GeoHL: A Point-Recommendation Approach for Spatiotemporal Data

ABSTRACT

Spatiotemporal 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 spatiotemporal 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 spatiotemporal 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 spatiotemporal data in various fields of science. Analysis of such data is interesting as it is grounded on reality: each record represents a specific location and time. Moreover, understanding patterns and trends provides analysis insights leading to improved user planning and decision making. Some instance applications of spatiotemporal data are smart city management, disaster management and autonomous transport.

Traditionally, an exploratory analysis scenario on spatiotemporal data is described as follows: the analyst visualizes the data using an off-the-shelf product (e.g., Tableau[1], Spotfire[2]). Then she looks at different parts of data for interesting patterns and trends. With the growing size of spatiotemporal datasets, this classical approach is not practical anymore: geographical points are scattered everywhere and the analyst cannot effectively observe insights.

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

EDBT 2016

DOI: 10.475/123_4

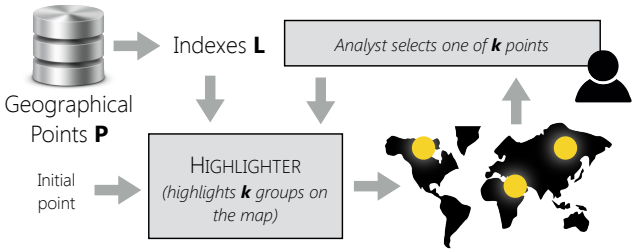

Figure 1: GeoHighlight Framework

To overcome this challenge, visualization environments offer a plethora of operations to filter out data. In practice, this doubles the problem: the analyst is left alone in a huge space of data and operations. In an exploratory context, the principled challenge for the analyst is *"what to see next"* during the analysis process. A *guidance mechanism* is necessary to point out potential future directions of analysis.

Given a geographical point of interest, the question is then how to recommend other points to be considered in future analysis steps in form of guidance. In this paper, we focus on one specific guidance approach, i.e., highlighting k-best points given a point of interest. Those k points should have high quality. Quality is formulated as optimization of two dimensions: *relevance* and *diversity*. Optimizing relevance ensures that recommended points are in-line with what the analyst has already liked. Optimizing diversity results points which are as different as possible from each other and unveil different aspects of analysis. Example 1 illustrates a common case in practice.

Example 1. *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*[3] *(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 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 spatiotemporal data analysis in

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

recent years, we point out following challenges for guidance: *i. Genericness.* Considering the heterogeneous nature of spatiotemporal 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 spatiotemporal 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 spatiotemporal processing [10], sub-second interactivity is still missing.

There exist few instances of information-highlighting methods [5, 7, 9, 8]. However all these methods are *objective* and do not apply to the context of spatiotemporal guidance where user feedback is involved. In terms of recommendation, few approaches focus on spatial dimension [2, 4] while the context (user feedback) 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.

2. 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.

3. PROBLEM STATEMENT

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.

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 [1] 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.

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

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 forumlate highlighting as an optimization-based problem where we optimize diversity and respect a bound on relevance.

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 expresses interest in obtaining high quality points in two dimensions at the same time (relevance and diversity) which makes the problem more challenging.

4. ALGORITHM

We propose a solution for GEOGUIDE by inspiring from both recommendation [6] and visual highlighting [5, 7] 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

Algorithm 1: HIGHLIGHTER Algorithm

Input: $p \in \mathcal{P}$, σ, k, $tlimit$
Output: $\mathcal{S}_p$

```
1  S_p ← get_top_k(L^p)
2  p_next ← get_next(L^p)
3  while (tlimit not exceeded ∧ relevance(p, p_next) ≥ σ)
   do
4      for p_current ∈ S_p do
5          if diversity_improved(S_p, p_next, p_current) then
6              S_p ← replace(S_p, p_next, p_current)
7              break
8          end
9      end
10     p_next ← get_next(L^p)
11 end
12 return S_p
```

easily extended to regions using point-clustering methods such as k-means.

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 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.

5. ILLUSTRATIVE SCENARIO

We illustrate an application of GEOHIGHLIGHT in a realistic scenario for New York taxi dataset[4]. This dataset has been frequently exploited for urban analysis (e.g., [3]). 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 100K unique points for the sake of clarity of results. We employ HIGHLIGHTER (Algorithm 1) with following parameters: $\sigma = 0.7$, $k = 5$ and $tlimit = 200ms$.

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

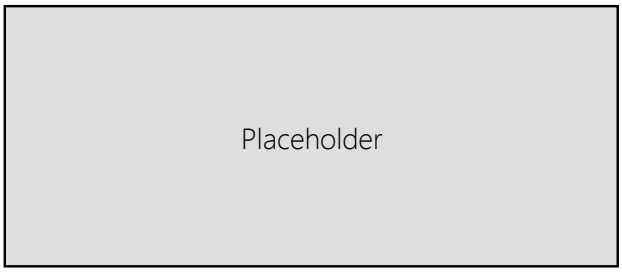

Figure 2: Application of GEOHIGHLIGHT on New York Taxi dataset

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. The point depicts a drop-off at *270 West 43rd Street* on January 9, 2014 at 10PM. HIGHLIGHTER then provides 5 relevant points to the selected point (Figure 2). Among 5 highlighted points, Lucas selects the 3rd one, i.e., a pick-up at *Columbus Avenue* near Central Park occurred approximately at the same time of the first selection. 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. Lucas looks for a good drop-off point which is in a neighborhood of the previous selection as the cab-idle station. Lucas selects the second highlighted point at *183 Duane Street* at [time?] as other highlights are around airport and train stations which have less taxi requests in the afternoon. This selection contributes to the heavy cab request in Manhattan island at that time of the day.

6. 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 to measure the influence of relevance, size constraint k, and time limit on execution time, and second, a user study to evaluate how efficient analysts can gain insights in our framework.

Experiment Settings. Unless otherwise stated, we use the same settings discussed in Section 5. 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.

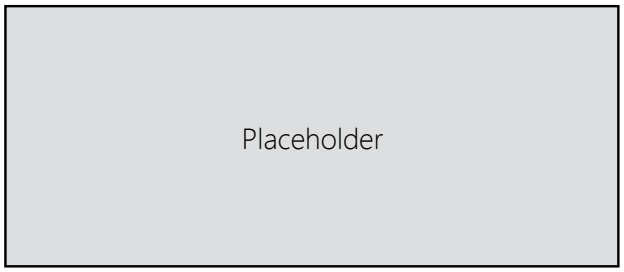

Figure 3: Performance Evaluation

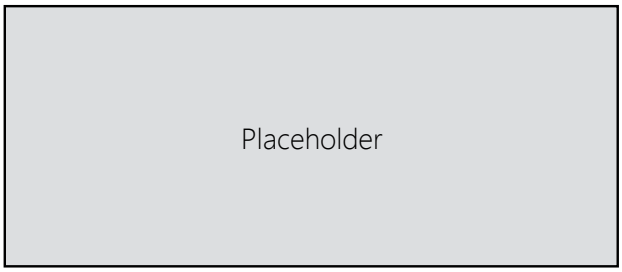

Figure 4: User Study

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 ($tlimit = 10s$) in order to evaluate the effect of relevance and size constraint (k) on execution time.

Figure 3 left illustrates the effect of size constraint by varying k from 2 to 500. We observe that [observation based on experiment results]. Figure 3 right illustrates the effect of relevance threshold by varying σ from 0.1 to 1.0. We observe that [observation based on experiment results].

User Study. The principled question that we ask ourselves is whether GeoHighlight is useful for analyst in practice. To answer this question, we designed a user study with 24 participants. 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. 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 miles. 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.

We observe that in general, it takes XXX steps to reach a defined goal in GeoHighlight. Level of expertise improves the analysis length in average by XXX step. This shows that the guidance component helps analysts discover their data and quickly reach to the goal. 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 match profiles.

7. 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 showed the efficiency and usability of our framework in form of scenarios, 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.

8. REFERENCES

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

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

[3] 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.

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

[5] J. Liang and M. L. Huang. Highlighting in information visualization: A survey. In *2010 14th International Conference Information Visualisation*, pages 79–85, July 2010.

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

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

[8] 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.

[9] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2016.

[10] J. Yu, J. Wu, and M. Sarwat. Geospark: A cluster computing framework for processing large-scale spatial data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 70. ACM, 2015.