A Guidance-based Visualization Framework for Spatiotemporal Data

Behrooz Omidvar-Tehrani
Department of Computer Science
The Ohio State University
omidvar-tehrani.1@osu.edu

Plácido A. Souza Neto, Gustavo Guerino
Federal Institute of Rio Grande do Norte
IFRN, Brazil
placido.neto@ifrn.edu.br, gustavo.guerino@academico.ifrn.edu.br

***Abstract*—This paper presents GEOHIGHLIGHT, an interactive guidance-based visualization framework for spatiotemporal data. The main objective of our proposal is provide an environment to guide analysts through the challenge due the huge size and diversity of spatiotemporal data in order to retrieve important information by guidance techniques. To overcome this challenge, in general, visualization environments offer a plethora of operations to manipulate data (filter, aggregate, etc.). In practice, this duplicates 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. To fulfil this need in geo analysis, we propose a *guidance mechanism* to point out potential future directions of analysis.**

I. INTRODUCTION

There exists, nowadays, a great amount of spatiotemporal data in various fields of science. Understanding patterns and trends through visualizing spatiotempral data improves analyst planning and decision making. Given a geographical point of interest, the main question for an analyst is then how to recommend other points to be considered in future analysis steps in form of guidance. In this paper, our proposal 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.

In this paper, we argue the need for a guidance-based visualization framework for spatiotemoral data by considering following desiderata for this environment.

Genericness. The framework component should be agnostic (making no assumption) about the dataset type, attributes and distribution.

Limited Options. The guidance approach should provide a limited set of recommendations because too many options distract the analyst.

Relevance. The guidance approach should deliver results which have similar characteristics to what the analyst has already liked.

Diversity. Recommendations should represent distinct regions so that the analyst can observe different aspects of data and decide for the next analysis iteration.

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. Despite progress in efficient spatiotemporal processing [6], sub-second interactivity is still missing.

We inspire from both recommendation [3] and visual highlighting [2], [4] methodologies and propose GEOHIGHLIGHT as a solution to aforementioned challenges. GEOHIGHLIGHT is a visualization framework for spatiotemporal data which guides the analyst throughout the process towards interesting points. In this paper, we propose this guidance approach for analysis of huge datasets with geographic informations. So, the analyst considers the guidance and picks a direction for the next analysis iteration.

II. GEOHIGHLIGHT

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.

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.

Figure 1 illustrates the general structure of GEOHIGHLIGHT. We observe that from a dataset with spatialtemporal information points, an analyst can select one of k points to search for similar other points in order to proceed with specific decision like: (i) considering a Taxi analysis for pick-up a new passenger after a drop-off, *which are the best points in the neighborhood to pick-up new client?*; or (ii) considering a bicycle ride analysis, *which are the best points a user can leave the bicycle considering that this user will pick-up the bike in a specific location and plan to pedal for 1 hour?*

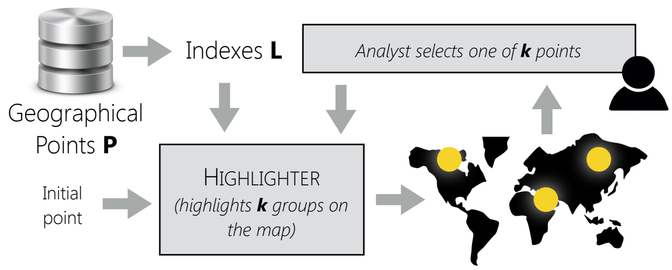

Fig. 1. GEOHIGHLIGHT Framework

III. DEMONSTRATION

This demonstration allows to observe thoroughly the inner working of GEOHIGHLIGHT, and to illustrate the use of the environment from the highlighting and guidance results in order the help the data analyst. The demonstration platform and its architecture were implemented as following: (i) the GEOHIGHLIGHTengine is implemented in Python [1], and (ii) the graphical user interface is based on Node.js, using different libraries like *D3.js*[2], *Google Maps API*[3] *and Ajax*[4]. The client-side part allows the client to choose the values for a clearly-defined set of initial parameters to visualize the information about the execution of GEOHIGHLIGHTwith the chosen values.

Demo. We consider for this demonstration two realistic scenario for New York taxi[5] and bicycle[6] datasets. These datasets has been frequently exploited for urban analysis (e.g. in [1]). The New York taxi 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 New York bicycle dataset contains information from 2013 to 2016, and 15 attributes such as start station id and end station id, their respective geographical informations, trip duration and distance. These scenarios illustrate how an analyst can achieve an exploratory analysis goal. We preprocessed the original datasets and considered a subset of 20K unique points for the sake of clarity of results.

Scenario 1. The scenario consists of a data scientist, we name Lucas, 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 goals 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. Considering this, we present how GEOHIGHLIGHTcan support Lucas by retrieving quality information to his analysis. Figures 2 and 3 present the GEOHIGHLIGHTdashboard for this scenario, where the analyst can choose a initial point from a set of points. Then, the framework the presents k other potential points that may be consider in Lucas analysis from the set of parameters (σ, k and a $timeLimit$) defined. The same process can be made again in a *second pass*.

[1] *https://www.python.org/*
[2] *https://d3js.org/*
[3] *https://developers.google.com/maps/*
[4] *http://api.jquery.com/jquery.ajax/*
[5] *https://data.cityofnewyork.us/view/gn7m-em8n*
[6] *https://s3.amazonaws.com/tripdata/index.html*

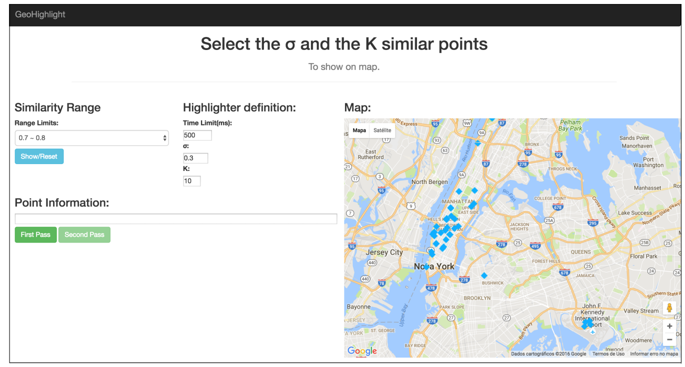

Fig. 2. GEOHIGHLIGHT Dashboard [New York Taxi Scenario] - View of a set of taxi pick-up and drop-off points.

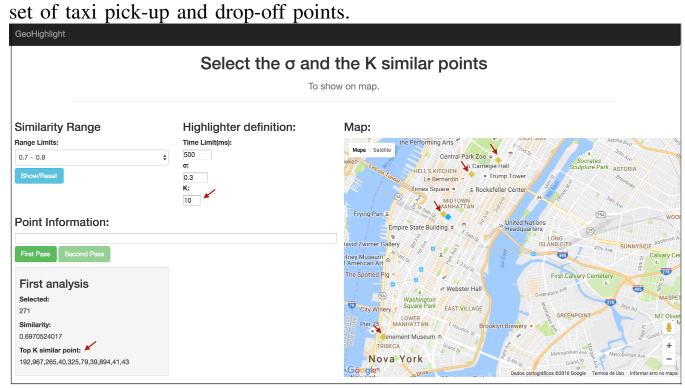

Fig. 3. GEOHIGHLIGHT Dashboard [New York Taxi Scenario] - Choosing a specific point from σ, k and a $timeLimit$.

Scenario 2. Another data scientist, we name Shadi, whose task is to optimize New York bicycles station location in New York by analysing the rides from January until September 2016. How to identify problems considering the day and time for drop-off bike rides in full stations. She must to identify the problematic bike stations to pick-up and drop-off .

IV. CONCLUSION

In this demo we have presented GEOHIGHLIGHT, a generic and interactive framework for visualization of spatiotemporal data. Some future extensions include the integration of generic query approach based on [5], and we also want to consider an analyst profile vector which is built during interactive steps and will be exploited to return more analyst-tailored results.

REFERENCES

[1] 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.

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

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

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

[5] M. Vartak, S. Rahman, S. Madden, A. G. Parameswaran, and N. Polyzotis. SEEDB: efficient data-driven visualization recommendations to support visual analytics. *PVLDB*, 8(13):2182–2193, 2015.

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