How to Evaluate Exploratory User Interfaces?

Tatiana Gossen, Stefan Haun, Andreas Nürnberger
Data & Knowledge Engineering Group, Faculty of Computer Science,
Otto-von-Guericke-University Magdeburg, Germany
{tatiana.gossen,stefan.haun,andreas.nuernberger}@ovgu.de

ABSTRACT

Usability evaluation is an integral part of user interface software development. We discus how to apply existing evaluation methods to exploration tools supporting complex information needs. Evaluation of such complex systems is very challenging and requires collaboration with domain experts for creating scenarios and participation. Furthermore, complex information needs are usually vaguely defined and require much user time to be solved. In order to evaluate these tools more efficiently four components are essential: a standardized evaluation methodology, benchmark data sets, benchmark tasks and clearly defined evaluation measures. As an outlook of this position paper, we propose a method which can serve as a starting point to develop a methodology for evaluation of exploration tools supporting complex information needs.

Keywords

usability evaluation, benchmark scenarios, exploratory search

1. INTRODUCTION

In this paper we discuss issues related to evaluation of exploration tools supporting complex information needs (*CIN-ET*). Our starting point are systems designed for exploration of large, high-dimensional and heterogeneous data sets. The *Jigsaw* [4] system for investigative analysis across collections of text documents, the *Enronic* [7] tool for a graph based information exploration in emails and the *CET* [5] for efficient exploration and analysis of complex graph structures are some examples of exploration tools. The research question which we targeted is how to evaluate such systems.

The most important functionality of exploration tools supporting complex information needs is to support users in the creative discovery of information and relations that were overlooked before in data sets (e.g. document collections). With an evaluation it should be proven that—using the tool—users are able to satisfy their complex information needs effectively, efficiently and with positive attitude.

Evaluation methods which can be used vary and consist of formal usability studies in the form of controlled experiments and longitudinal studies, benchmark evaluation of the underlying algorithms, informal usability testing and large-scale log-based usability testing [6]. There is also some research in the area of automatic evaluation of user interfaces [12]. We consider an automatic approach, but it is not clear if this would work for CIN-ET evaluation.

2. EVALUATION CHALLENGES

Since CIN-ETs are complex systems [10], evaluation of them is very challenging. The first challenge is to create an appropriate scenario for evaluation. The tasks must be complex enough to represent a realistic situation. Such realistic exploratory tasks require much time (weeks or even months) to be solved. Lab experiments are limited in time, therefore a "good balance" between time and the right level of complexity is crucial for lab user studies. Longitudinal studies overcome lab experiments drawbacks like strong time limitation and artificial environment. Researchers motivate the community to conduct long-term user studies because they can be well applied for studying the creative activities that users of information visualization systems engage in. [11]

CIN-ETs are often designed to be used by experts with domain-specific knowledge, e.g. molecular biologists, who are more difficult to find than participants without special skills or knowledge. Thus, the second challenge is recruiting the participants. This should be a group of people which represents the end users. It requires either collaboration with scientific institutions or some incentive (like money) to engage their participation [10]. In the study preparation step collaboration with domain experts is also needed to help the researchers in creation of appropriate scenarios.

Controlled lab studies and longitudinal studies require an involvement of target users. The well established usability aspects which are evaluated in these studies, are *effectiveness*, *efficiency* and *satisfaction* [1, 6]. In the context of CIN-ET evaluation, one can express effectiveness in the amount of discovered information, efficiency in time to find new facts or in importance of the made discovery and satisfaction in the user's rating of the tool's comfort and acceptability [3].

3. METHODOLOGICAL SHORTCOMINGS

By evaluating CIN-ETs we can either focus on the tool examination or carry out a comparative evaluation. Most researchers concentrate on evaluating their own tool to gain a deeper understanding of user interactions with it. However, the results do not provide such important information if or under what conditions their tool outperforms alternative tools for the same purpose. We found only one publication [8] that proposed an experimental design and a methodology for a comparative user study of complex systems.

To be able to compare and rank a CIN-ET among similar ones, benchmark data sets and tasks for user studies

SIGIR Workshop on "entertain me": Supporting Complex Search Tasks, July 28, 2011, Beijing.

are essential [9]. Suppose we wanted to repeat the study conducted in [8] to compare our tool to theirs, we would need the document collection and the task solution used by the authors. However, this data is not available to the public, so we cannot compare the results. A promising direction here is the *Visual Analytics Science and Technology* (VAST) contest[1] which offers data sets of different application domains with description and open-ended domain specific tasks. These tasks should be solved with the help of specific software within the contest. After the contest the solutions are made public, making the data available to evaluations.

Additionally, clearly defined evaluation measures are also important in order to evaluate exploration tools more efficiently. These could be measures from different domains, e.g. information retrieval and human computer interaction, but new measures are still necessary in order to capture the amount of discoveries in document collections or how creative a solution is. The task solution itself can be very complex, so we need a way to account for answers which are only partially correct or complete.

One can draw an analogy between user evaluation of exploration tools and IR automated evaluation of ranking algorithms. The latter requires a set of test queries, a document collection with labels according to relevancies (e.g. TREC) and a measure (e.g. Average Precision) [6], while CIN-ET user evaluation requires a benchmark data set, a benchmark task with a standard solution and an evaluation measure.

4. BENCHMARK EVALUATION

In the following we propose an evaluation method for discovery tools, consisting of two parts: The first part is a "small" controlled experiment with about 5–10 participants. The purpose of this is to collect qualitative data using user observations like audio/video recording and interviewing the participants afterwards. We actually do not need a special task to be solved by the participants. The assignment can be to discover new information using the software. From this study we collect data about learnability improvements and user satisfaction.

The second part is an online study, in which the software is provided to the participants as an online application. The participants can access the tool from their own working environment and spend as much time as they like with the tool, even working discontinuously. After that they can use an online questionnaire to provide the task solution and usability feedback. Participants are motivated to solve a thrilling task using the tool. We assume that the VAST benchmark data with an investigative task (from IEEE VAST 2006 Contest) can be used as a benchmark data set and a benchmark task. The tool interactions of each participant are logged on the server side. We can analyze them to get the time spent by participants to get the solution and interaction patterns. The outcome of the study also contains the number of participants who succeeded in solving the task in comparison to all participants who tried.

The described method is only the first step in the creation of a good methodology. It still has several drawbacks. The first problem is to get an appropriate participants' number. It is not easy to stimulate the participation even with money and if it would work the study becomes cost consuming. One possible solution lies in automatic evaluation (see, e.g., [2]). We could simulate exploration process on different levels and for diverse tasks. However it is not clear how to model a *creative* exploration process, which is important in the case of CIN tasks like creative information discovery. We also do not have a clear understanding how to judge the success of the search given a complex information need. Thus, the question about evaluation measure remains.

5. CONCLUSIONS AND FUTURE WORK

We proposed a method which can serve as a starting point to develop a methodology for CIN-ET evaluation. However, several aspects are yet unclear. This applies to evaluation methodology, in particular the possibility to evaluate the CIN-ET automatically, and evaluation measures. We would like to motivate the community and make the researchers pay attention to the fact that evaluation of CIN-ETs should be carried out using a standardized evaluation methodology in combination with benchmark data sets, tasks and measures. Only then CIN-ET designers can evaluate their tools more efficiently.

6. REFERENCES

[1] *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs). Part 11 - guidelines for specifying and measuring usability.* 1998.

[2] L. Azzopardi, K. Järvelin, J. Kamps, and M. Smucker. Proc. of SIGIR 2010 Workshop on the Simulation of Interaction: Automated Evaluation of Interactive IR (SimInt 2010). ACM Press, 2010.

[3] N. Bevan. Measuring usability as quality of use. *Software Quality Journal*, 4(2):115–130, 1995.

[4] C. Görg and J. Stasko. Jigsaw: investigative analysis on text document collections through visualization. In *DESI II Works.*, 2008.

[5] S. Haun, A. Nürnberger, T. Kötter, K. Thiel, and M. Berthold. CET: a tool for creative exploration of graphs. In *Proc. ECML/PKDD*, pages 587–590, 2010.

[6] M. Hearst. *Search user interfaces.* Cambridge University Press, 2009.

[7] J. Heer. Exploring Enron: Visualizing ANLP results. 2004.

[8] Y. Kang, C. Goerg, and J. Stasko. How can visual analytics assist investigative analysis? Design implications from an evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 2010.

[9] C. Plaisant. The challenge of information visualization evaluation. In *Proc. of the working conference on Advanced visual interfaces*, pages 109–116. ACM, 2004.

[10] J. Redish. Expanding usability testing to evaluate complex systems. *Journal of Usability Studies*, 2(3):102–111, 2007.

[11] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proce. of AVI works. on BEyond time and errors: novel evaluation methods for inf. vis.*, pages 1–7. ACM, 2006.

[12] S. Stober and A. Nürnberger. Automatic evaluation of user adaptive interfaces for information organization and exploration. In *SIGIR Works. on SimInt'10*, pages 33–34, Jul 2010.

[1] http://hcil.cs.umd.edu/localphp/hcil/vast11/