A Survey on Interaction Log Analysis for Evaluating Exploratory Visualizations

Omar ElTayeby
University of North Carolina at Charlotte
oeltayeb@uncc.edu

Wenwen Dou
University of North Carolina at Charlotte
wdou1@uncc.edu

ABSTRACT

The trend of exploratory visualization development has driven the visual analytics (VA) community to design special evaluation methods. The main goals of these evaluations are to understand the exploration process and improve it by recording users' interactions and thoughts. Some of the recent works have focused on performing manual evaluations of the interaction logs, however, lately some researchers have taken the step towards automating the process using interaction logs. In this paper we show the capability of how interaction log analysis can be automated by summarizing previous works' steps into building blocks. In addition, we demonstrate the use of each building block by showing their methodologies as use case scenarios, such as how to encode and segment interactions and what machine learning algorithms can automate the process. We also link the studies reviewed with sensemaking aspects and interaction taxonomies selection.

Keywords

Interaction log analysis; explorative visualizations; insight-based evaluations; sensemaking process; interaction taxonomies

1. INTRODUCTION

The process of evaluating visualizations, in the past couple of years, has shaped many related aspects in the VA field. VA researchers have conducted evaluations in order to reach the following general goals: Understand how insights are derived, and improve the design, usability, aesthetics and visualization cognitive aspects (more details can be found in [34]).

Many of the research works have gone "beyond analyzing time and error", which is the most primitive quantitative method that can be applied. Time and error are important measures for evaluating the user's performance and the tool's usability [27, 3]. With the growing demand of developing and evaluating visualization for "exploratory" uses researchers have adopted insight-based evaluations. The needed evaluations for exploratory analyses are more complex than just counting the number errors and measuring the completion time in user studies.

This survey emphasizes the importance of evaluations that used interaction logs gathered from provenance tools. In the 2014 BELIEV workshop Smuc [41] addressed the connection between error and insights analysis. He divided Reason's model with respect to interactions according to the type of insights gained. Using the same concept we can see that there is a lot of research potential in reaching the goals mentioned previously by analyzing the interaction logs. In the recent years, the trend of analyzing user interaction logs has been growing, thus, we survey papers on visualizations' interaction log analysis and its related work to identify technical challenges and methodologies, and propose directions for future work.

2. SURVEY METHODOLOGY

Our sole focus is on the methodologies that used interaction logs to evaluate visualizations; therefore, in this section we start with defining interactions in the visualization context. However, in the next section we give an overview on different data sources that are commonly used besides interaction logs.

Interaction in the context of visualization refers to "the dialog between the user and the system as the user explores the data set to uncover insights" [43]. In the papers we surveyed, interactions refer to actions that are taken by the user to maneuver the visualization components. Logging of such actions produces interaction logs.

Firstly, we carried out two rounds for collecting the surveyed papers: in the first round we collected papers that are related to interaction logs analysis, either for the purpose of evaluation or provenance; in the second round we collected papers in which their research informs the analysis of interaction logs (such as user interaction taxonomies, sensemaking models, etc.) Secondly, we summarized those papers to identify similar goals and divide them as shown in section 4. Thirdly, we identified common methods used in each paper, then compared between those methods in order to construct the building blocks presented in section 5. For each common step we compare between the methodologies presented in the papers, despite the fact that, their frameworks did not follow the same order of steps.

3. DATA SOURCES

In the context of automating the evaluation of visualizations, researchers proposed approaches to analyze a combi-

BELIV '16, October 24 2016, Baltimore, MD, USA

DOI: http://dx.doi.org/10.1145/2993901.2993912

nation of multiple data sources. In this section we give an overview on the data types used in previous works, which are used with interaction logs in the analysis: eye-movement, think-aloud protocol or note taking, and screen video capture of the user study session.

Ivory and Hearst [19] have organized the methods of usability evaluation, for user interfaces in general, into a taxonomy according to the method's class and type, automation type and effort level. For each method type they recommended an automation type. Since visualization tools are a more specific type of user interface, the automation methods available are restricted to the analyses which are possible to implement, especially for understanding the process of deriving insights. The most promising data types that can be used for automating the evaluation are the interaction logs and the eye-movement data, the other two types mentioned are used for the purpose of qualitative analysis to confirm and cross-validate the findings. However, as the verbal records and screen video capture are necessary to mark the insights that the user have made, they require a lot of manual work which needs to be minimized from the automation perspective.

Interaction logs hold useful potential for usability studies in many research and industrial areas; i.e. Salesforce, Google Analytics and many other companies use the interaction logs of users to analyze their behavior with websites and search engines [7, 18, 31]. On the evaluation side of visualizations, interaction logs provide detailed information about the sequence of steps which the users have taken in order to arrive at their findings and reveal the spatio-temporal patterns [15, 6, 12, 9, 29]. Kang et al. [23] extracted the activity patterns of users for an investigative analysis case study in Jigsaw. Blascheck et al. [4] leveraged both interaction logs and eye-movement data and applied visual analytics techniques for detecting spatio-temporal patterns. Most commonly the user's logs consist of a timestamps, views (in case of multiple view visualizations) and the data which the user interacted with, and the type of action or action details [15, 4, 16, 9].

Eye-movement eye-tracking technology have grown in the past decade for HCI usability studies [35, 20], eye tracking devices facilitated the detection of eye-fixations, gaze durations, and saccadic velocities and amplitudes [39]. HCI usability studies utilized the eye-movements data in quantitative and qualitative comparisons between user interface design choices. Andrienko et al. [2] juxtaposed the structure of the geographical movement data to the eye-tracking data structure in a spatio-temporal sense. They summarized the analysis tasks of previous work into two main categories: one focusing on the areas of interest (AOIs) and another on the movements. The first entails the user's attention and therefore the eye-fixation is extracted. On the other hand, the jumps (saccades) are extracted for analyzing the movements. They listed several analysis methods for both tasks such as summarizing the map of spatial distribution demoting user's attention areas, clustering similar users temporarily, and extracting events from trajectories. Methods applied for analyzing movement data [1] are the original inspiration for the originality of these methods. Back to our main set of research goals for VA system evaluations, the eye-movement data provides information on the user's strategy [25], which is important for understanding the sensemaking process [14]. Such cognitive information can also be inferred from the interaction logs. Therefore, both data types are complementary and can easily be used for automating the evaluation methods.

Think-aloud records & Screen video capture [21] both data types are used to provide qualitative explanations for the interaction logs analysis; some approaches used them as a complementary analysis to confirm the conclusions from the quantitative analysis [15, 40]. In user studies the users reports their findings and thinking process either vocally or by taking notes, which is sometimes considered distracting according to [13]. Thus, other approaches [4] prefer to capture screen videos when the user study involves tasks that are sensitive to interruptions and could interfere the user's performance. Despite the advantage of the screen video capture has for not interfering the user's analysis process, the think-aloud records reveal more semantic information about the user's intentions. Think-aloud protocol is much less intrusive, since note taking requires user's attention for writing down their thoughts [11]. Both methods require post-hoc manual analysis, where the researcher encodes the actions or sets of actions either into a higher level of actions from the taxonomies mentioned before or to the corresponding low-level interaction logs.

4. RESEARCH GOALS

Since this survey is focused on analyzing the interaction logs with respect to evaluating visualizations and user performances, we classify previous works' research goals into to three evaluation scenarios [26]. Our categorization is inspired by Lam et al. [26] in which they provided 7 types of evaluation scenarios instead of evaluation methods. Our evaluation scenarios using interaction logs include: evaluating user performance and strategy (section 4.1), understanding insight generation and sensemaking (section 4.2), and evaluating visualization design (section 4.3).

4.1 Evaluating user's performance & strategies

The most common set of goals that researchers have focused on in the past few years concern two aspects: one is the user's performance and behavior; the other is the ability of the visualizations on facilitating users' reasoning processes. Measuring the user's behavioral model addresses the problem of traditional design methods "one size fits all" [44]; not all users have the same cognitive ability or personal traits to use the same functionalities provided by the visualization. Visualizations should be flexible enough to be easily used by users with different technical skills and cognitive abilities. Ziemkiewicz et. al [44] pointed out the importance of analyzing the user's thinking process which would lead to designing visualizations that are able to extend according to the user's cognitive ability and hit a wide variety of audience. The application of such analysis is useful for designing adaptive and personalized interfaces, as it is the main practice of HCI researchers [28]. For example, HARVEST [12] is a tool that automatically detects user's patterns in order to provide dynamic visualization recommendations. In addition, it provides a history panel for reusing previous analytic thinking in new contexts and with new data. Ziemkiewicz et. al also emphasized on the challenges that face researchers when mapping the effects of different visual designs choices on users' different personality traits and cognitive abilities, and thus, researchers in the community are eager to provide

visual mapping schemas. As they [44] surveyed research efforts towards the challenge of adapting visualization designs for user's differences, they divided the factors affecting the user's performance into visual literacy, cognitive and personality factors (traits). One of their goals is to build models of users' personality and cognitive ability by logging their interactions.

Continuing this research direction, Brown et al. [6] aimed to infer user's locus of control, extraversion, and neuroticism from user interactions. Additionally, they predicted the user's task performance in real-time using logged user interactions with a gamified visualization, "Finding Waldo". To explore the question of how users approach investigations using visual analytical tools, Kang et al. [23] compared two of Jigsaw's settings with traditional tools. The main goal is to test their evaluation methodology for investigative analysis visualizations, which led to design suggestions for Jigsaw and new evaluation metrics. While Dou et al. [9] and Lipford et al. [29] analyzed the interaction logs to recover the user's reasoning process, their goal is to explore whether it is possible to glean higher-level reasoning process from analyzing user interaction logs. Dou et al. [9] constructed an Operation Analysis Tool (OAT) to analyze the interaction logs from a user study on WireVis [8], and Lipford et al. [29] aimed to use this OAT for their user study's subjects to help them recall their strategies and findings. The subjects have indicated much more confidence in their ability to recall their initial analysis strategies when using the OAT in comparison when not using it. The results of this user study are a legitimate proof that interaction logs can be used for identifying the strategies that are supported by the VA systems. Along with researching the user's reasoning process, Blascheck et al. [4] built a VA tool for evaluating the efficiency of visualizations by automatically detecting patterns of users' interaction, eye tracking and think-aloud data. This combination of data sources enables evaluators to find deeper insights about the user's behavior. They used a heuristic-based approach to evaluate their VA tool, where experts reviewed their interaction logs from two case studies on VarifocalReader [24] and Word Cloud Explorer [17]. The case studies helped them to understand the analyst's behavior in navigating hierarchical visualizations.

4.2 Understanding insights generation & sensemaking process

Understanding the sensemaking process when using visualization tools is an important goal that researchers have been seeking recently; Sacha et al. created knowledge generation models special for visual analytics [38]. The main challenge when validating such models is collecting information about user's cognition state. Unfortunately, the only technique for gathering this information are think-aloud and note taking which are impractical when coding the collected information for large number of participants in user studies. Thus, researchers have endorsed the idea of translating interactions to insights, which we suggest, needs to be standardized using available taxonomies.

From interactions to insights mapping perspective Saraiya et al. [40] aimed to study user's exploratory behavior by analyzing factors that affect the insight generation. Their evaluation aimed to investigate explorative behavior only; they focused on testing the capability of the tool to support hypothesis generation. They defined 8 characteristics of insights, which can be used for evaluating visualization tools in general and some of those characteristics were inspired from their prior work [32, 33]. Although, the study in [40] was tailored for Bioinformatics applications, Reda et al. [37] used a reduced set of the insight characteristics from Saraiya et al. Reda et al.'s concept is based on separating the mental and interaction states and finding the transitions between them; this separation is seen when the users offload their cognition onto the interface and perceive new information from it. Their future goal is to investigate the higher-level patterns for these transitions to link these transitions with general sensemaking models.

On the other hand, Gomez et al. [11] extended Saraiya et al.'s [40] method to design a hybrid of both insight- and task-based methodologies, LITE (Layered Insight- and Task-based Evaluation). They used it to evaluate spatiotemporal visualizations proving that Saraiya et al.'s method is applicable to evaluating visualizations other than Bioinformatics visualizations. North et al. [33] addressed the problems for applying benchmark tasks for the purpose of insight-based evaluations, while Gomez et al. were able to overcome these problems.

4.3 Evaluating visualization design

In user studies, measuring the sole effect of individual visualization components is quite challenging when it comes to separating different factors that support and hinder insight generation. Guo et al. [15] applied a hybrid evaluation approach to find interface design factors that affect insights generation. The main research question that motivated this work came from understanding how the interactions lead to insights by analyzing the user's interaction logs and the influence of the visualization design on the insights gained. As a case study they applied their approach on a visual analytics tool, and were able to give design recommendations for eliminating the difficulties that prevented insights to be gained. Their data-driven (interaction logs) design recommendations were high-level and supported by qualitative analysis results. The nature of these findings waived their need from separating different possible factors that affect insight generation.

Besides HARVEST, Tome [10] is another framework for automating the evaluation of interface design using interaction logs from crowd of users. The main advantage of Tome is providing user's time-completion predictions after making design changes from interaction logs that were collected before making those changes; evaluators do not have to repeat their crowd user studies. Another advantage is evaluating the sole effect of each interaction individually on a low-level. This advantage is a contribution towards overcoming the main challenge we address in this subsection. In contrast with Guo et al.'s high-level approach for design recommendations, Tome has taken an atomic approach to overcome the same challenge.

5. BUILDING BLOCKS

We identify common steps from prior methods as building blocks for constructing a framework to analyze interaction logs. We chose the term building blocks as opposed to a pipeline for the reason that not all steps have to be taken or the order could change. Figure 1 presents the building blocks. Each block is an abstraction of multiple methods that have been applied to interaction log analysis. To

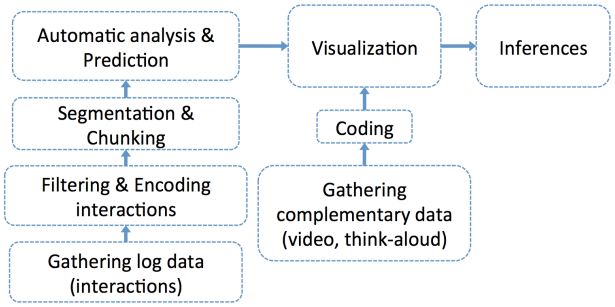

Figure 1: Building blocks for evaluations using interaction log analysis.

Table 1: Methods used for each building block

Building block	Methods
Gathering log data	Recording interaction logs and eye-movement
Gathering complementary data	Recording screen video capture and think-aloud protocol
Filtering	Post-hoc or ad-hoc processing
Encoding	Mapping raw interactions to taxonomy categories
Segmentation and Chunking	N-gram, LCS
Automatic analysis & Prediction	SVM, decision trees, PCA
Visualization techniques	Timeline, stacked multiple timelines, Dendrograms, decision trees

provide more details, we discuss methodologies within each building block. We summarize the methods relevant to each building block in Table 1, and elaborate on these methods in the subsections. Note that many papers present methods that fit into multiple building blocks, therefore the relevant pieces are mentioned in multiple subsections. We hope this summary provide an overview of the state-of-the-art efforts towards analyzing interaction logs for evaluation.

The structure of our building blocks is inspired from Guo et al.'s [15] evaluation pipeline. We made an effort to separate the collection of interaction logs and other complementary data since their analysis methods are fairly distinct. In addition, we use broad terminologies to describe the process of each block in order to fit the wide range of surveyed methods.

5.1 Encoding interactions & filtering

Before predicting the interaction patterns or measuring performance using the interaction logs, researchers perform two important steps to increase the accuracy of their models: they filter redundant and byproduct interactions, and then map the interface's specific interactions into standardized categories. Encoding the interactions requires taxonomies to produce understandable patterns that entail cognitive meaning. Previous research works that developed interaction taxonomies have helped the community to apply powerful evaluation methods using interaction logs and bridged the gap between quantitative and qualitative evaluations. Addressing the need of bridging this gap is the main reason behind the emergence of taxonomies.

Taxonomies enable categorization of interactions in order to convert between low- and high-level tasks, and there is no such perfect taxonomy that covers all scopes and applications. In their literature review, Brehmer et al. [5] addressed the challenge of analyzing tasks for identifying user behavior and classified the developed taxonomies into three different groups that focused on: low-level tasks, high-level tasks, and user behavior. They summarized many of the recent works and compared between their applications and scopes to construct a multi-level typology that abstracts visualization tasks using the how, why and what questions to describe the tasks.

Taxonomies were developed for different purposes and visualization applications, and therefore, the choice of the taxonomy is a crucial step. The most two important factors for choosing taxonomy are the accuracy of mapping the raw interactions and the coverage of the taxonomy categories on all of the raw interactions. In some cases, visualization interactions do not exactly match any particular taxonomy; therefore, researchers would modify the taxonomy's categories to meet their needs. These modifications depend on the level of the taxonomy and its categories' definitions, since definitions of similar categories might differ in different taxonomies. The comparison between Guo et al.'s [15] and Blascheck et al.'s [4] choices of taxonomies is a good example for reasoning their decisions. Guo et al. evaluated a system that investigates networks of relationships between distinct types of entities, such as people, documents and keywords. Yi et al. [43] provide the adequate categorization of interactions for graph-based visualizations; therefore, Guo et al. used their taxonomy, added the "Retrieve", and removed the "Encode" actions in order to represent their interactions. On the other hand, Blascheck et al. examined multi-layer textual visualizations, VarifocalReader [24] and Word Cloud Explorer [17]. These interactions require abstract representations related to text retrieval operations [5] using taxonomies that are more general. In fact, Brehmer et al. [5] considered Yi et al.'s [43] taxonomy as low-level interactions when they compared it with other taxonomies, and used its definitions of "Select" and "Annotate" interactions.

Some previous efforts transcribed verbal records besides encoding the interactions [9, 29] while others only needed to count and characterize insights using the think-aloud collected [23, 40, 11]. Interestingly to approach the gap between the interactions and the sensemaking model, Reda et al. [37] have mixed an interaction coding schema with a mental state schema in a one state transition diagram. They were able to elaborate the transition between the observations, hypothesis, goal formation and the interactions. The cognitive side was revealed with respect to the interactions. In the perspective of the four-tier model proposed by Gotz and Zhou [13], the tasks and subtasks are equivalent to the mental states described by Reda et al. in [37].

Last but not least, some works have used raw interactions only without encoding them. For example, in TOME, Gomez and Laidlaw [10] used the Keystroke-Level Model (KLM) to collect interaction history, Heer in [16] used a hybrid of state and action model to group simple interactions under five categories: shelf, data, analysis, worksheet and formatting, and Brown et al. [6] used n-gram to extract

patterns from raw interactions without categorizing them according to any taxonomy (since their analysis was space-oriented).

5.2 Segmentation & chunking

Interaction representation does not stop at the encoding stage; individual interactions need to be grouped in order to form a semantic description which can be interpreted by evaluators or used as an input for predictive algorithms. The manual effort done by Gotz and Zhou [13] for grouping interactions in the form of trails enabled searching for the appropriate chunks of interactions that give meaning. They provided simple linear trails which provide meaningful chunks. The nonlinear trails provide the appropriate link between those chunks to find semantic meaning. Similarly, Guo et al. [15] have grouped the interactions into segments using a greedy algorithm after extracting all possible subsequences with length greater or equal to three as the candidate patterns. This method is considered fully automatic; the algorithm determines which interactions should be grouped into separate chunks or segments. In addition, Heer et al. [16] presented a history visualization of chunks of interactions that were extracted using temporal rules (time-dependent). These rules were designed manually from their observations on cases in their empirical data usage for Tableau.

To derive meaningful semantics from low-level interactions, Gotz and Zhou [13] have divided the interaction schema into four layered tiers. Their model's idea is universal in the sense of hierarchical linear division and conquering tasks. One contribution is highlighting the importance of the action layer that connects the semantic layers (task and subtasks) with the events. They were able to achieve this connection by combining the actions into groups to represent them as subtasks. They introduced the notion of trails to make such connection according to the classes of intentions which they defined. A similar example to Gotz and Zhou's trails is the patterns extracted in Guo et al.'s [15] experiment, while the difference is that the patterns were automatically extracted. In the sense of mapping low-level interactions to high-level semantics, Guo et al. have encoded the insights into three categories: fact, generalization, and hypothesis, and calculated the correlations between the interactions and those insight categories using Pearson's r. As a suggestion, we see an advantage in forming another extra tier between the action and subtasks tiers to enhance the action representation. The extra tier can be constructed with the help of automatic pattern detection; given that Guo et al.'s interaction patterns and Gotz and Zhou's trails are on the same tier level. Hopefully in the future, more efforts could build different catalog of trails to be reused across different applications.

5.3 Automatic analysis & prediction

Automated predictions and analysis are one of the core building blocks applied to help analysts derive inferences visually and/or statistically. In Brown et al.'s work [6], they fed their machine learning algorithms with three distinct data representations. As for the n-gram representation, they were able to draw conclusions on the users' performance by running a decision tree. The n-grams seem to work well with navigation related interactions as input to decision trees for deriving conclusions. Their other two representations (state- and event-based) were used as the input to the popular support vector machines (SVM) algorithm using different kernels. The accuracy of SVM on time completion was plotted along the experiment duration, in order to evaluate the method's performance in real-time. They also suggested using Boosting, sequence alignment methods, such as LCS in [4], and random process models which is used by Reda et al. [37].

In contrast to using n-gram as an intermediate feature, Blascheck et al. [4] extracted n-grams from the interaction logs to detect patterns as a final interpretation for the evaluators to investigate. The n-grams were extracted to infer similarities between users who have the same interaction patterns. As another bottom-up analysis, they used the Longest Common Substrings (LCS) to show those similarities and find longest time duration in which users were behaving similarly. The problem with LCS, is that it employs a very strict criteria for comparing such diverse outcome from users, thus, All Common Substrings were provided as another option. Privitera and Stark, [36] developed a methodology for automatically identifying AOIs (or regions of interests (ROIs)) using the sequence of eye fixations on images and compared those sequences using Levenshtein distance. Similarly, Blascheck et al. hierarchically clustered the eye movement data to identify similar scanpaths using Levenshtein distance.

Other than machine learning algorithms, Blascheck et al. provided basic search function for analysts to look up interaction subsequences of the exact match, and fuzzy search to relax the search restrictions using wildcards. Several works have used rule-based approaches to mine patterns and predict user performance, for example, HARVEST [12] detects patterns to construct a library of interaction rules based on their frequencies. Those rules were later used to recommend the adequate visual technique. In some cases, visualization designs force users to execute multiple interactions that are unnecessary and could be replaced by fewer ones. Gomez and Laidlaw [10] used CogTool [22] to predict the time reductions that could be achieved, where those predictions are also rule-based.

5.4 Visualization techniques

Taking advantage of the power of VA, researchers have designed VA systems to display and analyze interaction log data collected for evaluating other VA systems. This is how the name of the paper "VA squared" [4] is originated. The timeline visualization technique is the notion adopted by event visualization tools, such as LifeFlow [42] and EventFlow [30]. Blascheck et al. [4], Guo et al. [15], and Dou et al. [9] have adopted the timeline visualization of interaction sequences for showing the sequence of events. The most basic timeline visualization used within those three papers is in figure 2 (a), which is designed by Guo et al. [15]. Each participant is presented by a row and each top level action is color coded. A more complicated visual representation appears in Dou et al.'s [9]; their OAT shows the users' interactions with the three views of WireVis [8] and the investigation depths for one user (figure 2 (b)). Blascheck et al. [4] have displayed users within bounding boxes separately, and within each bounding box each of the tool's view correspond to a timeline, where the AOIs and interactions are plotted in parallel (figure 2 (c)). Another two important views in Blascheck et al.'s tool are the participant list that shows the dendrogram result of the hierarchical clus-

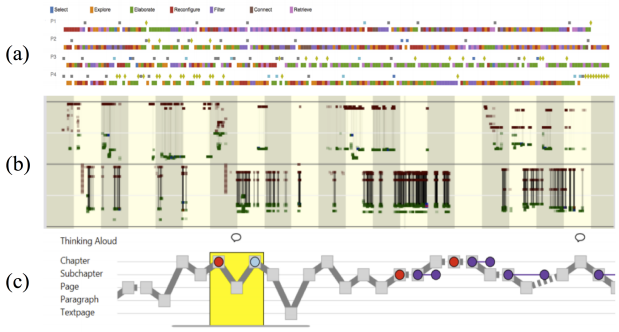

Figure 2: Timeline visualizations of interaction logs by: (a) Guo et al. [15], (b) Dou et al. [9] and (c) Blascheck et al. [4]

tering algorithm for grouping users, and the AOI list view that shows the transition between AOIs. In terms of the relationship between actions and between abstract actions with coded insights, Guo et al. [15] visualized the transition matrix per user and all users. One advantage of combining different data sources is linking between the meaningful quantitative and qualitative analysis. In their future work, Dou et al. [9] envisioned the usefulness of making video segments available in parallel to the interaction for evaluators to investigate. Blascheck et al. [4] applied this idea by providing speech bubbles for evaluators to view the video of interactions and voice recording as synchronized annotations. Such integrated information provides the evaluators a ground truth for verifying their hypothesis when analyzing interaction logs.

Brown et al. [6] showed the decision tree's result, where the leaf nodes are the output label classes (the user is fast or slow), the internal nodes are the n-grams extracted, and the edges represent how the decisions were taken according to the counts of the n-grams which are represented in the internal nodes. However, the resulting tree shown in their figure is more of a decision list rather a decision tree. They were also able to link between the user's traits and speed with their search behavior, which is depicted by the path taken inside the image to find Waldo. The transition between the viewpoints is visualized to differentiate typical behavior of users who are fast versus slow, and who impose external versus internal locus of control. The thickness of the line encodes the number of users who have gone through the transition.

6. CONCLUSION

In summary, we showed efforts that have used interaction logs for evaluating visualizations by summarizing their research goals from the perspective of three evaluation scenarios. In order to compare between the methods used in those efforts we constructed building blocks that summarize their common steps. Our building blocks integrate these common steps to demonstrate the automation of the evaluation process. The structure of the building blocks is a generalized form of the surveyed frameworks for future use; researchers can remove and add blocks according to their analysis goals and collected data.

Generally in user studies, the limited number of participants is a major bottleneck for evaluating explorative visualizations; they require deeper and more manual analysis than evaluating task-based visualization. Our intention is to encourage automating evaluation methods in order to clear this bottleneck. Many user studies' credibility depend on having large number of participants, thus, automating the evaluation process will provide the ability of hosting crowdsourcing experiments.

7. REFERENCES

[1] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel. A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages & Computing*, 22(3):213–232, 2011.

[2] G. Andrienko, N. Andrienko, M. Burch, and D. Weiskopf. Visual analytics methodology for eye movement studies. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2889–2898, 2012.

[3] E. Bertini, A. Perer, C. Plaisant, and G. Santucci. *BELIV'08: Beyond time and errors: novel evaluation methods for information visualization*. ACM, 2008.

[4] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl. Va 2: A visual analytics approach for evaluating visual analytics applications. *IEEE transactions on visualization and computer graphics*, 22(1):61–70, 2016.

[5] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[6] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics*, 20(12):1663–1672, 2014.

[7] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN systems*, 27(6):1065–1073, 1995.

[8] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 155–162. IEEE, 2007.

[9] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering reasoning process from user interactions. *IEEE Computer Graphics & Applications*, 29(3):52–61, 2009.

[10] S. Gomez and D. Laidlaw. Modeling task performance for a crowd of users from interaction histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2465–2468. ACM, 2012.

[11] S. R. Gomez, H. Guo, C. Ziemkiewicz, and D. H. Laidlaw. An insight-and task-based methodology for evaluating spatiotemporal visual analytics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 63–72. IEEE, 2014.

[12] D. Gotz, J. Lu, P. Kissa, N. Cao, W. H. Qian, S. X. Liu, and M. X. Zhou. Harvest: an intelligent visual

analytic tool for the masses. In *Proceedings of the first international workshop on Intelligent visual interfaces for text analysis*, pages 1–4. ACM, 2010.

[13] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009.

[14] T. M. Green, W. Ribarsky, and B. Fisher. Visual analytics for complex concepts using a human cognition model. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pages 91–98. IEEE, 2008.

[15] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics*, 22(1):51–60, 2016.

[16] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE transactions on visualization and computer graphics*, 14(6):1189–1196, 2008.

[17] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii International Conference on System Sciences*, pages 1833–1842. IEEE, 2014.

[18] R. N. Hunter. Successes and failures of patrons searching the online catalog at a large academic library: a transaction log analysis. *RQ*, pages 395–402, 1991.

[19] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR)*, 33(4):470–516, 2001.

[20] R. Jacob and K. S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, 2(3):4, 2003.

[21] M. W. Jaspers, T. Steen, C. van Den Bos, and M. Geenen. The think aloud method: a guide to user interface design. *International journal of medical informatics*, 73(11):781–795, 2004.

[22] B. E. John, K. Prevas, D. D. Salvucci, and K. Koedinger. Predictive human performance modeling made easy. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 455–462. ACM, 2004.

[23] Y.-a. Kang, C. Gorg, and J. Stasko. Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 139–146. IEEE, 2009.

[24] S. Koch, M. John, M. Wörner, A. Müller, and T. Ertl. VarifocalreaderâĂŤin-depth visual analysis of large text documents. *IEEE transactions on visualization and computer graphics*, 20(12):1723–1732, 2014.

[25] K. Kurzhals, B. Fisher, M. Burch, and D. Weiskopf. Evaluating visual analytics with eye tracking. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pages 61–69. ACM, 2014.

[26] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[27] H. Lam, P. Isenberg, T. Isenberg, and M. Sedlmair. Proceedings of the fifth workshop on âĂIJbeyond time and errorsâĂŤnovel evaluation methods for visualizationâĂİ(beliv 2014, november 10, paris, france). In *Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV 2014)*, 2014.

[28] P. Langley. User modeling in adaptive interface. In *UM99 User Modeling*, pages 357–370. Springer, 1999.

[29] H. R. Lipford, F. Stukes, W. Dou, M. E. Hawkins, and R. Chang. Helping users recall their reasoning process. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 187–194. IEEE, 2010.

[30] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics*, 19(12):2227–2236, 2013.

[31] D. Nicholas, P. Huntington, N. Lievesley, and R. Withey. Cracking the code: Web log analysis. *Online and CD-Rom review*, 23(5):263–269, 1999.

[32] C. North. Toward measuring visualization insight. *IEEE computer graphics and applications*, 26(3):6–9, 2006.

[33] C. North, P. Saraiya, and K. Duca. A comparison of benchmark task and insight evaluation methods for information visualization. *Information Visualization*, page 1473871611415989, 2011.

[34] C. Papadopoulos, I. Gutenko, and A. Kaufman. Veevvie: Visual explorer for empirical visualization, vr and interaction experiments. *IEEE transactions on visualization and computer graphics*, 22(1):111–120, 2016.

[35] A. Poole and L. J. Ball. Eye tracking in hci and usability research. *Encyclopedia of human computer interaction*, 1:211–219, 2006.

[36] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence*, 22(9):970–982, 2000.

[37] K. Reda, A. E. Johnson, J. Leigh, and M. E. Papka. Evaluating user behavior and strategy during visual exploration. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pages 41–45. ACM, 2014.

[38] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613, 2014.

[39] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

[40] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics*, 11(4):443–456, 2005.

[41] M. Smuc. Just the other side of the coin? from error to insight analysis. *Information Visualization*, page

1473871615598641, 2015.

[42] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. Lifeflow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1747–1756. ACM, 2011.

[43] J. S. Yi, Y. ah Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007.

[44] C. Ziemkiewicz, A. Ottley, R. J. Crouser, K. Chauncey, S. L. Su, and R. Chang. Understanding visualization by understanding individual users. *IEEE computer graphics and applications*, 32(6):88–94, 2012.