User Learning In Interactive Data Exploration

Sanad Saha*
Oregon State University
Corvallis, United States
sahasa@oregonstate.edu

Nischal Aryal*
Oregon State University
Corvallis, United States
aryaln@oregonstate.edu

Leilani Battle
University of Washington
Washington, United States
leibatt@cs.washington.edu

Arash Termehchy
Oregon State University
Corvallis, United States
termehca@oregonstate.edu

***Abstract*—Users explore large, complex datasets to find interesting hypotheses and previously unseen insights. In this process, known as data exploration, users often generate database queries *without any precise goals or concrete information need*, posing challenges for database systems that assume the user has a clear intent a priori. In response, system developers often model users' exploration strategies over time, which could enable the system to predict and adapt to users' subsequent actions. However, current models generally treat users' exploration behavior as *static*, whereas in reality, users *dynamically change their behavior* in response to what they *learn* during exploration. In this paper, we present an analysis of existing data exploration logs to quantify shifts in users' data exploration strategies over time. Our analysis confirms that users shift their behavior over time, and state-of-the-art learning algorithms struggle to adapt to this evolution, revealing new avenues for building more accurate models of user exploration behavior within data exploration systems.**

I. Introduction

Interactive data exploration (IDE) is often an *open-ended* process where the user has no clear objective and delves into the data to discover intriguing observations. However, exploration can also involve *goal-directed tasks* with precise analysis goals or a concrete hypothesis in mind [1]. Thus, data exploration spans a wide spectrum of potential analysis tasks.

Database management systems (DBMS) often struggle to support IDE workloads due to the wide range of potential queries analysts can generate. Furthermore, the ad-hoc nature of exploration queries can easily thwart standard optimization techniques, leading to slow DBMS query response times. Studies have shown that slow response times can lead to user frustration and even abandonment of exploration tasks [2].

For a smoother IDE experience, analysts often use data exploration systems (DES), which help them to explore large datasets by comprehending their information needs and guiding them towards their desired information through user-friendly visual summaries [1]. A common approach to designing DES is to model a user's exploration behavior to provide customized support matching the user's visualization interests, preferences, and analysis strategies. Leveraging this user model, DES can recommend visualizations from relevant, interesting data areas and suggest exploratory operations that may help users attain results effectively and efficiently. Furthermore, the DES can further improve interactive user experience by prefetching promising data regions [3].

*Both authors contributed equally to this work.

However, an implicit assumption in many DES is that the user's analysis goals and strategies to reach them are fixed, i.e., will not change much over time [3]. But recent research in information search indicates that users learn to improve their exploration strategies by modifying their keyword queries over time [4]. We posit that the evolutionary nature of IDE shares strong parallels with query-based information searching.

The goal of this study is to investigate the evolution of users' learning behavior during open-ended and goal-directed IDE tasks. We perform statistical analyses to determine whether or not such evolution exists. Our results suggest that *analysts do in fact change their decision-making strategies during exploration*, indicating that their analysis behaviors may be driven by what they learn as they explore. Furthermore, we employ state-of-the-art learning algorithms to investigate whether they can adapt to users' evolving exploration behavior and predict their future interactions. Access to these models will help DES adapt to users' learning behaviors and assist them in IDE. We make the following contributions: (a) Our preliminary findings on real-life IDE tasks demonstrate meaningful shifts in users' exploration strategies. (b) We connect users' learning behavior during data exploration to their preference for fixed or dynamic exploration policies. We argue that it is advantageous for the DES to adapt to users' preferences because it can improve users' experience and the quality of the results. (c) Our empirical analysis of state-of-the-art learning methods for modeling human learning, successfully applied in other fields, including reinforcement learning and neuroscience, reveals shortcomings in adapting to users' dynamic information needs.

TABLE I
Characteristics of Data Exploration Tasks

Characteristics	imMens user study [5]	ForeCache user study [3]
Exploration need	Open-ended	Goal-directed
Prior experience	15min	Training subtask
Time restriction	30min	No
Task Complexity	Querying Summarized Plots	Pan/Zoom 2D map

II. User learning in data exploration tasks

It is observed that the objective and properties (i.e., characteristics) of *data exploration tasks* may significantly influence users' interactions with an interactive DES [1]. Rather than conducting a separate user study to collect data, we observe that visualization researchers [3], [5] have already conducted influential studies in this space and have publicly shared their study data. We select two of these datasets (Table I) to study the effect of user learning on user's exploration strategies.

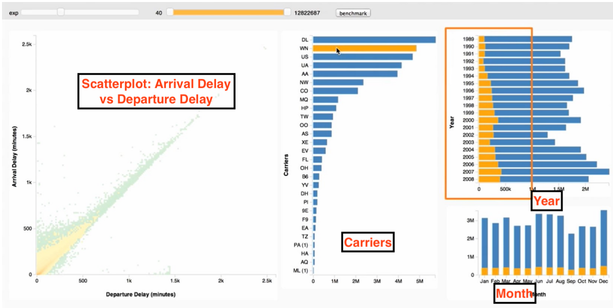

Fig. 1. Exploring different data areas with a DES [5].

To analyze how and whether users' exploration strategies evolve while completing these tasks, we first examine *users' interactions* with the corresponding DES. For example, to complete the imMens exploration task [5], at each step t, the user selects one visualization v_t to interact with from a pool of K visualizations and analyze the information in the interface (Figure 1). To better understand the evolution of user behavior, we statistically analyze changes in users' exploration strategies. First, we divide each user's exploration session into two distinct phases. Then, we perform statistical tests using linear mixed-effects models [5] to examine the evolution of users' exploration strategy (i.e. probability of interacting with each visualization) between the two phases. We chose to use a 50-50 split, as it represents the most fundamental level at which we would expect to observe changes in strategy *We find that users learn about data during exploration, and their data focus shifts, which causes significant changes in their exploration strategies in the later half compared to the initial.*

We define how users explore the dataset and make informed choices of exploration strategies, as *the learning problem*. We draw parallels between users' learning problem and objective functions from well-studied online learning frameworks, such as Multi-Arm Bandits (MAB) and Markov Decision Process [6]. *For instance, in the imMens task (Figure 1), the user's decision to select one visualization from a set of K is consistent with a K-arm MAB problem.* Like a bandit agent, the user must learn an optimal policy that maximizes the information received by interacting with visualizations over time (i.e., reward). As neither visualization researchers nor database researchers have tested for the exact objective function humans use during IDE scenarios, our investigation encompasses a range of online learning frameworks.

III. Performance Evaluation

We utilize popular learning algorithms from reinforcement learning, economics, cognitive psychology, and neuroscience to model users' exploration behavior (See [6] for algorithm and modeling details). Furthermore, user learning may vary in different stages of data exploration. To capture this nuance, we set different training thresholds for our algorithms. We establish a consistent measure of how well these learning algorithms can adapt to users' exploration strategy by empirically evaluating their performance in predicting users' future interactions (e.g., the next visualization in the imMens exploration task).

We find that for *goal-directed tasks*, users' interactions are predicted with 68% accuracy by the Reinforce algorithm (Figure 3). Selected algorithms struggle more for *open-ended tasks* (Figure 2). Where the top-performing model, Contextual multi-arm bandit, has an average accuracy of 53.2%. This shows that the user strategies and learning schemes' are *correlated with the task characteristics* (Table I).

Our analysis shows that it is possible to model the evolution of users' exploration behavior for the *average user* (Reinforce in Figure 3 is the top model for 70% of the users). We also see *simple algorithms* modeling user exploration behavior better than complex ones, e.g., Greedy with 50% accuracy (Figure 2). These findings allude to a broader hypothesis that *current learning models may not be able to capture the evolution of users' exploration behavior, especially in open-ended tasks.* Subsequent studies are needed to develop more accurate models of users' data exploration behavior and learning-aware DES.

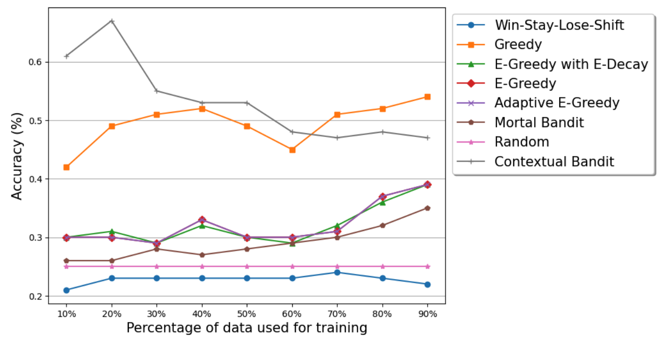

Fig. 2. Algorithm accuracy on different training data in open-ended task

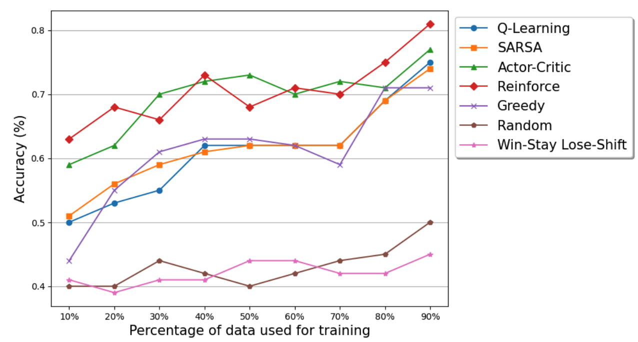

Fig. 3. Algorithm accuracy on different training data in goal-directed task

References

[1] L. Battle and J. Heer, "Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau," in *Computer graphics forum*, vol. 38, no. 3. Wiley Online Library, 2019.

[2] Z. Liu, B. Jiang, and J. Heer, "immens: Real-time visual querying of big data," in *Computer Graphics Forum*, vol. 32, no. 3pt4. Wiley Online Library, 2013, pp. 421–430.

[3] L. Battle, R. Chang, and M. Stonebraker, "Dynamic prefetching of data tiles for interactive visualization," in *Proceedings of the 2016 International Conference on Management of Data*. ACM, 2016.

[4] B. McCamish, V. Ghadakchi, A. Termehchy, B. Touri, and L. Huang, "The data interaction game," in *Proceedings of the 2018 International Conference on Management of Data*, 2018, pp. 83–98.

[5] Z. Liu and J. Heer, "The effects of interactive latency on exploratory visual analysis," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2122–2131, 2014.

[6] S. Saha, N. Aryal, L. Battle, and A. Termehchy, "How does user behavior evolve during exploratory visual analysis?" 2023. [Online]. Available: 10.48550/arXiv.2312.09407