A USE CENTRED FRAMEWORK FOR EVALUATION OF THE WEB

Annelise Mark Pejtersen and Mark Dunlop
Centre for Human Machine Interaction, Risø National Laboratory

Raya Fidel
Graduate School of Library and Information Science, University of Washington

Experimentation on the usability and effectiveness of web search engines, and the web as a whole, can lead to misleading results if the evaluation work is carried out too early (e.g., planning to evaluate the functionality of the web in advance of a test of the interface readability) or without setting a suitable set of constraints, or *boundary*, around the experiments (e.g. not suitably controlling the work domain and strategies when evaluating interface readability). The boundary along which the cut is made varies considerably, depending on the aim of the experiment. Therefore a suitable framework for identifying the experimental boundary conditions must be applied. This short paper proposes the use of the framework shown in Figure 1 for defining different types of evaluation boundaries when evaluating the worth of web search engines and World Wide Web information seeking in general.

The innermost boundaries on Figure 1 correspond most closely to the traditions of experimental psychology. The remaining boundaries move the context successively further from the actor to encompass more and more of the total work content and context to give increasingly complete and realistic evaluation methods. The various categories of evaluative experiments will now be presented with reference to the use of the framework for specifying the boundary conditions in order to evaluate how well web search engines, and browsers as a whole, match the defined demand at each level. The categories are labelled with reference to the boundaries shown in Figure 1: Evaluation of how well a system match the users' sensory-motor characteristics, the users' cognitive capabilities and mental processes, the cognitive decision task and strategies, the work task situation and the work environment and the social organisation of work. Examples for each layer are based on several empirical studies and observations from different use domains.

It follows that a systematic evaluation according to the scheme shown in Figure 1 will proceed systematically inwards from global to local features. At the innermost boundaries, evaluation is focused on the match of the form of the displayed information to the users preferred strategies and tactics for coping with work requirements in an effective and for them acceptable way.

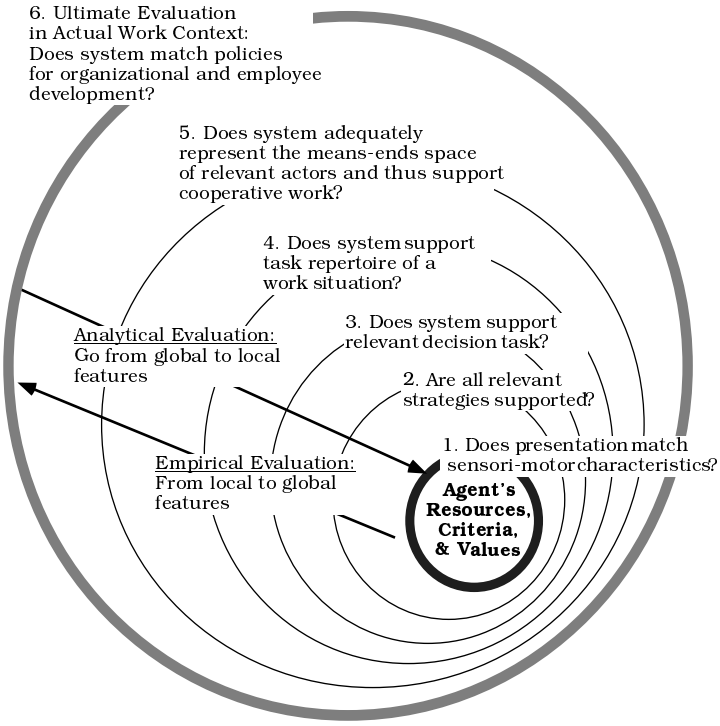

Figure 1: The figure demonstrates how different evaluation questions can be asked at the various levels described in the framework for evaluation

Does the Web Match User Characteristics

Experiments with the ergonomics of the system at this boundary will examine whether the physical configuration of a system, the equipment, and its arrangement at the users' work place correspond to the anthropometric and sensory characteristics of the user group.

This level also addresses evaluation principles that are important for the understandability of the information flow in the communication between the system and the user. Numerous experiments and empirical data have been published during the last decade in human factors handbooks, checklists and guidelines leading to general design principles.

Examples at this level include problems users have in understanding the syntax of search engines (and in particular boolean logic), lack of understanding on how keyword matching works and getting lost in a complex network because links do not adequately describe their destination and loading times are slow so short term memory is overloaded.

Does the Web Support Mental Strategies

This dimension introduces the concept of mental strategies. A strategy is a category of cognitive task procedures which is based on a particular kind of mental model and the related interpretation of information, and on a particular set of tactical rules.

The question here is to compare task requirements with the cognitive resource profile and the subjective performance criteria of the individual actors. Are the mental strategies that can be used for each of the decision functions supported? This evaluation is done by a detailed analysis of the actual work performance of several individuals in different situations. The evaluation of the support of characteristics of the various strategies is performed with respect to subjective performance criteria such as time needed, cognitive strain, amount of information required, and cost of failure.

Our studies have highlighted five main strategies for finding information on the web: browsing, analytical, empirical, known site and similarity.

Example problems with different strategies include:

- Problems browsing when homepages are not structured inline with users' tasks;
- Analytical searching can fail if searching an unknown domain where the user is not fluent with synonyms/related-terms to revise their query with;
- Empirical searching can get into difficulty when a previous search approach fails because the search engine index has changed or pages have moved.
- Known site strategy relies heavily on pages not changing and bookmarks working which can fail with script based pages.
- Similarity searching is not typically supported at all although there are some moves to support this (e.g. within Apple's Sherlock and AltaVista image finder).

Does the Web Adequately Support The Cognitive Decision Tasks

Going from boundary 2 to 3 results in a focus on studies of *problem solving* and *decision situations*. Here, the cognitive processes are more complex, and more constrained by the task environment, but less constrained by the experimental tool. A basic question to be asked at this level is: Does the system effectively support the cognitive decisions that have to be made during task performance? Does the system support the actor's decision making and exploration, situation analysis, goal evaluation and planning supported for familiar as well as less familiar situations? Can decisions be made smoothly without any interruptions?

The use of web browsers and search engines to support decision tasks is often left with the user to manage and they will develop their own strategies.

Problems of smooth integration at this level may occur, for example, because the time delay, and in particular the unpredictability of loading time, disrupt work processes and make it difficult to schedule (Johnson).

Does the Web Adequately Support The Task Situation

Does the system adequately support the actual work task situation, is its capacity adequate? The question is here whether the system supports the entire task repertoire, are the tools adequate, their functionality sufficient and does the information cover the complete work task space? Experiments may serve to evaluate whether information is available about the basic concepts of the system and its overall architecture. Is it possible to navigate among tasks, and to pursue several, different task related goals?

One important aspect of support for task situations is integration of web browsing and searching with the normal working environment. This is improving but still lacks in many ways, for example when writing academic papers references are often searched for on the web but there is often no easy way to add references to one's personal database nor to the paper one is writing without considerable reformatting.

Does the Web Adequately Support Role Allocation and Co-operative Work

There is no functionality in the Web that directly supports role allocation and co-operative working. The Web does, however, provide a standard document and communication channel though HTML and HTTP. As such, it is used extensively for remote co-operative working. However if two authors are working on a document together, lack of direct support leaves versioning, ownership and security problems completely in the hands of the authors. This often leads to extensive telephone or e-mail communication to backup co-working on shared documents though web pages. For searching, approaches such as collaborative filtering, path models or recommender systems (e.g. Chalmers et al 1998, Glance 1998) are starting to make information seeking on the web a much more collaborative and group exercise but have still to be adopted by large user groups (with exceptions of on-line shopping examples such as Amazon.Com). These new technologies also introduce interesting evaluation challenges as the success of such techniques must be measured within a group.

Field Evaluation in Actual Work Environment

As the ultimate step evaluation at boundary 6 in the actual work context will address the question: Does the web match policies for organisational and employee acceptance and development? How is its impact on the work context and the quality of the work situation? Does the web support several coherent work task activities and the co-operative co-ordination of activities among several users, maybe in different departments of the organisation, and does it

support interaction and co-ordination with institutions outside the organisation?

In engineering design, the product being designed should effectively match the goals and constraints of a variety of domains, including marketing and sales, distribution, manufacturing, and maintenance and repair. The consequence of this development is the need for web systems that are able to support dynamic co-operation in a complex network of co-operating decision makers, including engineers, managers, and subcontractors, having very different expertise and professional backgrounds, concerns and objectives, time horizons, and information needs within an organisation and across organisations. Therefore, access to information on past business practices, available resources and future plans, etc. that are often found in a variety of departments within a company need to be organised and structured in a way that is compatible with their task requirements.

This is recognised as a major problem for professional, co-operating engineers who face increasing difficulty in coupling their work related information needs to the content of heterogeneous information sources on the web. The web browsers use a uniform icon based interface and standard telecommunications protocols to support access to multiple databases.

However, these systems do not support the semantic coupling of engineers' information needs to the semantics of information sources, i.e. they do not support access to information that is organised and displayed as means and ends that are compatible with co-operating end users' tasks and job experience.

Limitations of Human Factors Guidelines as a tool to evaluate the web

As defined here, traditional guidelines are prescriptive recommendations for identifying and incorporating human factors requirements into a product/system in a way that improves the overall quality of the design.

Guidelines can range from the very specific to the very general i.e., from hard data concerning specific ergonomic features to more general design criteria to checklists, structured questionnaires, flow diagrams of the design process, etc. Usually these various types of aids are general purpose and therefore context free in content issued as they are by governmental and defence agencies or by user groups within a large encompassing domain.

Traditional design guidelines apparently do very little to foster intuition and creativity and/or a re(usable) knowledge base. On the contrary, they represent an attempt at providing advice about human factors issues to people who have little or no human factors expertise. As a result, experienced designers will tend to ignore guidelines as a basis for evaluation and design and base their design decisions on the factors that they are used to considering. This suggests that human factors will be put off until someone realises that there is a problem, but by that time it may be too late to remedy the situation in a cost-effective manner.

The specific design solutions that are developed for any one application depend on a host of unpredictable factors that are unique to that particular problem. Since guidelines are intended to be generally applicable across a wide number and variety of design problems, they cannot possibly capture the rich sensitivity to context that is required for effective design. The point is well described by Gould (1988): "Guidelines cannot deal with choices that are highly dependent on context, as many choices in interface design are. Human performance adapts strongly to details of the task environment. There are simply so many details, and this adaptation is so little understood, that guidelines cannot hope to anticipate all of this" (p. 782). Evaluation and design is heavily context dependent whereas guidelines attempt to be context free. An excellent example of this point has been illustrated by Grudin (1989) who has pointed out the limitations of one the most frequently mentioned design criteria, namely that interfaces should be designed to be consistent. Through the use of several detailed examples, Grudin demonstrates convincingly that "interface consistency is a largely unworkable concept; the more closely one looks, the less substance one finds" (p. 1164).

The basic reason for this is that there are other considerations which may be more important and which should therefore override consistency as a design criterion. Most of these other factors are contextual, depending upon the characteristics of the user and the work context. Thus, it is the users' work contexts that should be the primary constraint on evaluation and design, not interface consistency. We would go one step further in hypothesising that the difficulties identified by Grudin with consistency generalise to most, if not all, evaluation and design criteria. Exceptions are always possible and design tradeoffs need to be made with a deep and clear understanding of the available resources as well as the constraints imposed by the work context.

This is not to say that traditional guidelines are useless; only that they are extremely limited as a sole source of support for the evaluation and design.

Another problem has to do with the trade-off between specificity and generality. Results from surveys indicate that, in certain situations, some designers would rather have specific rules that they can follow in a rote manner than general principles which must be instantiated and translated before they can be directly applied (Smith 1988). This is probably particularly true for designers who do not have a background in human factors and therefore the expertise necessary to make the jump from general principles to design decisions in an effective manner. The obvious difficulty with making guidelines too specific is that generality is lost.

The reverse problem has to do with the question of whether our understanding is deep enough to even justify specific recommendations.

As a result of this inherent and necessary tension between the specific and the general, it is generally acknowledged that guidelines must be interpreted within the context defined by the design problem in question, and that the use of guidelines needs to be mediated by good judgement. Therefore, a more productive form of guidance will have to instil and utilise a sense of framework issues as described above.

References

Chalmers, M., Rodden, K., and Brodbeck, D. "The Order of Things: Activity-Centred Information Access", *Proc. WWW7*, pp. 359-367, Brisbane, April 1998.

Glance, N., "Putting recommender systems to work for organizations", In: *Proc. of Workshop on Recommender Systems, AAAI'98,* AAAI Technical Report WS-98-08, July 1998

Gould, J.D., "How to Design Usable Systems". In Helander, M. (ed): *Handbook of Human-Computer Interaction*, pp 757-789, Amsterdam 1988.

Grudin, J., "The Case Against User Interface Consistency", *Communications of the ACM*, 32, pp 1164-1173, 1989.

Johnson, C. W., "Time And The Web: Representing Temporal Properties Of Interaction With Distributed Systems", *People And Computers X: Proceedings of HCI'95*, pp39-50, edited by M. Kirby and A. Dix, Cambridge University Press, 1995.

Smith, S.L., "Standards versus Guidelines for Designing User Interface Software". In Helander, M. (ed): *Handbook of Human-Computer Interaction*, pp 877-889, Amsterdam 1988.