User Evaluation Methods for Visual Web Search Interfaces

Orland Hoeber
Department of Computer Science
Memorial University
St. John's, NL, Canada A1B 3X5
hoeber@cs.mun.ca

Abstract

In recent years, numerous visual Web search interfaces have been developed in the research community. However, the user evaluations of these interfaces have been performed using a wide range of methods, making it difficult to compare and verify the relative value of the proposed advancements. In this paper, we survey these evaluation methods, and propose a stepped evaluation and refinement model for the systematic study and enhancement of visual Web search interfaces. We suggest that this stepped model can be generalized to support the evaluation of other information visualization systems that target exploratory or knowledge-centric domains.

1 Introduction

Novel Web search interfaces that employ information visualization and interaction techniques are gaining popularity in the research literature [2, 6, 9, 11, 12, 14, 17]. However, it seems that few of these systems are being evaluated with human subjects in a systematic manner. The lack of empirical evidence regarding the utility of such interfaces makes it difficult for both academic and corporate researchers to evaluate the potential value that these interface improvements would provide to their systems and products [16, 23]. As a result, few advancements have been adopted within the public Web search market.

User evaluations provide researchers with a means to verify and validate design assumptions, confirm or reject hypotheses, and make comparisons between different systems and techniques. Within the domain of human-computer interaction, there are numerous accepted methods and procedures that can assist researchers as they evaluate their work. Many of these techniques are covered in textbooks in the domain of human computer interaction [18, 24, 27, 28]. Unfortunately, applying user evaluation methods in the study of Web search interfaces is not as well understood as in other domains.

The difficulty with evaluating Web search interfaces is related to the knowledge-centric nature of the domain. Finding the single top relevant document in the search results set is not of particular interest; the list-based representation of the existing search engines support this task adequately without the need for visual representations. Instead, novel Web search interfaces commonly focus on exploratory search tasks, or tasks where the user's information need is ill-defined. In these situations, the goal is to support the searcher as they consider and evaluate documents until they have gained *enough* knowledge about the target task. Measuring this knowledge and determining when a task is complete is a difficult aspect of any user study that focuses on such interfaces.

This work builds upon the research by others on the topic of evaluating information visualization systems [1, 4, 7, 16, 23, 25, 26, 29, 34]. We present a cross-section of evaluation methods available for researchers to validate their designs for visual Web search interfaces. Our primary contribution is the focus on the specific difficulties of evaluating visual Web search interfaces, and the presentation of a stepped evaluation and refinement model that includes multiple methodologies and incremental improvement of the interface under investigation. We propose that this model can be useful in the evaluation of not only information visualization systems targeted at Web search, but also for other exploratory domains such as geo-visualization, visual analytics, and exploratory data analysis.

2 Inspection Methods

Many researchers have proposed the use of inspection methods as part of their evaluation methodology for information visualization systems [23, 25, 34, 35, 38]. In general, there are two styles of inspection method: those that are based on a set of design guidelines, such as *heuristic evaluations*; and those that are based on performing specific tasks, such as *cognitive walkthroughs*. Both are normally conducted by one or more expert evaluators or researchers, either independently or in teams.

Heuristic evaluations provide a set of guidelines against which the prototype is evaluated [19]. In general, these are presented as a set of usability principles which apply to all user interfaces. The reviewer verifies that the system promotes heuristics such as visibility, control, consistency, error prevention, and recognition.

Cognitive walkthroughs differ slightly, in that they are inspection methods that are focused on the tasks that users will perform with the interface [20]. The goal is for the reviewer to act like an end-user conducting specific tasks with the system. As each step in a task is performed, the reviewer considers what the users would know and do, and whether they would have difficulties in performing the steps required to fulfill their tasks.

Both Freitas et al. [8] and Zuk et al. [38] have suggested that inspection methods can be extended to include the consideration of specific issues relevant to information visualization. Their extensions include evaluating the codification of information, the completeness of the information being visually represented, and the spatial organization of the objects in the visual display. Some of these extensions may be more relevant to heuristic evaluations; others are dependent on the task and therefore are more appropriately considered as part of a cognitive walkthrough. With respect to the codification of information, we believe that the primary concerns should be related to the methods for encoding information [15], and more specifically, the correct use of colour [32].

For the inspection of Web search interfaces, both methods are valuable. The heuristic evaluations provide a framework for evaluating the design choices made in the development of the interface. This can be especially valuable when the requirements for visualizing the Web search information are not well defined [35], or during exploratory design. The cognitive walkthrough methods can be very beneficial in understanding whether the visual representations of the information are helpful for the searcher's tasks, or just a distraction. Further, since such interfaces are often very interactive, it is important to not just focus on the core usability issues as promoted by heuristic evaluation methods, but also to evaluate the interactive nature of the systems.

These methods are low-cost and relatively low-effort, and can be conducted on preliminary prototype implementations allowing the researcher to identify and fix problems early in the development process. However, it may be difficult to choose an appropriately complex and comprehensive set of tasks for the cognitive walkthrough. In addition, there is a lack of empirical evidence produced; the end result is not a validation of the design, but instead an analytical evaluation and identification of potential problems that need to be addressed.

3 Laboratory Studies

Laboratory studies are the most common method for evaluating information visualization systems. Such studies can range from those that evaluate low-level perceptual aspects of a particular visualization technique or component using simple detail-oriented tasks, to the evaluation of complete information visualization systems using complex high-level tasks [4]. Quantitative measures such as time to task completion, error rates, and accuracy are commonly measured. In addition, qualitative measures relating to subjective satisfaction and opinions regarding specific features are also captured.

The design of such studies are guided by the *scientific method*. Hypotheses are developed, independent variables are identified and controlled, dependent variables are measured, and statistics are applied to verify or refute the hypotheses. Usability and information visualization goals drive the research questions and hypotheses, and subsequently affect the choice of independent and dependent variables.

The use of laboratory studies is promoted in most papers on the topic of evaluating information visualization systems [1, 4, 7, 21, 23, 25, 26, 34]. In general, problems with such studies include the fact that they are time consuming, expensive, and difficult to design [7, 34], and that participants are observed for a short period of time [23]. Such studies often make use of students as participants, even though they may not accurately reflect the target user group [1, 25]. A further critique of laboratory studies is that they focus on the abilities of novice users to quickly learn and use the system under investigation [34]. Since the design of laboratory studies is outside of the scope of this paper, we refer the reader to Andrews [1], Carpendale [4], and Ellis & Dix [7] for practical advice regarding the evaluation of information visualization systems in a laboratory setting.

Studies of visual Web search interfaces tend to focus on evaluating nearly-complete research prototype implementations. While the critiques of the participants in a laboratory study may be a concern in the general case, it is not as problematic when studying visual Web search interfaces. Since such interfaces are generally designed for public use, the ability for people to quickly learn and use the interface effectively is a paramount concern that is worthy of evaluation. Pre-task questionnaires can allow researchers to judge the prior Web search experience and analyze the data accordingly.

There are two further issues that are often difficult to resolve when evaluating visual Web search interfaces: choosing appropriate tasks, and deciding when the tasks have been successfully completed. Choosing trivial or easy tasks can result in the participant avoiding using the visual features that are the focus of the evaluation. Choosing difficult tasks without providing the participant with the prior knowledge required to judge the relevance of the documents can result in participants having an inability to complete the task. Designing a task that will trigger the active engagement needed to evaluate and explore the information is challenging [7, 21, 23].

Our experience has been to provide the participants with a scenario for the search activity, along with a clearly articulated information need. This is similar to the simulated work tasks proposed by Borlund [3]. Choosing tasks from collections such as the TREC 2005 Hard Track [1] alleviates the burden of having to create difficult yet understandable tasks; however it is important to ensure that the starting query is representative of the short queries commonly used in Web search [2]. Another alternative is to have the participants use self-identified search topics [22, 30, 33]. However, making comparisons between the performance of participants then becomes difficult.

Task completion can be especially difficult to determine in this domain. Since Web search is a knowledge-centric activity, how does one decide when a participant has gained enough knowledge about a topic? While it may be possible to quiz the participants at the end of the session, separating the ability of the participant to learn about the topic from the support the visual Web search interface provides may not be possible. Somehow, we must be able to estimate the amount of knowledge the participants are gaining as they perform their assigned search activities.

A method we have found to be effective is based on participants assigning relevance scores to the documents they consider [30, 33], without viewing the actual documents. Measurements of task completion times are based on the participants having identified a pre-determined number of relevant documents (e.g., 10 relevant documents). By comparing the relevance scores of the participants to *ground truth* relevance scores assigned by a panel of experts, measurements such as error rates and accuracy can be calculated. Others have suggested allowing participants to continue searching until they are satisfied [2, 30, 33] or to limit the number of documents that can be considered [22, 36]; however these techniques make it difficult to measure the time to task completion.

Although there are many challenging aspects to conducting laboratory studies on visual Web search interfaces, the one key benefit is that they produce empirical results that are replicable. Careful selection of tasks and participants can provide evidence regarding the utility of the system for searchers of various capabilities and tasks of varying complexity. Further, laboratory studies are well-suited to conducting a structured comparison between multiple visualization alternatives (although care should be taken in situations where the participants are familiar with one interface and not the other [1]). An added benefit is that by simply watching the participants as they work through the assigned tasks, we can gain insight into how the visual Web search interface is working and how readily the participants accept and use the visual representations and tools. Others have noticed a similar benefit in the evaluation of their own information visualization systems [23].

4 Field Trials

Field trials are an evaluation method that studies how a small group of participants are able to use the system under investigation in their normal work environment doing real tasks. As noted by Shneiderman and Plaisant, “the physical and social environments are inextricably intertwined with the use of information and computing technologies” [28]. The goal here is to verify that the system can operate as expected in the *real world*.

The normal procedure is to use ethnographic methods and think-aloud protocols to observe and evaluate how the system is being used, how effective the participants are at formulating mental models, and whether the participants are experiencing any difficulties. Since it is difficult to make direct comparisons between participants and tasks, and the small number of participants makes verifying the statistical significance of the results infeasible, quantitative measures are not as valuable as in laboratory studies. Instead, qualitative measures carry much more weight, with satisfaction being an important indicatory of success [23].

Some have suggested that focus groups used in collaboration with field trials can provide a rich source of qualitative information regarding the information visualization system [16]. It is also possible to extend the field trials over longer periods of time in order to gain an understanding of how the participants transition from novice to expert users [29].

When evaluating visual Web search interfaces, field trials can be very useful when there is a sub-group of end-users that require special consideration. For example, an interface may be designed for the specific needs of expert researchers conducting exploratory searches. Others might be designed for novice users, or for those with vision impairments. In all of these cases, field trials with a small group of participants can provide valuable insights into the learnability, usability, and utility of the interface.

The key benefit of a field trial is that the results are based on realistic usage settings and tasks. In essence, the search tasks are no longer assigned to the participants; instead they are able to search for anything they like. The goal is to keep the data collection methods open-ended, focusing on the insight gained by the participants [21]. Within such studies, the subjective satisfaction measure is an important indicator of how well the proposed interface may be received by

[1] http://trec.nist.gov/data/t14_hard.html

the target market. However, it can be difficult to replicate or generalize the results, since they are tied directly to a specific setting. When analyzing the results, the differences between the participants and their tasks must be considered carefully.

Prior to running a field trial, it is important that the visual Web search interface be sufficiently complete and bug-free. Any problems with the stability of the system can result in severely decreased subjective reactions, since it is often difficult for participants to separate the value of the visual representation from the problems that occurred due to bugs. As such, early research prototypes may not be suitable for field trials.

5 Longitudinal Studies

Longitudinal studies are designed to allow participants to engage in learning and using the system under investigation over an extended period of time. Such studies may span multiple days or weeks. During this time, the participants make exclusive use of the system for all activities related to the target tasks that are supported.

Participants in such studies are commonly recruited from the target user population. Since there is no need for the investigator to actively observe or monitor the participants (unlike field trials), it is possible to support a large number of participants in a single study. However, the coordination of the participants and ensuring that they remain actively involved in the study requires careful planning and monitoring.

The data in longitudinal studies can be collected by logging the usage activities of the participants, administering questionnaires, conducting interviews, and/or conducting focus groups. The usage logs, which will be discussed in more detail in the next section, allows the data to be collected remotely and automatically. Daily, biweekly, or weekly questionnaires (perhaps administered online) allow the subjective reactions of the participants to be captured throughout the course of the study. Questionnaires may be based on existing instruments designed to capture impressions of learnability and ease-of-use, such as the *technology acceptance model* [5]. Interviews and focus groups can also be conducted throughout the study period to capture qualitative impressions on the utility of the system within the real-world setting.

Longitudinal studies are especially well-suited to the evaluation of visual Web search interfaces. Since searching the Web is the type of activity that occurs in the workplace, school, and home, and at various times throughout the day, field trials may not capture all aspects of use. Longitudinal studies provide valuable insight into the utility of the visual representations and interaction throughout the full breadth of Web search activities.

The key benefits to longitudinal studies is that they capture activities and impressions of the system during real-world use. There is a moderate level of control, in terms of measuring the activities and subjective reactions, and choosing the participants. The studies are not constrained by a specific location or time, making them more flexible than other evaluation methods. The evidence captured is empirical, and the study can be readily replicated with additional participants. Further, by carefully timing the frequency of the questionnaires, interviews, or focus groups, the ease by which the participants are able to learn to use a system can be captured.

However, compared to the other evaluation methods, longitudinal studies are difficult to manage and time consuming. Care must be taken in the analysis of the results, given that the tasks performed may be different among the various participants. There is also the risk of attrition that must be managed (i.e., participants dropping out of the study prior to its completion); our experiences have found that rewarding the participants both at the beginning and end of the study can be effective.

6 Instrumentation and Log Analysis

The instrumentation of an interface consists of adding additional features that log the activities users perform with the system. In some cases, these logs may be stored on the local machine; in other cases, they may be sent over the Internet and stored on a server. By conducting a log analysis, we are able to determine the extent to which the features of the interface are being used. With enough participants in a study, aggregate trends and patterns can emerge.

Although we treat instrumentation and log analysis as a separate category of user evaluation, the use of instrumentation as a means for capturing the breadth of interaction has been employed in laboratory experiments [25], as elements of field trials [23, 29], and during longitudinal studies [10].

Log analysis can be especially valuable in studies of people using Web search engines. There have been numerous studies on the abilities of people to craft queries and evaluate search results based solely on search engine logs [13, 31]. For visual Web search interfaces, it is vitally important to not just log the queries, but also to log how the interactive visual features of the system are being used [37].

An important concern that should not be ignored with using instrumentation and log analysis techniques is protecting the identity and actions of the participants in the study. Certainly, if participants feel that their every action is being monitored and recorded, and that this information is tied directly to them as individuals, they may not use the system in a normal manner.

The key value in recording user activity is in the richness of the data collected. It is possible to re-construct the activities the users undertook during a session for the purposes of analyzing exceptional cases; it is also possible to conduct statistical analysis on the use of specific features of the system. However, care must be taken to ensure that the information needed for the analysis has been properly instrumented; this can be a significant technical challenge in certain settings. In some cases, complex interactions can become very difficult untangle and analyze.

7 Stepped Evaluation and Refinement Model

Each of the methods discussed in this paper can provide valuable information regarding the efficiency, effectiveness, and utility of a visual Web search interface. We propose that rather than using them in an ad-hoc fashion, they be used as part of a comprehensive structured process to both improve the quality of the system and to provide empirical evidence regarding the utility of the proposed methods, both in controlled and real-world settings (see Figure 1).

The use of multiple methods in an evaluation process is not a new idea. Tory & Möller suggested a combination of laboratory studies and usability inspection [34, 35]. Rester & Pohl suggested usability inspections followed by laboratory studies, field trials, and finally an assessment of transferability [25]. Plaisant suggested that laboratory studies could be augmented by longitudinal studies and field trials [23]. North proposed the combination of laboratory studies and open-ended evaluations similar to field trials [21]. Our contribution here is a systematic method for the evaluation and subsequent refinement of knowledge-centric interfaces such as those for visual Web search. In particular, we explicitly include steps of refinement and re-development as a direct outcome of the evaluations, moving the interface from a preliminary research prototype to a fully-operational beta-release.

This stepped model begins with the design and development of a preliminary research prototype implementation of the system. Inspection methods that consist of both information visualization-specific heuristic evaluations and cognitive walkthroughs provide concrete and actionable information regarding which aspects of the interface are working as planned and which need more work. Since the goal is to refine the interface prior to conducting further testing, we can consider the inspections as formative evaluations. Depending on the extent to which changes are made, it might be necessary to conduct subsequent rounds of inspections, and make further refinements to the research prototype.

The second level of evaluation are the laboratory studies. These are conducted on the refined research prototype, with the expectation that all the major usability issues have been resolved, and the researchers are confident that their visual representations and interaction methods will be beneficial to the end-users. While it may be necessary in some cases to conduct component-level laboratory experiments, in most cases with visual Web search interfaces, evidence of the utility of the techniques employed will come from evaluating the full system. It may also be useful to design the study such that a comparison of visualization alternatives can be achieved.

The outcome of the laboratory experiments will include empirical evidence that validates some hypotheses, refutes others, or provides inconclusive or mixed results. The subjective opinions of the participants is also an important measure. As with most research, some of the things that are tried will work as expected; others will not. Rather than stopping here (as much of the research literature does), it is important to verify that the hypotheses hold in real-world settings. However, most research prototypes are not sufficiently complete to support real-world testing. Therefore, a necessary step is to re-develop or further refine the research prototypes as fully-operational beta-release systems.

There are two choices for evaluating the beta-release system in real-world settings: field trials and longitudinal studies. If there is a small sub-population of users for whom it is important to verify the system in a real-world setting, it can be beneficial to conduct field trials. The primary outcome is subjective evidence regarding the efficacy of the system and the satisfaction of the participants.

Longitudinal studies allow for the collection of empirical evidence using a large groups of participants over extended periods of time. Through the use of logging facilities, information regarding how the system is being used can be captured and analyzed. Repeated questionnaires during the course of the study can provide evidence of the learnability of the system. All of this information is captured in the context of real-world use, providing grounded insight into the value of the information visualization techniques.

The primary drawback of this model is the time and effort it would take to apply it in a research or industry project. Clearly, working through multiple iterations of inspections, laboratory studies, and then field trials and/or longitudinal studies is not trivial. However, the benefit of this model is that it allows researchers to determine the value of the proposed technique in an incremental manner, increasing the quality and completeness of the prototype implementation along the way. Flaws in the fundamental hypotheses and negative aspects of the implementation can be identified early and corrected as needed. The final outcome includes comprehensive empirical evidence of the value of the visualization method in both controlled and real-world settings.

We have used this stepped model of evaluation and refinement in our own research on visual and interactive interfaces to Web search. Prototype implementations of two techniques were developed, inspected, refined, and evaluated in laboratory studies: HotMap [12] and WordBars [11]. Based on the outcomes of these studies, the posi-

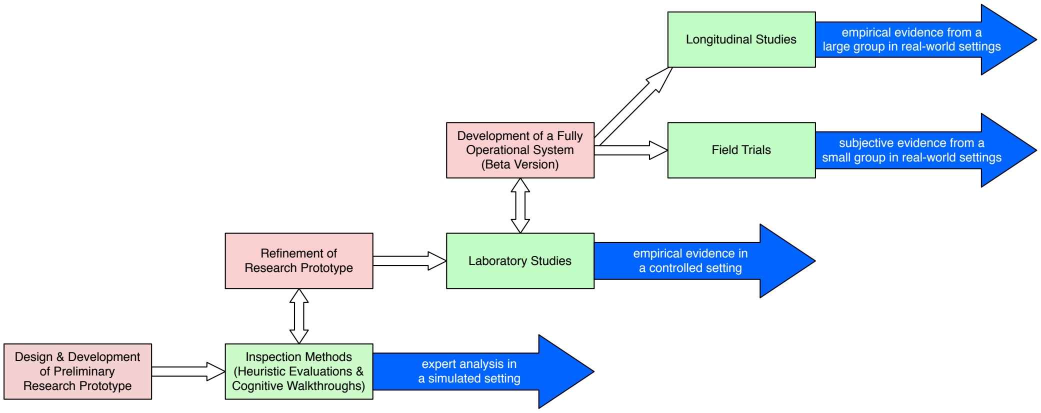
Longitudinal Studies
empirical evidence from a large group in real-world settings
Development of a Fully Operational System (Beta Version)
Field Trials
subjective evidence from a small group in real-world settings
Refinement of Research Prototype
Laboratory Studies
empirical evidence in a controlled setting
Design & Development of Preliminary Research Prototype
Inspection Methods (Heuristic Evaluations & Cognitive Walkthroughs)
expert analysis in a simulated setting

Figure 1. The stepped evaluation and refinement model, wherein the evaluation methods lead to further refinement and development of the information visualization system. The final outcome is empirical evidence from a large group of real users.

tive elements of each of these prototypes have been combined into a public beta version of these research efforts: *theHotMap.com*. This system has subsequently been evaluated in a longitudinal study, providing empirical evidence of the benefits of the techniques in real-world settings [10].

Although this model was formulated within the context of evaluating visual Web search interfaces, we believe it can be applied to the evaluation of other interfaces for which laboratory studies do not provide sufficient evidence regarding the value of the proposed techniques. Within knowledge-centric domains (including Web search), the formulation of representative tasks is difficult, as is the measurement of the knowledge gained by the participants. Following this model, researchers are able to evaluate and refine aspects of the interface in a structured manner, leading to a stable and fully-operational beta-version of the system that is suitable for real-world evaluations. While the inspection and laboratory study methods will provide some evidence regarding the value of the techniques used in the interface, the true value (and problems) will be revealed through real-world use during the field trials and longitudinal studies.

8 Conclusions

Although there have been substantial research advancements in the realm of visual Web search interfaces in recent years, the evaluation of these methods have generally been ad hoc and unstructured. We believe that this has made it difficult to examine the benefits that these advancements can provide to the Web search industry.

In this paper, we have provided a cross-section of research methods that can be used to evaluate visual Web search interfaces. We have combined these methods into a stepped evaluation and refinement model which provides a systematic process of study and improvement of knowledge-centric interfaces such as those present in Web search. We believe that by following this model, researchers will be able to incrementally increase the quality of their research prototypes, and produce evidence regarding the value and constraints of their proposed methods both within controlled settings and the real world. Our hope is that this evidence will lead to the adoption and integration of more research-initiated information visualization methods within the domain of Web search, as well as other knowledge-centric domains.

Acknowledgements

This work has been made possible through the author's Discovery Grant provided by the Natural Science and Engineering Research Council of Canada (NSERC).

References

[1] K. Andrews. Evaluating information visualizations. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2006.

[2] E. Berenci, C. Carpineto, V. Giannini, and S. Mizzaro. Effectiveness of keyword-based display and selection of retrieval results for interactive searches. *International Journal on Digital Libraries*, 3(3):249–260, 2000.

[3] P. Borlund. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[4] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, editors,

Information Visualization: Human-Centered Issues and Perspectives, LNCS 4950, pages 19–45. Springer, 2008.

[5] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Quarterly*, 13(3):319–340, 1989.

[6] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. VisGets: Coordinated visualizations of Web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205–1212, 2008.

[7] G. Ellis and A. Dix. An explorative analysis of user evaluation studies in information visualization. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2006.

[8] C. Freitas, P. Luzzardi, R. Cava, M. Winckler, M. Pimenta, and L. Nedel. On evaluating information visualization techniques. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, 2002.

[9] T. Heimonen and N. Jhaveri. Visualizing query occurrence in search result lists. In *Proceedings of the International Conference on Information Visualization*, pages 877–882, 2005.

[10] O. Hoeber, D. Schroeder, and M. Brooks. Real-world user evaluations of a visual and interactive Web search interface. In *Proceedings of the International Conference on Information Visualization*, 2009.

[11] O. Hoeber and X. D. Yang. Evaluating WordBars in exploratory Web search scenarios. *Information Processing and Management*, 44(2):485–510, 2008.

[12] O. Hoeber and X. D. Yang. HotMap: Supporting visual explorations of Web search results. *Journal of the American Society for Information Science and Technology*, 60(1):90–110, 2009.

[13] B. J. Jansen and U. Pooch. A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3):235–246, 2001.

[14] B. Kules, J. Kustanowitz, and B. Shneiderman. Categorizing Web search results into meaningful and stable categories using fast-feature techniques. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 201–219, 2006.

[15] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.

[16] R. Mazza and A. Berré. Focus group methodology for evaluating information visualization techniques and tools. In *Proceedings of the International Conference on Information Visualization*, pages 74–80, 2007.

[17] T. N. Nguyen and J. Zhang. A novel visualization model for web search. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):981–988, 2006.

[18] J. Nielsen. *Usability Engineering*. Academic Press, 1993.

[19] J. Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 152–158, 1994.

[20] J. Nielsen and R. L. Mack. *Usability Inspection Methods*. John Wiley & Sons, 1994.

[21] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006.

[22] S. Nowicki. Student vs. search engine: Undergraduates rank results for relevance. *Libraries and the Academy*, 3(3):503–515, 2003.

[23] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 109–116, 2004.

[24] J. Preece, Y. Rogers, and H. Sharp. *Interaction Design: beyond human-computer interaction*. Johyn Wiley & Sons, 2002.

[25] M. Rester and M. Pohl. Methods for the evaluation of an interactive infovis tool supporting exploratory reasoning processes. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2006.

[26] K. Risden, M. P. Czerwinski, T. Munzner, and D. B. Cook. An initial examination of ease of use for 2d and 3d information visualizations of Web content. *International Journal on Human-Computer Studies*, 53(5):695–714, 2000.

[27] M. B. Rosson and J. M. Carroll. *Usability Engineering: scenario-based development of human-computer interaction*. Morgan Kaufmann, 2002.

[28] B. Shneiderman and C. Plaisant. *Designing the User Interface*. Addison-Wesley, 4th edition, 2005.

[29] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2006.

[30] A. Spink. A user-centered approach to evaluating human interaction with web search engines: an exploratory study. *Information Processing and Management*, 38(3):401–426, 2002.

[31] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[32] M. C. Stone. Representing colors as three numbers. *IEEE Computer Graphics and Applications*, 25(4):78–85, 2005.

[33] L. T. Su. A comprehensive and systematic model of user evaluation of Web search enginges: II. an evaluation by undergraduates. *Journal of the American Society for Information Science and Technology*, 54(13):1193–1223, 2003.

[34] M. Tory and T. Möller. Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 10(1):72–84, 2004.

[35] M. Tory and T. Möller. Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications*, 25(5):8–11, 2005.

[36] L. Vaughan. New measures for search engine evaluation proposed and tested. *Information Processing and Management*, 40(4):677–691, 2004.

[37] O. Zamir and O. Etzioni. Grouper: a dynamic clustering interface to web search results. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 31(11-16):1361–1374, 1999.

[38] T. Zuk, L. Schlesier, P. Neumann, M. Hancock, and S. Carpendale. Heuristics for information visualization evaluation. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2006.