User-Oriented Evaluation Methods for Interactive Web Search Interfaces

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

Xue Dong Yang
Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
yang@cs.uregina.ca

Abstract

Although significant efforts have been devoted to the study and evaluation of information retrieval systems from an algorithmic perspective, far less work has been performed on the evaluation of these systems from the user's perspective. This is certainly the case for Web information retrieval, where the major search engines continue to utilise interfaces that have not changed substantially since their introduction. One of the challenges in developing new Web search interfaces is the evaluation of these systems in comparison to one another, as well as in comparison to the popular Web search engines. In this paper, we highlight some of the methods used in the literature for evaluating Web search systems, and present a summary of the methods that we have found to be effective in dealing with the challenges of evaluating intelligent and interactive Web search interfaces.

1. Introduction

In many user-oriented research fields, such as human computer interaction and information visualization, conducting user studies is not only an accepted practice, it is an expectation. 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 these domains, 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 [10, 14, 13, 12].

A common procedure for conducting user evaluations with software is to assign the participants various tasks to perform, measure the time it takes to complete the tasks along with the number of errors made, and make observations about how the software is being used. Subjective reactions by the participants can also be collected in order to measure their confidence and satisfaction in completing the task, and the ease of use of the software. The assigned tasks are often representative of the operations that users commonly perform in the real-world use of the software. Alternately, tasks may be assigned to explore specific new features of the system in order to understand whether these features are indeed an improvement over the existing practice.

Applying user evaluation methods in the study of Web search interfaces is not as well understood as in other domains. It has been noted that "the study of end-user searching on Web search engines is still in its infancy" [17]. In the literature describing user-oriented evaluations within this domain, the methods employed cover a wide range with respect to the experimental design, task assignment, and measurements. In this paper, we survey these methods and comment on the techniques we have found to be particularly useful in the evaluation of interactive Web search tools.

2. Experimental Design

Experimental design is a critical aspect of any user-oriented evaluation: poor experimental design leads to ambiguous and biased results that are hard to analyse; good experimental design leads to clear and unbiased results that are statistically verifiable. In addition to designing the experiment such that independent variables can be manipulated and dependent variables can be measured, a number of important experimental design decisions must be made.

When designing the experiments for the comparison of multiple Web search interfaces, care must be taken to minimize the ability for participants to transfer what they have learned in one interface to the operation of the other interfaces. While this can be ensured using a between-subjects design (i.e., each participant is only exposed to one interface), it is more common to conduct a within-subjects experiment (i.e., each participant is exposed to each interface) requiring fewer participants. The potential bias is addressed by randomizing the order in which the participants are exposed to the interfaces [17, 11].

Whenever a within-subjects design is chosen for the interface variable, it is possible for participants to learn or remember the results from a previously assigned Web search task. To address this situation, some studies provide multiple search tasks, and vary the order in which the participants are exposed to this independent variable through group assignment [1, 17]. Performing multiple Web search tasks increases the amount of time between conducting the same task with the different interfaces, and reduces the ability of the participants to remember the specific details of the search results.

In some studies, participants were allowed to follow the links in order to view the documents [15, 17]. In doing so, there is a risk that participants will be able to learn about the search task as they view these documents. While this is a beneficial effect in real-world Web searching, it has a biasing effect in user evaluations. Further, participants may be able to remember the visual layout of relevant documents in subsequent searches. This can be avoided by restricting participants from following the hyperlinks to view the documents, requiring that they considered only the information provided within the Web search interfaces being studied.

3. Web Search Tasks and Search Results

In the studies on Web search interfaces, some researchers provided specific search tasks and queries for their participants to conduct [1, 19], whereas others allowed participants to choose their own search topics [20, 15, 17, 11]. While allowing participants to choose their search topic provides a more realistic evaluation, it does not readily lend itself to making systematic comparisons of the Web search interfaces under controlled conditions.

Care must be taken when choosing a set of Web search tasks for participants to perform using the candidate Web search interfaces. While a number of test collections exist that include queries for the evaluation of algorithms used in information retrieval systems (e.g., TREC 2005 HARD Track[1]), the queries tend to be long, complex, and very specific. Since most Web searchers use queries consisting of one to three terms [9, 16], it is not realistic to use these test collection queries as-is. An alternative is to use the long queries as the descriptions of the information need, and write shorter queries for these topics [1]. Alternately, researchers may write their own Web search tasks to address specific features of their interfaces, such as the ability to handle question-answering searches, or sift to through very ambiguous sets of search results.

Regardless of how the search task is devised, a critical feature of a user-oriented evaluation of an intelligent Web search interface is that the users have an adequate understanding of the assigned information need. Clearly, if a participant does not understand what they are searching for, they will have a very difficult time deciding the relevance of the documents in the search results. It is beneficial to spend time at the beginning of each task providing explanations and answering questions. This can help alleviate situations where participants misinterpret the meaning of the search topic.

When evaluating multiple user interfaces for Web search results, it is very important to ensure that these interfaces all provide access to the same results set. If a live search engine is used, it is entirely possible that the search results may change during the course of the study. Therefore, it is desirable that the search results be cached prior to starting the study, and that these search results be provided to each competing interfaces in the study.

4. Measures

Making measurements of dependent variables as the independent variables are manipulated is one of the fundamental procedures in conducting user evaluations. Commonly, we wish to determine whether our changes have allowed the participants to perform their search tasks faster, more accurately, with less errors, with more satisfaction, and with higher confidence. The later two items from this list can be measured using questionnaires, the first three require relevance judgements to be made and for there to be clear completion criteria for each task. In this section, we will discuss the issues related to the measurement of the dependent variables when conducting a study to compare Web search interfaces.

4.1 Relevance Judgements

When evaluating a Web search interface, whether it is an independent evaluation of a single system or a comparison of multiple interfaces, researchers must be able to measure the participants' judgements in regards to the relevance of documents in the search results set. The methods by which this relevance judgement data is measured varies greatly within the literature.

Some have used Web search logs as the source of relevance judgement data [20]. However, this assumes that all documents viewed are equally relevant to the information need. Further, this technique does not provide any information with respect to how many non-relevant documents were considered in the course of the evaluation.

In user evaluations of traditional information retrieval systems, binary relevance judgments are common [3, 18]; these have also been used in the evaluation of Web information retrieval systems [1]. Binary relevance decisions as-

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

Table 1. A relevance scale for measuring the participants' relevance judgements.

Score	Description
4	This document is relevant.
3	This document is probably relevant.
2	This document might be relevant.
1	This document is not relevant.

sume a very simple model of document relevance (i.e., that a document is either relevant or not relevant to the information need). Clearly, in this model, there is no room for partial relevance, or for the degree of confidence in the relevance to be expressed by the participants.

Some studies have asked participants to provide the top search results in the order of relevance to their information need [11, 17, 19]. Although this technique can allow statistical comparisons to be made between the participants' orders and the order provided by the search engines (with the goal of determining the quality of the search results) [11], it assumes that the participants are able to provide such an order. This may not be possible if there are a large portion of non-relevant documents, or if the number of documents to be ranked is large. Further, the order of relevant documents may not be as important as finding a set of high-quality relevant documents tasks such as exploratory searching.

Another alternative is to allow the participants to make relevance judgements using a finite relevance scale. Both three-point [17] and four-point scales [15] have been reported in the literature. Scales such as the one in Table 1 are easy to understand and remember by the participants, and take into account different levels of confidence in the search results. While it would be possible to use a scale with a larger number of choices, keeping the number relatively small makes it easier for the participants to remember the meanings of the scores, and reduces the amount of time the participants spend deciding which relevance score to assign to a particular document.

Some have suggested the use of a continuous relevance scale, requiring participants to mark on this scale how well each document in the search results fulfills their information need [2, 15]. While this technique has also been used in the study of traditional information retrieval systems [8], it is not clear whether participants can effectively use such a tool to indicate relevance. Further, since a common analysis method is to quantize the locations indicated by the participants on this relevance scale [2], the result is equivalent to providing a multi-point relevance scale.

4.2 Time to Completion

In order answer research questions such as "Does our new Web search results interface allow users to evaluate search results faster than Google?", we must be able to measure how long it takes participants to complete the same assigned Web search tasks using the competing interfaces. In order to measure this time to completion, clear completion criteria must exist for the tasks. Without such criteria, time comparisons among users may not be valid.

In many studies, participants were allowed to continue searching until they were satisfied (or had given up) [1, 15, 17]. Other studies only provided a limited number of documents for the participants to consider [11, 19]. Of these studies, only one provided an analysis of the time taken to complete the tasks [17]. However, since there was no clear completion criteria, this timing data has little meaning. Some participants may have been very engaged in evaluating the search results, spending a large portion of time in a very successful and productive manner; others may have become frustrated and stopped after a short period of time.

In order to effectively answer research questions related to the time efficiency of an interface, tasks with equivalent completion criteria must be performed. For example, using the relevance scores in Table 1, it is possible to specify a task as being completed once the participant has given ten documents a relevance score of either three or four. While this may not be a natural method for users to decide whether they have successfully fulfilled their information need, it provides a consistent way to conclude the task, independent of the interface being used.

4.3 Subjective Measures

Collecting subjective measures is a valuable way of measuring participant feellings towards satisfaction, confidence, and ease of use. This data is often collected in the form of a questionnaire administered at the end of the evaluation session. Commonly, participants will be asked to rate their degrees of agreement on a Likert scale with respect to a number of statements regarding their tasks and the features of the software.

While collecting these subjective measures at the end of the session is often most convenient, there are other times in a user evaluation of Web search interfaces when subjective measures may also be taken. For example, some have administered evaluation questionnaires at the end of each search task [1, 17]. Such measures taken immediately following a task can be used to determine the participants' subjective opinions about a specific task conducted using a specific interface. Measures taken at the end of the sessions can be used to determine participants' subjective opinions about the general features of the tasks and interfaces.

4.4 Data Collection Methods

While it is very tempting to develop automatic data collection methods, it is not advisable to do so without considering the consequences of this decision. One of the largest pitfalls in data collection is to develop a method that interferes with the assigned tasks. This is especially problematic when the tasks are timed. For example, one study of a traditional information retrieval system included relevance rank sliders within the interface [8]. Clearly, these sliders were in direct competition with the interface being tested.

Others have required the participants to print out the search results pages for further evaluation [17, 11]. While this may be a valid method for evaluating existing search engines, new prototype systems may not include the ability to print, or may introduce interactive features that promote the manipulation and exploration of the search results [6, 5].

In other studies, participants completed paper forms with their relevance score data for each document considered [15, 19]. While this is a low-overhead method that does not directly compete with the Web search interface, it does compete with the Web search tasks being performed by the participants. For example, in order to log the relevance score for a particular document, the participant must change their focus from the interface to the paper form, and then back again to continue the task.

Manual data entry methods that require the participant to speak the relevance scores and the investigator to log these reduces the impact of collecting the data on the Web search task being performed. This verbal protocol is least intrusive than the methods discussed above. As an added bonus, this method of data collection allows the researcher to carefully watch the users as they perform their Web search tasks, providing valuable insight into the Web search techniques and strategies employed by the participants. The automatic logging of user interactions within the interface may also be beneficial, so long as it doesn't interfere with the participants' primary tasks of conducting searches with the interface.

5. Experiences with HotMap and WordBars

Evaluating intelligent interfaces for Web search poses special problems in the design of the experiments, the tasks assigned to the participants, and in the collection of the measures. In particular, tools such as HotMap [6] and WordBars [5] support users in interactively manipulating the search results as they seek relevant documents. These tools are designed to support exploratory searching, providing a range of actions that are much more complex than the common list-based representation of search results. As a result, many of the simpler techniques that have been used to evaluate static Web search interfaces are not as effective for the evaluation of intelligent and interactive Web search interfaces. In this section, the decisions we have made in designing and conducting user evaluations of the prototype systems [4, 7] are provided.

Experimental Design

1. A within-subjects design requires fewer participants and allows for the direct comparison between interfaces on a participant-by-participant basis.

2. Assigning the order participants use the interfaces in a pseudo-random order reduces the biasing effects of using one interface before another.

3. Performing multiple search tasks in a pseudo-random order reduces the learning effects of having previously conducted a search on the assigned tasks.

4. Requiring participants to decide relevance based on the information provided in the interface reduces their ability to learn about the topic by viewing specific documents.

Web Search Tasks and Search Results

1. Providing specific search tasks that all participants conduct supports the evaluation of the specific type of searching supported by the interface (i.e., using exploratory tasks to evaluate exploratory Web search interfaces).

2. Writing queries based on the search topics from existing test collections reduces the work required to prepare appropriate search tasks

3. Describing each task in detail and answering questions posed by participants places each participant at an approximately equal level with respect to their prior knowledge on each search topic.

4. Caching the search results for each task ensures that each participant reviews an identical set of documents during the course of the evaluation.

Measures

1. Using a four-point relevance judgement scale captures varying degrees of confidence in the relevance of documents considered.

2. Providing a clear completion criteria for each task supports the comparison of the interfaces based on time measurements.

3. Collecting subjective reaction measures after the completion of each search task with each interface provides specific details regarding the participants' feelings at each stage in the study.

4. Collecting subjective reaction measures at the conclusion of the study provides an overall view of the participants' feelings with respect to the interfaces in general.

Data Collection Methods

1. Having the participants speak the relevance scores of the documents being considered, rather than having them log this information themselves using the interface or a paper-based form, reduces the interference of the data collection methods with conducting the assigned search tasks.

2. Logging the spoken relevance scores manually allows the experimenter to closely watch how each participant uses the systems.

6. Conclusions

Conducting user-oriented evaluations of intelligent and interactive Web search interfaces is difficult due to the high degree of interactivity promoted by such systems. While numerous studies have been conducted to evaluate Web search systems both in isolation and as comparisons among competing systems, there is little consensus on the proper methods to use. This paper presents a review of the methods used in a selection of user-oriented studies. A summary of the techniques we have found to be effective in the evaluation of our systems are provided, with the focus on controlled experimentation with multiple interfaces for Web search.

References

[1] 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.

[2] H. Greisdorf and A. Spink. Median measure: an approach to IR systems evaluation. *Information Processing and Management*, 37(6):843–857, 2001.

[3] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[4] O. Hoeber and X. D. Yang. A comparative user study of web search interfaces: HotMap, Concept Highlighter, and Google. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[5] O. Hoeber and X. D. Yang. Interactive web information retrieval using WordBars. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2006.

[6] O. Hoeber and X. D. Yang. The visual exploration of web search results using HotMap. In *Proceedings of the International Conference on Information Visualization*, 2006.

[7] O. Hoeber and X. D. Yang. Evaluating wordbars in exploratory web search scenarios. *Information Processing and Management*, 2007 (to appear).

[8] P. J.-H. Hu, P.-C. Ma, and P. Y. Chau. Evaluation of user interface designs for information retrieval systems: a computer-based experiment. *Decision Support Systems*, 27(1):125–143, 1999.

[9] 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.

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

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

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

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

[14] B. Shneiderman. *Designing the User Interface*. Addison-Wesley, 1998.

[15] 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.

[16] 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.

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

[18] A. G. Sutcliffe, M. Ennis, and J. Hu. Evaluating the effectiveness of visual user interfaces for information retrieval. *International Journal on Human-Computer Studies*, 53(5):741–763, 2000.

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

[20] 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.