is involved. In terms of recommendation, few approaches focus on spatial dimension [6, 8] while the context and result diversification are missing.

In this paper, we propose a generic interactive analysis approach for guiding analysts towards potential interesting points. The analyst considers the guidance and picks a direction for the next analysis iteration.

[4] *https://www.yelp.com/*

2. PROBLEM STATEMENT

Data Model. We consider a spatiotemporal database $\mathcal{D}$ consisting $\langle \mathcal{P}, \mathcal{A} \rangle$ where $\mathcal{P}$ is the set of geographical points and $\mathcal{A}$ is the set of point attributes. For each $p \in \mathcal{P}$, we consider a tuple $\langle lat, lon, alt, t \rangle$ where lat, lon and alt denote p's geographical coordinates (latitude, longitude and altitude respectively), and t is the timestamp. The set $\mathcal{A}_p$ contains attribute-values for p over the schema of $\mathcal{A}$. For instance, on a bike-sharing dataset, $\mathcal{A}_p = \langle$`female`, `young`, `hybrid-bike`$\rangle$ on the schema $\mathcal{A} = \langle$`gender`, `age`, `type`$\rangle$ denotes that p is associated to a young female cyclist who rides a hybrid bike. The set $\mathcal{A}$ is domain-dependent and defines the semantics of a spatiotemporal dataset.

In this paper, we address the problem of *generic guidance* in spatiotemporal data: "what is the process of guiding analysts in iterative analysis steps on any spatiotemporal dataset?" In other words, we are interested in an approach which highlights a set of k points that the analyst should consider in the next analysis iteration. This should not be a heuristic-based data-dependent highlighting, but a generic approach which is applied on any spatiotemporal dataset. We describe the desiderata of generic guidance approach as follows.

D1. Genericness. The guidance component should be agnostic (making no assumption) about the dataset type, attributes and distribution. In other words, the guidance approach should not be a function of any property of data.

D2. Limited Options. The set of k highlighted points should not be very large because too many options distract the analyst.

D3. Relevance. The fundamental difference between highlighting and k-NN spatial queries [5] is that, in the former, the focus is on k points which have similar characteristics to p, hence relevant. For instance, consider a taxi ride in New York for a young male customer for an itinerary of 10 kilometers and \$3 tip. In contrary to thousands of kilometers of geographical distance, the ride is very similar to another one in San Fransisco for a middle-age male customer for an itinerary of 8 kilometers and \$2.5 tip. Given two points p and p', we define *relevance* as follows.

$$relevance(p, p') = average_{a \in \mathcal{A}_p \cup \mathcal{A}_{p'}}(sim(p, p', a)) \quad (1)$$

The similarity function $sim()$ can be any function such as Jaccard and Cosine. Each attribute can have its own similarity function (as string and integer attributes are compared differently.) Then $sim()$ works as an overriding-function which provides encapsulated similarity computations for any type of attribute.

D4. Diversity. A guidance approach should also consider coverage of all points: k highlighted points should represent distinct regions so that the analyst can observe different aspects of data and decide for the next analysis iteration. Hence, k points should be diverse. Given a set of points $s = \{p_1, p_2 \dots\}$, we define *diversity* as follows.

$$diversity(s) = average_{\{p,p'\} \subseteq s | p \neq p'} distance(p, p') \quad (2)$$

The function $distance(p, p')$ operates on geographical coordinates of p and p' and can be considered as any distance function of Chebyshev distance family such as Eucledian. However, as distance computations are done in *spherical* *space* using latitude, longitude and altitude, it is au-naturel to employ Harvestine distance shown in Equation 3.

$$distance(p,p') = [acos(cos(p_{lat}).cos(p'_{lat}).cos(p_{lon}).cos(p'_{lon}) + cos(p_{lat}).sin(p'_{lat}).cos(p_{lon}).sin(p'_{lon}) + sin(p_{lat}).sin(p'_{lat}))] \times earth\_radius \quad (3)$$

D5. Interactivity. The exploratory nature of the analysis requires the guidance component to be involved in an interactive process. Hence the analyst can investigate and refine different aspects of spatiotemporal data in iterative steps. For being interactive, the guidance component should be efficient so that the train of thought of analyst would not be broken during the analysis process.

Following aforementioned desiderata, we formulate highlighting as an optimization-based problem on relevance and diversity dimensions.

PROBLEM 1 (GEOGUIDE). *Given an input point p and a threshold σ, the problem is to return top-k points denoted S_p where $|S_p| = k$ and $\forall p' \in S_p, relevance(p,p') \geq \sigma$ and $diversity(S_p)$ is maximized.*

Problem 1 is hard due to the huge space of spatiotemporal data: for any given point p, an exhaustive search over all other points is necessary to find k points with maximal relevance. Moreover, the problem investigates in two dimensions at the same time (relevance and diversity) which makes it more challenging.

3. ALGORITHM

We propose a solution for GEOGUIDE by inspiring from both recommendation [10] and visual highlighting [9, 11] methodologies. GEOGUIDE requires an efficient algorithm for dynamically analyzing and comparing geographical points. We propose GEOHIGHLIGHT as a solution for the generic guidance problem in spatiotemporal data (Figure 1). Although GEOHIGHLIGHT operates on points, its functionality can be easily extended to regions using point-clustering methods such as k-means.

Intuitively, GEOHIGHLIGHT is layer on top of raw visualization which highlights k representatives at each analysis iteration. The representatives reflect implicit feedback of the analyst in previous steps. We define a feedback vector which contains all point and region attributes. Example of point attributes are price range and cuisine (in Yelp dataset) and example of region attributes are average altitude and speed in a region (in flight data). Initially, zero is assigned to all attributes. Whenever the system receives an implicit feedback on any of those attributes, it will augment it.

At each iteration, the feedback vector is compared with all points to find the ones which are more similar to analyst feedback. To speed up comparison, we exploit boolean representations. We convert both the feedback and point identity to booleans and make boolean comparisons which are indeed faster.

We dont want to be slow in price of values and insights. So we make our algorithm best effort and bound it to a time limit. In time limit, it starts from most similar options only to previous options and they also become diverse.

GEOHIGHLIGHT operates in two steps: PREPARATION and HIGHLIGHTER. In order to speed up computing relevance in online execution, we pre-compute an inverted index for each single geographical point in $\mathcal{P}$ in the offline PREPARATION step (as is commonly done in Web search). Each index $\mathcal{L}_p$ for the point p stores all other points in $\mathcal{P}$ in decreasing order of their relevance with p. Thanks to the parameter σ, we only partially materialize the indexes.

Algorithm 1: HIGHLIGHTER Algorithm

```
   Input: $p \in \mathcal{P}$, $\sigma$, $k$, $tlimit$
1  $\mathcal{S}_p \leftarrow get\_top\_k(\mathcal{L}^p)$
2  $p_{next} \leftarrow get\_next(\mathcal{L}^p)$
3  while ($tlimit\ not\ exceeded \land relevance(p, p_{next}) \geq \sigma$) do
4      for $p_{current} \in \mathcal{S}_p$ do
5          if $diversity\_improved(\mathcal{S}_p, p_{next}, p_{current})$ then
6              $\mathcal{S}_p \leftarrow replace(\mathcal{S}_p, p_{next}, p_{current})$
7              $break$
8          end
9      end
10     $p_{next} \leftarrow get\_next(\mathcal{L}^p)$
11 end
12 return $\mathcal{S}_p$
```

Algorithm 1 illustrates the online execution step of GEOHIGHLIGHT so called HIGHLIGHTER. The algorithm is a single greedy procedure that solves the GEOGUIDE problem. HIGHLIGHTER is called at each interactive step of GEOHIGHLIGHT (as in Figure 1). The algorithm admits as input a point $p \in \mathcal{P}$ and returns the best k points denoted $\mathcal{S}_p$.

To comply with the desiderata **D5**, we consider a time limit parameter *tlimit* in Algorithm 1. In a *best-effort* strategy, the algorithm bounds user waiting time by *tlimit* to return the best possible results by then.

HIGHLIGHTER begins by retrieving the most relevant points to p by simply retrieving the k highest ranking points in $\mathcal{L}_p$ (line 1). Function $get\_next(\mathcal{L}_p)$ (Line 2) returns the next point p_{next} in $\mathcal{L}_p$ in sequential order. Lines 3 to 11 iterate over the inverted indexes to determine if other points should be considered to increase diversity while staying within the time limit and not violating the relevance threshold with the selected point. Since points in $\mathcal{L}_g$ are sorted on decreasing relevance with p, the algorithm can safely stop as soon as the relevance condition is violated (or if the time limit is exceeded).

The algorithm then looks for a candidate point $p_{current} \in \mathcal{S}_p$ to replace in order to increase diversity. The boolean function $diversity\_improved()$ (line 5) checks if by replacing $p_{current}$ by p_{next} in $\mathcal{S}_p$, the overall diversity of the new $\mathcal{S}_p$ increases.

4. ILLUSTRATIVE SCENARIO

We illustrate an application of GEOHIGHLIGHT in a realistic scenario for New York taxi dataset[5]. This dataset has been frequently exploited for urban analysis (e.g. in [7]). The dataset contains 173,179,759 records of taxi trips and 18 attributes such as pickup and dropoff date/time, passenger count and trip distance. The scenario illustrates how an analyst can achieve an exploratory analysis goal. We preprocessed the original dataset and considered a subset of 20K unique points for the sake of clarity of results. We

[5] *https://data.cityofnewyork.us/view/gn7m-em8n*

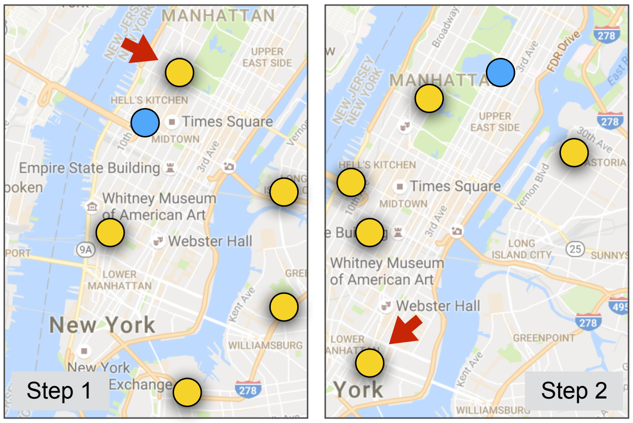

Figure 2: Application of GeoHighlight

employ Highlighter (Algorithm 1) with following parameters: $\sigma = 0.2$, $k = 5$ and $tlimit = 200ms$.

Consider Lucas, a data scientist whose task is to optimize New York taxi trips. Focusing on cab-idle locations, he wants to discover which neighborhoods work the best for which drivers to increase the overall availability. Also, he wants to discover how drivers should choose their next cab-idle station to be more available. Lucas employs GeoHighlight and follows a case-by-case inspection as his analysis methodology by analyzing and learning from historical data.

He begins the analysis by selecting a point from the most crowded region in New York, i.e., Times Square (Figure 2 left). The point depicts a drop-off at *"3 Times Square, New York, NY"* on January 9, 2014 around 9PM (the blue point in the figure). Highlighter then provides 5 relevant points to the selected point (yellow points in the figure). Among 5 highlighted points, Lucas selects a pick-up at *"West 53rd Street"* near Central Park occurred approximately at the same time of the first selection (the point marked with an arrow in the figure). This pick-up has a potential to enchain with the first choice (i.e., a drop-off) to engage the driver in a larger distance.

In the next step, Highlighter shows 5 other points relevant to the new selection (Figure 2 right). Lucas looks for a good drop-off point which is in a neighborhood of the previous selection as the cab-idle station. Lucas selects a highlight in downtown as others are around the train station which have often less taxi requests at nights. This selection contributes to the heavy cab request in Manhattan island at that time of the day. Note that in both steps, the k results are a compromise between relevance and diversity.

5. EXPERIMENTS

We evaluate the efficiency and usefulness of GeoHighlight in an extensive set of experiments. We consider two types of experiments: first, a performance study measures the influence of relevance and size constraint thresholds on execution time. Second, we measure the usefulness of our framework in a user study.

Experiment Settings. Unless otherwise stated, we use the same settings discussed in Section 4. All experiments are implemented in Python (functionality) and JavaScript D3 (visualization) on a 2.8GHz Intel Core i5 machine with an 16GB main memory, running OS X 10.9.2.

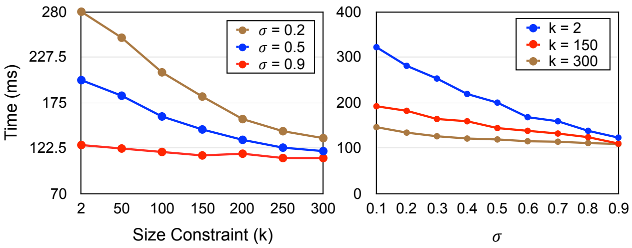

Figure 3: Performance Evaluation

Performance Study. GeoHighlight is designed for exploratory context where interactivity is a need. The "best-effort" greedy approach of Highlighter (Algorithm 1) guarantees to return the best possible results within a time limit. We consider a large time limit ($tlimit = 2s$) in order to evaluate the effect of relevance and size constraint thresholds on execution time.

Figure 3 left illustrates the effect of size constraint by varying k from 2 to 300. In general, larger values of σ provides more freedom for the algorithm hence more time-consuming. An interesting observation is that increasing k leads decreasing execution time. This is because in larger sets, there exist fewer opportunities for increasing diversity, hence Highlighter terminates early.

Figure 3 right confirms that lower values of σ decreases execution time as they provide more flexibility for diversity improvement. For larger values of k (i.e., when $k > 100$), the influence of σ on execution time becomes insignificant.

User Study. The principled question that we ask ourselves is whether GeoHighlight is useful for analysts in practice. To answer this question, we designed a user study with 24 participants (students in Computer Science). Half of the participants know the New York region well (experts) and the other half have a limited knowledge (novice). In our user study, we define a task for each participant and ask him/her to fulfill the task using both GeoHighlight and Tableau (as the most advanced spatiotemporal visualization tool). Then we measure the cardinality of steps to reach the goal.

We define two tasks, *T1: finding a point in a requested location*, and *T2: finding a point with a requested profile.* As an example for *T1*, we ask participants to find points in the Central Park area. An example of *T2* is to find a drop-off point with $2 tip whose trip distance is 3 kilometers. Participants may begin their navigation from three different starting points: *I1: close to the goal*, *I2: far from the goal*, and *I3: random.* We evaluate the effect of expertise, goal and starting point on the analysis length. Figure 4 illustrates the results for novice (left) and expert (right) participant.

We observe that in general, it takes in average 10.7 steps to reach a defined goal in GeoHighlight, i.e., 33 steps less than Tableau. This shows that the guidance component helps analysts discover their data and quickly reach to the goal. Level of expertise improves the analysis length in average by 4 steps. Interestingly, starting points do not have a huge influence. It is potentially due to the diversity component which provides distinct options. We also observe that *T2* is an easier task than *T1*. This is potentially due to similarity component where the analyst can request options similar to what she has already seen and greedily moves to

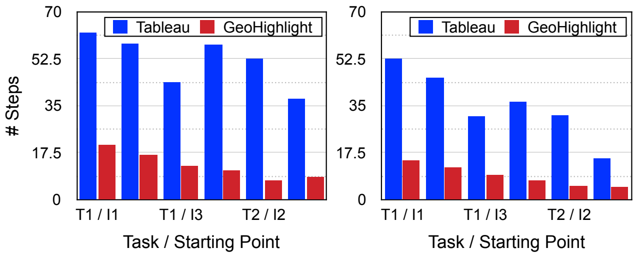

Figure 4: User Study

match profiles.

6. CONCLUSION

We addressed the problem of generic guidance and introduced GeoHighlight, the first efficient interactive highlighting approach in spatiotemporal data. We formulated our problem in form of a constrained optimization and proposed Highlighter, a greedy algorithm to highlight k-best points for a given point of interest within a time limit. We discussed genericness of our approach by materializing few examples from restaurant and taxi datasets. We also showed the efficiency and usability of our framework in form of performance experiments and user study. There are several directions of improvement for this work. Specifically, we want to consider an analyst profile vector which is built during interactive steps and will be exploited to return more analyst-tailored results.

7. REFERENCES

[1] Airbnb dataset.
http://insideairbnb.com/get-the-data.html.

[2] Nyc citi bike dataset.
http://www.nyc.gov/html/tlc/html/about/trip˙record˙data.shtml.

[3] Nyc taxi and limousine commission dataset.
http://www.nyc.gov/html/tlc/html/about/trip˙record˙data.shtml.

[4] Yelp dataset.
http://www.nyc.gov/html/tlc/html/about/trip˙record˙data.shtml.

[5] A. M. Aly, W. G. Aref, and M. Ouzzani. Spatial queries with k-nearest-neighbor and relational predicates. In *SIGSPATIAL*, page 28. ACM, 2015.

[6] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel. Recommendations in location-based social networks: a survey. *GeoInformatica*, 19(3):525–565, 2015.

[7] J. Freire, A. Bessa, F. Chirigati, H. T. Vo, and K. Zhao. Exploring what not to clean in urban data: A study using new york city taxi trips. *IEEE Data Eng. Bull.*, 39(2):63–77, 2016.

[8] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *ICDE*, pages 450–461, 2012.

[9] J. Liang and M. L. Huang. Highlighting in information visualization: A survey. In *2010 14th International Conference Information Visualisation*, July 2010.

[10] B. Omidvar-Tehrani, S. Amer-Yahia, and A. Termier. Interactive user group analysis. In *CIKM*, pages 403–412. ACM, 2015.

[11] A. C. Robinson. Highlighting in geovisualization. *Cartography and Geographic Information Science*, 38(4):373–383, 2011.

[12] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.

[13] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *TVCG*, 22(1), 2016.

[14] J. Yu, J. Wu, and M. Sarwat. Geospark: A cluster computing framework for processing large-scale spatial data. In *SIGSPATIAL*, page 70. ACM, 2015.