An Interactive Guidance 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 urban management and social science. Discovering patterns and trends in this data provides improved insights for planning and decision making in several applications such smart city and disaster management. However, exploratory analysis of such data is a challenge due to its huge size 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 develop GeoGuide, an interactive guidance approach for spatial data. GeoGuide captures the feedback of analysts and exploits it to highlight potentially interesting analysis options. Demonstration attendees experience the web-based implementation of GeoGuide in various scenarios.

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 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 literature in spatial data analysis has a focus 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., [8]). 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 because *i*. she doesn't know what to investigate next in an exploratory context and *ii*. she may get distracted and miss interesting points by visual clutter caused by huge point overlaps. In other words, although iteration transitions (between one analysis step to the other) can be performed efficiently, the decision which forms a transition, remains unclear.

There exist few instances of information-highlighting methods in the literature [7, 10, 12, 11]. 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 [3, 6] while the context and result diversification are missing.

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 duplicates the problem: the analyst is left alone in a huge space of spatial 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.

Contribution. We demonstrate GeoGuide, an interactive framework to highlight a subset of geographical points based on analyst feedback. Although GeoGuide operates on points, its functionality can be easily extended to regions using point-clustering methods. The highlighted set facilitates the decision-making process by providing guidance on what the analyst should potentially concentrate on. The set of highlights is deliberated over high quality. We consider two quality metrics in GeoGuide: *relevance* and *diversity*. First, each highlighted point should be relevant to historical choices of the analyst. Second, highlights should be geographically diverse to unveil different aspects of analysis. Both quality metrics are interdependent to compute the set of highlights.

Despite literature contains several instances of feedback exploitation to guide the analyst in further analysis steps (e.g., [4]), the common used approach is the top-k processing methodology in order to prune the search space based on the explicit feedback and recommend a small subset of interesting results of size k. A clear distinction and contribution of GeoGuide is that it doesn't aim for pruning, but leveraging the actual data with potential interesting re-

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

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org.
Proceedings of the VLDB Endowment, Vol. 10, No. 6
Copyright 2017 VLDB Endowment 2150-8097/17/02.

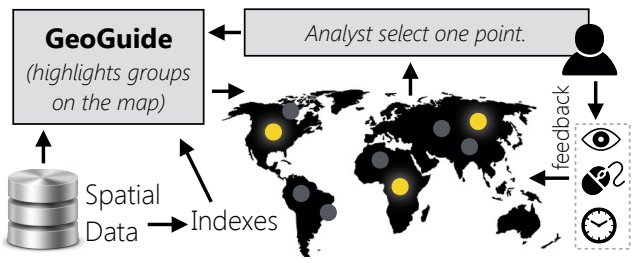

Figure 1: GeoGuide Framework

sults that the analyst may miss due to the huge volume of spatial data. While in top-k processing algorithms, analyst choices are limited to k, GeoGuide has a freedom of choice where highlights get seamlessly updated with new analyst choices. We present a system overview in Section 2 and our demonstration plan in Section 3.

2. SYSTEM OVERVIEW

Given a dataset with a set of spatio-temporal information points, our system is able to process and generates highlighted informations base on analyst preferences and behaviour. Our proposed framework is able to highlight different information based on specific data attributes, highlighting, for instance, each points by size or color intensity. Using GeoGuide framework the analyst can also define a subset of points to be highlighted over the dataset by a simple filtering action. The functionalities of GeoGuide are an inspiration from both recommendation [9] and visual highlighting [7, 10] methodologies. GeoGuide is a layer on top of a raw visualization to guide analysts in large-scale spatial data analysis. Figure 1 illustrates the main components of GeoGuide architecture.

The following example illustrates the challenge of our approach in practice.

Example 1. *Liam is planning a short trip to Paris. He decides to rent a home-stay from Airbnb website*[3]*. He is open to any type of lodging and he wants to explore different options (i.e., exploratory analysis). He queries all available locations in Paris with a fair price. His query results in 3000 locations. As he has no other preferences, an exhaustive investigation needs scanning each location independently which is nearly infeasible. In case he wants to focus on a smaller set of options, it is not clear which subset he needs to look at. While he is looking at primary locations in the list, he shows interest in having "balcony" as amenity and being close to Eiffel tower. An ideal system can capture this feedback in order to short-list a small subset of remaining locations that Liam should consider as high priority.*

In our framework, we consider a spatial 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 $\langle lat, lon, alt \rangle$ which denotes p's geographical coordinates (latitude, longitude and altitude respectively). 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 spatial dataset.

We also define a feedback vector $\mathcal{F}$ on the schema $\mathcal{A}$ initialized by zero. The vector gets updated by $\mathcal{A}_p$ whenever the analyst shows interest in a geographical point p. Feedback vector is always kept normalized, i.e., $\Sigma_{v \in \mathcal{F}}(v) = 1.0$. Unlike the literature which mainly focuses on explicit feedback (where the analyst should clearly reflect her likes and dislikes), we investigate on implicit feedback. This enables the system to capture *what the analyst may miss* instead of what the analyst has clearly investigated before. We consider different ways to capture implicit feedback.

- **Gaze Tracking.** During spatial data analysis, it is often the case that analysts look at some regions of interest but forget to provide an explicit feedback. For instance in Example 1, while Liam is focusing on home-stays close to the Eiffel tower, he also looks at farther locations with easy train access. However, he never clicks on those points. We call this latent signal, *gaze*. It shown in [5] that gaze has a strong correlation with "user attention". The signal can be captured by tracking eye movements aka saccades [1]. We employ ixLabs gaze tracking[4] as it only needs a simple web-cam to capture the gaze signal.

- **Cursor Tracking.** To address privacy issues of web-cam exploitation for gaze tracking, we consider an alternative option of tracking the mouse cursor. It is shown in [2] that mouse gestures have a strong correlation with "user engagement". Intuitively, a point receives a positive feedback if the cursor moves around it frequently.

- **Session Time.** In most spatial datasets, there is a profile page dedicated to each point. Examples are restaurant pages in Yelp and lodging pages in Airbnb. We consider the amount of time that the analyst spends in a page as an implicit feedback. For instance, if the analyst spends few minutes in a page for an Indian cuisine restaurant, this counts as positive feedback for this type of restaurants.

At each step of the analysis, GeoGuide highlights few points based on the feedback content $\mathcal{F}$. The highlighting decision is made based on two quality metrics, i.e., relevance and diversity.

Relevance. Highlights should be in the same line with analyst feedback (captured either by gaze, mouse cursor or session time). Note that we consider *contextual-based* relevance and not *distance-based* relevance. The reason originates from our data observation. For instance in a taxi dataset, consider a 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 relevant to another one in San Fransisco for a middle-age male customer for an itinerary of 8 kilometers and $2.5 tip. The relevance between a point p and the feedback vector $\mathcal{F}$ is defined as follows.

$$relevance(p, \mathcal{F}) = average_{a \in \mathcal{A}_p \cap \mathcal{F}}(sim(p, \mathcal{F}, a)) \quad (1)$$

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

[3] *http://www.airbnb.com*

[4] *http://www.xlabsgaze.com/*

which provides encapsulated similarity computations for any type of attribute.

Diversity. Highlighted points should also represent distinct regions so that the analyst can observe different aspects of data and decide based on the big picture. 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 Haversine distance shown in Equation 3.

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

GeoGuide employs a best-effort greedy approach to efficiently compute highlighted points. We consider an offline step followed by the online execution of GeoGuide. 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 step (as is commonly done in Web search). Each index $\mathcal{L}_p$ for the point p keeps all other points in $\mathcal{P}$ in decreasing order of their relevance with p.

During online execution, GeoGuide admits as input a point $p \in \mathcal{P}$ (the user explicit choice) and returns the set of highlights $\mathcal{H} \subset \mathcal{P}$. GeoGuide makes sequential accesses to $\mathcal{L}_p$ to greedily maximize diversity. Points in $\mathcal{L}_p$ get a weight using $\mathcal{F}$. Points with a larger weight (i.e., closer to the analyst feedback) have a higher chance to be in $\mathcal{H}$. To speed up comparisons with $\mathcal{F}$ vector, we exploit bit-wise comparisons. We convert both $\mathcal{F}$ and point p to boolean representations and compute relevance (Equation 1) using bit-wise operators.

GeoGuide does not sacrifice efficiency in price of value. We consider a *time limit* parameter which determines when the algorithm should stop seeking maximized diversity. Scanning inverted indexes guarantees the relevance even if time limit is chosen to be very restrictive. Our observations with several datasets show that we achieve the diversity of more than 0.9 with time limit set to $200ms$.

GeoGuide is implemented in Python (as the computation engine) and JavaScript D3 (as the visualization engine). Demonstration attendees can play the role of analysts in GeoGuide. We provide several spatial datasets in our demo session: Yelp dataset of restaurant check-ins with 229,907 geographical points, Airbnb dataset for short-term lodging with 4,200,000 points, New York taxi dataset with 173,179,759 points. Participants can also experience different types of feedback capturing such as gaze, mouse movement and session time. We describe three demonstration scenarios as follows.

3. DEMONSTRATION PLAN

Our demonstration plan consists of 5 parts. First, we would like to present to the VLDB attendees the diversity of parameters in three different scenarios and its datasets. Second, we will demonstrate the how to set the GeoGuide variables before start the analysis by choosing a specific point in the map. Third, the attendees will be able to effectively see the highlighted points generated by the environment, and its properties. Fourth, we will present how to align different filter types in order to improve the results. Fifth, we will present to the attendees the use of implicit and explicit feedbacks, e.g., (i) how we capture cursor tracking in order to capture implicit informations (ii) by explicit choosing different parameters to be highlighted in therms of size and colors.

Scenario 1. On New York taxi dataset, we demonstrate how GeoGuide can contribute to urban planning and fleet management. We consider an explicit goal of discovering which neighborhoods work the best for which drivers in order to increase the overall availability of cabs in the city. We show how a chain of cab stations can be picked by GeoGuide in diverse location of the city.

Scenario 2. On Airbnb dataset, we demonstrate how GeoGuide can contribute to approach a lodging of interest based on analyst's feedback. As instructed in the user study, we consider the concrete case of finding a cheap lodging solution with a balcony near Eiffel tower. The attendee will observe how feedback converges the exploration towards the goal very quickly.

Scenario 3. On Yelp dataset, e demonstrate how GeoGuide can contribute to reach an early consensus on a restaurant. The attendee will observe that his/her preferences will be immediately captured and reflected in future highlights. The attendee can experience session time feedback in this scenario.

1. Diversity of parameters in three different datasets.

2. Setting framework variables.

3. Highlighted points generated by the system.

4. Filtering different properties.

5. Implicit and explicit feedbacks.

To validate our design choices in GeoGuide (quality dimensions and feedback capturing), we design a user study with 24 participants (students in Computer Science). We define a task for each participant and ask him/her to fulfill the task using GeoGuide and Tableau (as the most advanced off-the-shelf visualization product). Then we measure the number of steps to reach the goal. We define two tasks, *T1: finding a point in a requested location* (e.g., find a home-stay in the Central Park area, New York), and *T2: finding a point with a requested profile* (e.g., find a cheap home-stay with balcony in Paris.) Participants may begin their navigation from three different starting points: *I1: close to the goal*, *I2: far from the goal*, and *I3: random*.

In Tableau, participants employ filtering and querying tools to reach their goals. In GeoGuide, participants benefit from relevant and diverse highlights and feedback capturing using cursor tracking. Figure 3 illustrates the results of this study. We report results for separate sub-populations: the left figure illustrates the results for novice participants (who don't know the location, be it Paris or New York) and the right figure illustrates expert's results.

We observe that in general, it takes in average 10.7 steps

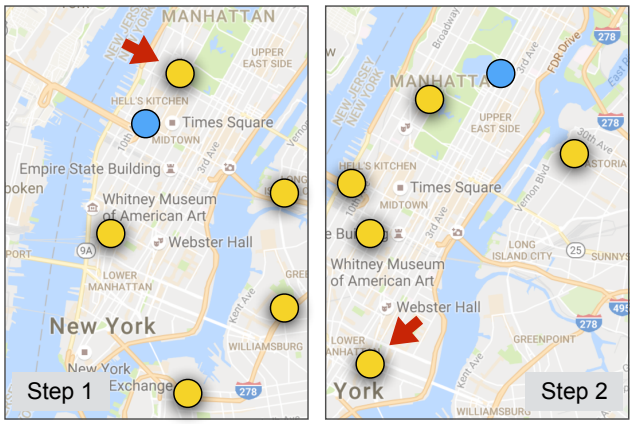

Figure 2: GeoGuide on New York Taxi Dataset

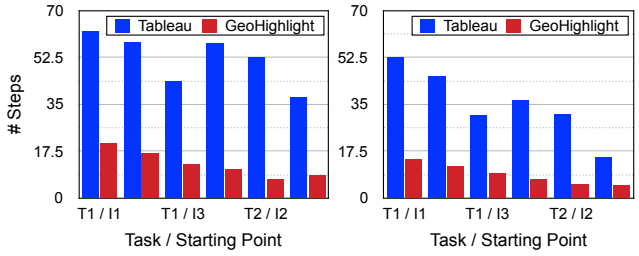

Figure 3: User Study

to reach a defined goal in GeoGuide, i.e., 33 steps less than Tableau. This shows that the highlighting component equipped with the feedback mechanism helps analysts discover their spatial data and 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 match profiles.

4. REFERENCES

[1] I. Arapakis, M. Lalmas, B. B. Cambazoglu, M.-C. Marcos, and J. M. Jose. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, 65(10):1988–2005, 2014.

[2] I. Arapakis, M. Lalmas, and G. Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1439–1448. ACM, 2014.

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

[4] M. Boley, M. Mampaey, B. Kang, P. Tokmakov, and S. Wrobel. One click mining: Interactive local pattern discovery through implicit preference and performance learning. In *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*, pages 27–35. ACM, 2013.

[5] M. H. Fischer. An investigation of attention allocation during sequential eye movement tasks. *The Quarterly Journal of Experimental Psychology: Section A*, 52(3):649–677, 1999.

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

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

[8] L. Lins, J. T. Klosowski, and C. Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, 2013.

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

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

[11] 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.

[12] 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.