EVALUATING HYPERMEDIA SYSTEMS

Moderator: Gary Perlman, The Ohio State University
Department of Computer and Information Science
2036 Neil Avenue, Columbus, OH 43210
614-292-2566, perlman@cis.ohio-state.edu

Panelists: Dennis E. Egan, Bell Communications Research
Kate Ehrlich, Sun Microsytems
Gary Marchionini, University of Maryland
Jakob Nielsen, Technical University of Denmark
Ben Shneiderman, University of Maryland

Hypermedia systems provide online access to complex networks of information with the goal of making it easier to find and use information. To validate the utility of their systems, several researchers and system developers have attempted to collect evaluation data on the usability and effectiveness of their systems and the features in their systems. Because of the potential complexity of hypermedia systems and the information structures they may represent, a variety of evaluation measures and methods have been used. These trade off the need for timely feedback in the development of new technology, the difficulty of controlling one or two variables in systems with dozens or hundreds of components, and the goal of gaining an understanding of hypermedia systems.

The key issues discussed by the panel include:

Ecological Evaluation of New Technologies Embedded in Complex Systems: How can the utility of new technologies be evaluated validly when they must be embedded in complex software systems that include a hardware platform, underlying user interface, and a myriad of functions? Are controlled experiments necessary and can they be performed economically? What problems can occur in naturalistic settings?

Measures of Learnability, Usability and Effectiveness: What performance measures are most useful? How does the choice of measure depend on the maturity of a system? on the tasks to be done with a system?

Application to Human-Computer System Evaluation in General: What have been some results about hypermedia systems as a result of *empirical* evaluation? How does the evaluation of hypermedia systems apply to the evaluation of general systems? What guidance can be given to designers and users of hypermedia systems?

DENNIS EGAN

Ecological Evaluation

I think that there are roles for at least four types of evaluation studies. Each has different goals and costs. Each is appropriate to a different phase of development. Each has a different degree of *ecological validity*.

1. Conceptual laboratory experiments. Sometimes it is important to establish a theoretical principal, or to find out in a rigorous way what causes a particular effect, or how large an effect might be. In these formal experiments, contrived laboratory conditions may not be strictly ecologically valid. For example, a "Wizard of Oz" prototype system might be evaluated before the technology required by the system actually exists. Experimental tasks may be much more highly structured than is possible in real life. Values of an experimental factor might exceed realistic bounds simply to establish a function between the factor and the dependent variable.
2. Quick, inexpensive prototype *debugging*. I agree with Nielsen that it is possible to learn a lot by observing a few people trying to use a prototype that still has *bugs* in it. The data here pass the *interoccular trauma test*--results just hit you between the eyes. Some difficulties users have are obvious, and sometimes users offer spontaneous observations and suggestions that are quite valuable. At this point design iterations are extremely frequent.
3. Evaluation of prototypes using prospective end users. While there is still an opportunity to change a new system, studies with some of its prospective end users can be used to fine tune it. By this point, major interface difficulties have been minimized, but the interface may still not be optimal for the target user group. The specific tasks, materials, and environment of the end users should be simulated as closely as possible. Subjective reactions as well as performance are important to measure.
4. Naturalistic observations. This type of study asks whether the system really meets a critical need, whether people choose to use or avoid the system, and how the embedded system affects users' efficiency, productivity, and satisfaction. Here, the system already has been installed. The next design iteration will have to wait until the next release, which may be a long time in the future.

Measures of Learnability, Usability, and Effectiveness

Depending on the goals of the evaluation (see above), a variety

of measures are useful. A conceptual laboratory study might wish to consider the asymptotic performance possible with a new design. A quick *debugging* study will almost certainly focus on the initial difficulties experienced by new users. A naturalistic study of an embedded system might consider how often the system is accessed for which tasks, and how the system changes productivity and job satisfaction.

In general, we probably have paid too little attention to affective measures that would indicate whether potential users would put forth the effort and cost associated with learning and using a new system. I also think that we need to compare performance based on current *low tech* methods of accomplishing target tasks with performance based on the new system.

Application to Human-Computer System Evaluation

Our experience with SuperBook includes evaluations of each of the four kinds I listed and uses a variety of measures. As a result, we have brought SuperBook from some conceptual work on verbal disagreement, rich indexing, and fisheye lens through several prototype designs to something that actually may be useful to people. It is important to note that our first real prototype produced performance that in some respects was worse than using printed paper materials. Successive analyses and design iterations have changed the SuperBook interface so that searching a 500-page reference text with SuperBook results in a 25% improvement in speed and accuracy over that possible with the text in printed form.

In my opinion, human-computer interface design will play an increasingly important role in determining which systems people adopt. As the interface becomes more important, I suspect that a new *consumerism* of interface design will develop. Potential users will become much more sophisticated in making choices about systems. Users will demand to see evidence of increased productivity, more useful functionality, and training support in addition to technical specifications and cost.

To designers I would say, "know the basic literature on human-computer interaction, allow your designs to be influenced by user studies, and be prepared to change designs based on the results." To users I would say, "demand to see the results of user studies."

KATE EHRLICH

Ecological Evaluation, or, Design of Applied Research

In our usability testing we strive to maintain a balance between addressing the near-term usability issues, and the more general applied research issues. In the context of our study of the Sun386i Help system, these usability issues included: observable difficulties learning or using the system, feature usage and problems associated with double-clicking on the mouse to traverse a hypertext link.

The real focus of the study, however, was on addressing the more general applied research issue of how people navigate in a hypertext system. Based on previous research we identified two major styles of navigation. One style, called browsing, is characterized by the user who searches through a large number of options until an appropriate topic is recognized. The other style, called analytical search, is characterized by the user who generates a short list of options based on recall of the topic. Based on previous research, we further conjectured that individual differences in visualization skills could affect the efficiency of finding information in a hypertext system. The study did indeed reveal several critical factors that influence information retrieval.

Below I describe the elements in the design of this study that I believe are important for carrying out such applied research:

- Formulation of clear, testable set of questions. The study was designed to address the specific question of navigational strategy rather than a vaguer question of the utility of our particular hypertext system.
- Translation of the questions into an operational form which lends itself to a controlled, replicable study. In our study we associated a browsing strategy with the use of the Table of Contents, and the analytical strategy with use of the Index. By providing these operational definitions for navigational strategy, we enable other researchers to carry out similar studies with other systems that have different characteristics or with a different user population or different subject matter in the text. The ability to replicate a study is one of the best methods for ensuring validity and for being able to generalize the results beyond the peculiarities of a particular system.
- The task given to the subjects has to be very carefully thought through. We ended up asking people to act as consultants and answer questions sent to them in mail by (fictitious) users. This task ensured that they used the Help system with some purpose in mind.
- The system needs to be solid and well-designed. It is very difficulty to conduct a meaningful study on a system or application that has known flaws. The Sun 386i Help system, for instance, had an excellent interface and was relatively easy to use.
- We also validated the "laboratory" study with a field study in which we monitored usage of the help system over a period of several months. Data for this study were collected from a software monitor and from a "diary" each person was asked to complete every time he/she used the help system.

Measures

The measures relate to the questions that are being addressed. We used observational techniques to learn about ease of use but we used more objective measures of success rate, response time and frequency of feature usage to measure other aspects of performance. Key to our study was the development of a software monitor that automatically recorded and time-stamped all user interactions with the system. This enabled us to get accurate, reliable data especially for our objective measures.

Application

The main results of the applied study demonstrated that people prefer to find information by browsing and that skill in visualization is strongly correlated with the speed and efficiency of finding information in a hypertext system. These results have some applicability to design of future systems in that they imply that systems should provide signposts for the information architecture of the content of the hypertext system as navigational cues to its users. From the usability slant of the study, we learned that people use a limited set of features and then stick to them. By comparing the results of the lab study and the field study we found that the particular features selected varied from one individual to another. We also found that there is a high error rate associated with double-clicking on the mouse to follow a hypertext link. This error rate may be due to mis-timing, to the user moving off the target between the first and second click or to a failure to be on the target at all. This level of detailed analysis is only possible from the software monitor data; observation alone cannot reveal the source of error in double-clicking.

GARY MARCHIONINI

Ecological Evaluation

There are three difficulties in measuring hypermedia effects on information seeking and learning: novelty, complexity, and interactivity. First, hypermedia are both novel and emerging environments requiring new kinds of literacy. Separating out the novelty effects from learning or retrieval effects is difficult. Moreover, since most systems for creating hypermedia are at present very primitive, important effects may not yet be facilitated or may be masked by system limitations. Second, information seeking and learning are complex tasks that are difficult to assess since criterion measures are themselves controversial. Although performance on fact retrieval tests is acceptable for the knowledge acquisition level of learning, most hypermedia applications aim at the analysis, synthesis, or evaluation levels of learning; assessing such learning is subjective and qualitative. Third, one of the essential characteristics of hypermedia use is the interaction among people and computers. It is increasingly apparent that what is important about such interactions is the process itself rather than some final product. Although we have powerful methods for assessing outcomes scaled on ordinal or even interval scales, methods for assessing patterns of interactions must be developed and tested.

In earlier work, we collected keystroke data unobtrusively, made observations, and conducted interviews with subjects. Behavioral data were used to describe gross, but distinct information seeking patterns termed analytic and browse strategies. In our current evaluation studies, we are taking a multi-faceted approach to evaluation. In our plans for evaluating the effects of Perseus on learning topics related to the ancient Greek world, we will use observations of groups and individuals, interviews with instructors and learners, document analyses, comparisons of the products of learning, and logs of learner-system interactions. We are developing tools for mapping keystroke or mouseclick data onto state spaces for tasks and system; and for representing the traces of these interactions graphically. Recognizing the limitations of each of these methods individually, we believe that constructing multiple views of complex interactions will allow us to at least provide rich baseline data for affecting future systems and applications, and more importantly, to develop an integrated understanding of the process of human-computer interaction itself.

Measures

We will include measures of system learning such as time to complete tasks and number of features used, and will assess measures of learning outcomes such as performance on examinations and assignments. However, we are most concerned with evaluating the quality of interactions. To this end, we will examine number of "conceptual moves" made, time invested in groups of moves (paths), verbal reports of subjects, and systematic participant observations.

Application

Results from previous studies provide evidence that users are guided by cognitive inertia (they accept defaults and minimal levels of system complexity) and prefer browsing strategies to highly planned analytical strategies. Results from the first years of the Perseus evaluation will be applied to the redesign and extension of subsequent releases of the environment, and longitudinal results will inform our understanding of learning and teaching in highly interactive electronic environments.

JAKOB NIELSEN

My main position on this panel is that one should follow the *discount usability engineering* method [Nielsen 1989a] in evaluating hypertext usability. It is not worth the effort to conduct sophisticated videotaped experiments as long as there are major catastrophes in the interface which can be found much more cheaply. Instead one should rely on fast iteration to debug the interface. I will give a few examples of this from my experience in developing a hypertext system with individualized context.

The need for discount methods is especially critical for the evaluation of large hypertexts (hundreds of thousands of nodes). Hypertexts basically have no regular structure, so usability problems may crop up in any individual node or link. Therefore the user interface design of a large hypertext is distributed over potentially millions of locations in the information space. Since it will be impossible to test the usability of all the nodes and links large hypertexts with traditional methods, we need to rely on heuristic methods [Nielsen & Molich 1990] in the development process instead. Detailed empirical evidence will have to come later from field use of the hypertext using methods like navigation logging and user relevance feedback (e.g. have buttons where users can click to indicate that "this link is **useless**").

Actually my true position is not as extreme as the one I have outlined here for the purpose of generating controversy on the panel. I do support the use of traditional laboratory-based experiments for purposes such as generating lists of usability heuristics for hypertext. I will also recommend field studies to supplement laboratory studies because hypertext usability is extremely dependent on individual user characteristics and the users' tasks [Nielsen 1989b]. Hypertext systems are similar to e.g. integrated software for business professionals in having their usability determined by embedded use in environments where users interpret the information in the nodes and links relative to their own knowledge and tasks. Therefore one will often not be able to predict the true usability of a hypertext by giving users artificial tasks where they cannot use situated skills.

For example, one of the questions asked by Nielsen and Lyngbaek [1989] was *How confident are you that you have found all the information of interest to you in the hypertext?* This question is extremely relevant for an assessment of the usability of the hypertext system in question. But this type of confidence rating would be meaningless if we had first forced users to spend a specific amount of time using the system and navigating to find the answers to a set of tasks defined by the experimenters. The hypertext ideal is to empower the user to be in control of the information.

GARY PERLMAN

Ecological Evaluation

It is demanding to evaluate the utility of a new technology if it must be embedded in a complete system. A hypertext system might include a window manager, editing capabilities, formatting capabilities, a query language, *etc.* In a complex system, it is exceedingly difficult to evaluate the utility of a single feature. User interface issues have a strong effect too. If a hypermedia capability is added in a way that makes it difficult to use, then it might appear that the capability is useless. In contrast, if non-hypermedia parts of a system are implemented with a poor user interface, say, a query language and no form-filling interface to database records, then hypermedia capabilities might look artificially attractive.

The method of specific advantages (Perlman, 1989) provides a practical means of making valid comparisons among complex alternatives. In the method, a task-specific, user-specific, or situation-specific advantage (by some measure) must be demonstrated for each feature/capability in a system. For example, it may be shown that in some situations users prefer to use a browsing strategy while in others they prefer to use keyword search. If, even after practice, users always prefer one over the other, it would not be clear if it were due to its intrinsic merit or some deficit of the other.

Controlling all but a few factors in *commercial* systems is difficult and impractical if a system is to be delivered to market, but lack of control allows for the possibility that there are features of questionable utility or of mediocre implementation. Controlled experiments may become more prevalent when we gain more experience with hypermedia systems and basic results such as those found on database systems or user interfaces can be demonstrated. With our current state of knowledge, systems vary so widely in their functionality that it may be acceptable to show that hypermedia systems have advantages over printed text and over simple page-turners with string or pattern search.

Measures of Learnability, Usability and Effectiveness

The addition of a new capability can be evaluated by such measures as how often it is used and how highly users rate it. If there are many ways to accomplish the same task, then these can be compared for frequency of use. Benchmark tasks with well defined goals are needed to evaluate competing systems, competing features within systems, or competing organizations of information. Defining the goals is often not easy, because we are dealing with complex information structures; reliable expert ratings of the correctness of a result are needed. During development a new system, there are likely to be many problems and observing a variety of users will suffice. After a system is released, measures of system or feature effectiveness are more appropriate. I have been refining a measure, based on signal detection theory, that integrates the positive result of finding relevant chunks of information (hits) with the negative result of finding irrelevant information (false alarms). This has allowed me to compute speed-accuracy tradeoffs for hypertext systems and other presentation formats.

Application to Human-Computer System Evaluation

Most of the concrete results with NaviText™ systems (Perlman, 1989) have been found by informal observation but confirmed by analysis of protocol information logged by the software. One result is that the ability to use an electronic version of a book depends on how well the user can map their knowledge of books onto the hypertext system. NaviText™ systems are not modelled after a book, and users without a explanation mapping book-use expertise to NaviText™ functions get confused about how to proceed, while users provided with such a mapping do not show the same problem. Another result is that in a *new* information structure, experience users of NaviText™ SAM use an outliner and browse links to learn about the structure of the information space, while when searching for information in a *familiar* space, a keyword search strategy is preferred. This recognition-to-recall transition was also verified by protocols. Another result supported by protocols was that experienced users of NaviText™ SAM would add promising looking links to a *queue* of possible chunks to examine, rather than follow them directly as in a *stack*. The BFS over DFS strategy prevents getting *lost in hyperspace* and obviates backtracking, which is somewhat awkward in NaviText™ SAM, raising a question of specific advantages: *Would the strategy arise if backtracking were easier?*.

Many methods used in the evaluation of NaviText™ systems apply to user interfaces in general, particularly the method of specific advantages. *Designers of systems* should design evaluation into their systems, allowing the logging of usage data for later analysis. In the requirements for hypermedia systems for particular applications, there should be *performance requirements* that compare the systems to *plausible strawmen* like the paper version of a document, or a simple online format such as a word-processor with a search function. After development, *users* should demand to see quantitative data showing the effectiveness of hypermedia systems. The OSU *hypermedia technology assessment project* is an attempt to classify the myriad of features of hypermedia systems, and to use a taxonomic inventory to compare dozens of systems donated for evaluation.

BEN SHNEIDERMAN

My own view is that there are three goals of evaluation:

1. to improve a specific hypertext by finding out if users can understand the structure of information and its scope. These evaluations could lead to revisions which might dramatically improve its usability.
2. to improve a specific hypertext system. Many issues of design can be changed so that the system becomes more usable - indexing techniques, pointing techniques, commands, etc.
3. to improve user interfaces in general. Hypertext research can lead to a better understanding of fundamental issues such as window management, screen readability, window size, pointing devices, menu structures, etc.

REFERENCES

Campagnoni, F. R. & Ehrlich, K. (1990) *Information Retrieval Using a Hypertext-Based Help System,* **ACM Trans. Office Information Systems.**

Egan, D. E. *et al* (1989) *Behavioral Evaluation and Analysis of a Hypertext Browser,* **Proceedings of CHI'89.**

Liebscher, P. & Marchionini, G. (1988) *Browse and Analytical Search Strategies in a Full-Text CD-ROM Encyclopedia.* **School Library Media Quarterly,** Summer, 223-233.

Marchionini, G. & Shneiderman, B. (1988) *Finding Facts vs. Browsing Knowledge in Hypertext Systems,* **IEEE Computer.**

Marchionini, G. (1989) *"Evaluating Hypermedia-Based Learning."* Presented at the NATO Advanced Research Workshop on Designing Hypertext / Hypermedia for Learning. Rottenburg, West Germany, July, 1989.

Marchionini, G. (1989) *Making the Transition from Print to Electronic Encyclopedias: Adaptation of Mental Models.* **Int'l Jl. of Man-Machine Studies, 30,** 591-618.

Marchionini, G. (1989) *Information-Seeking Strategies of Novices Using a Full-Text Electronic Encyclopedia.* **Journal of the American Society for Information Science,** 29(3), 165-176.

Nielsen, J. (1989a) *Usability Engineering at a Discount.* In G. Salvendy & M. J. Smith (Eds.) **Designing and Using Human-Computer Interfaces and Knowledge Based Systems,** Amsterdam: Elsevier Science, 394-401.

Nielsen, J. (1989b) *The Matters that Really Matter for Hypertext Usability.* **Proc. ACM Hypertext'89,** 239-248.

Nielsen, J. & Lyngbaek, U. (1989) *Two Field Studies of Hypermedia Usability.* **Proc. Hypertext'2 Conf.** 29-30.

Nielsen, J. & Molich, R. (1990) *Heuristic Evaluation of User Interfaces.* **Proc. ACM CHI'90.**

Perlman, G. (1989) *Asynchronous Design/Evaluation Methods for Hypertext TechnoLogy Development.* **Proceedings of ACM Hypertext'89,** 61-81.