▨ 연구논문

A Structured Model for Usability Evaluation of Software User Interfaces during the Design and Development Phases

- 설계 및 개발 단계에서 소프트웨어 사용자 인터페이스들의 사용성 평가를 위한 구조화 모형에 관한 연구 -

임 치 환*
Lim, Chee-Hwan

요 지

오늘날 인간-컴퓨터 시스템에서 상호작용 하는 컴퓨터 소프트웨어의 사용성에 많은 관심이 집중되고 있으며, 소프트웨어 인간공학 분야에서 이전 연구들은 소프트웨어 사용자 인터페이스들에 대한 사용성 평가의 중요성을 지적하고 있다. 소프트웨어 개발자들, 인터페이스 설계자들 또는 인간공학자들은 개발되는 시스템들이나 인터페이스 디자인들을 평가하는 일에 종종 직면하게 된다. 본 연구는 사용성 기준들과 척도들을 이용하여 사용자 인터페이스 디자인들의 평가를 위한 구조화 모형을 제시한다. 제시된 모형은 선별 단계와 평가 단계로 이루어진 두 단계 모형이다. 첫 번째 단계는 정성적인 기준들을 가지고 전문가들의 판단에 근거한 접근방법으로 가능한 인터페이스 대안들을 추려내어 합리적인 부분집합으로 줄이기 위한 단계이다. 두 번째 단계는 정량적인 기준들을 가지고 실 사용자에 근거한 접근 방법으로 객관적인 척도들을 가지고 첫 단계에서 제시된 대안들의 부분집합을 평가하는 것이다. 제안된 방법이 정보분석을 위한 데이터 베이스 시스템의 인터페이스 설계에 적용되었다. 본 연구에서 제안된 모형은 실제 평가자들에게 사용성 기준들과 척도들에 근거해서 최선의 인터페이스를 선정할 수 있는 구조화된 접근방법을 제시한다.

1. Introduction

The reasons for evaluating software user interfaces can be divided into some broad classes. These include assisting in design decisions, and measuring quality of use. In short, the distinction is between whether or not one is looking for new information about possible alternatives (e. g., user interface designs), or looking to report on the value of some measure for comparative purposes. In

* Dept. of Management Information System, Seowon University

some situations, one is interested in evaluation to select between two or more choices. For example, software developers or human factors engineers often confront the task of comparative evaluation among systems, versions or interface designs. In this type of evaluation multiple criteria are used because it is difficult to derive a single measure that effectively characterizes the overall usability of an interactive system.

Some authors have studied computer interfaces evaluation in terms of an integrated assessment of several interface characteristics. Mitta (1993) has suggested using the analytic hierarchy process (AHP) to rank-order computer interfaces based on multiple evaluative criteria. Stanney and Mollaghasemi (1995) also used the AHP to assess the relative importance of a realistic and an unrealistic desktop interface design. These approaches are able to allow simultaneous consideration of multiple criteria and to readily quantify consistency in the decision maker's judgments.

Mitta's approach, however, was based solely on subjective assessments. With this approach experimenters evaluated users with respect to their abilities to make satisfactory judgments regarding three criteria. Stanney and Mollaghasemi (1995) pointed out that the repeatability of this assessment procedure is questionable, and evaluated the interface attributes in terms of objective measures. The values of objective measures are mainly acquired through empirical assessment such as user testing or user trials. Because user testing is relatively time-consuming and expensive than subjective assessment, performing user testing for all alternatives may be difficult or expensive when multiple measures and several alternatives must be considered. Accordingly, a process for filtering possible alternatives to a reasonable subset is necessary for efficient evaluation.

The objective of this paper is to extend the earlier research, and to describe a structured approach for evaluating the usability of user interfaces. The decision framework as proposed in this paper, is made up of two phases, namely the prescreening phase (expert judgment-based approach) and the evaluation phase (user-based approach).

2. The First Phase

The first phase is mainly concerned with qualitative and subjective assessment. The objective of this phase is to filter possible alternatives to a reasonable subset. In this phase, experts mostly rely on their experience to make a judgment on the ergonomic quality of alternative interfaces. Lacking experience, they appraise alternatives with user interface design guidelines (Smith & Mosier, 1986; Shneiderman, 1992; Nielsen, 1993), established human factors principles and standards (e. g., ISO, ANSI, DIN, etc.), and criteria (Ravden & Johnson, 1989; Scapin, 1990). Expert judgment-based assessment has the advantage of providing an integrated view and using few resources. The

expert approach is based less on tasks to be performed by the tested system than on questions asked by user interface usability. Because this assessment inevitably requires some subjectivity, thus it is very important that the assessment be conducted by multiple experts (human factors specialists, system designers, software engineers, etc.).

This assessment also requires clear criteria against which to assess the quality of the user interfaces. Criteria should be agreed upon and identified before evaluation efforts begin. The sets of criteria currently available vary from one author to another in terms of number of criteria, degree of generality or specificity, and level of precision. The validity of these criteria has to be confirmed by fields and experimental studies with actual human factors specialists in interface design and evaluation.

As a result of the review and the comparison, eight criteria have been identified: suitability for the task (ST), user control (UC), flexibility (F), error management (EM), compatability (CP), self-descriptiveness (SD), consistency (CS), user workload (UW). If all criteria might be considered for alternatives, all criteria are not equally important: they should be weighted according to their relative importance. After defining the criteria, it is necessary to define possible alternatives. For example, the main alternatives in design process are various combinations of the types of user interface designs. Experts (i. e., members of the evaluation team) understand all alternatives under consideration and then evaluate them in terms of the usability criteria.

The proposed approach in the first phase uses absolute measurement AHP. Absolute measurement AHP is appropriate for situation where multiple criteria and many alternatives must be considered (Saaty, 1989). Absolute measurement AHP requires a pairwise comparison procedure between indicator categories (for each lowest level criterion) to establish the relative weights for these categories using eigenvector approach. A detailed procedure of this methodology may be found in Mullens, et al. (1995).

The objective of the first phase is to discard certain inferior alternatives and to reduce the number of alternatives under consideration. This gives us economical efficiency of analysis. After all, the high ranked (leading) alternatives are selected at the end of the prescreening phase. The results obtained from the prescreening phase are taken into the evaluation phase, which aims to evaluate the alternatives using objective measures.

3. The Second Phase

The objective of the second phase is to evaluate a subset of alternatives using quantitative criteria and to select the best alternative. The evaluation phase involves user-based assessment such as user testing, with quantitative criteria and measures. Since usability is a multi-dimensional concept that cannot be characterized by a single criterion, multiple measures are used in

usability assessment. A wide variety of quantifiable measures may be used, based on such factors as the specific interface to be tested, laboratory or field conditions, available test equipment, or aspects of the product.

Some of the measures used in usability testing include: time to complete task, number of tasks completed in a given time, ratio of successful interactions to errors, time spent recovering from errors, number of user errors, number of times a user expresses clear frustration during a test, number of times the interface misleads the user, proportion of users who say they would prefer using the system over that of some specified competitor, etc. (Nielsen, 1993; Whiteside et al., 1988).

ISO 9241-Part 11 (Guidance on usability) gives the following definition of usability (Bevan, 1995): "usability is measured by the extent to which the intended goals of use of the overall system are achieved (effectiveness); the resources such as time, money, mental effort that have to be expended to achieve the intended goals (efficiency); and the extent to which the user finds the overall system acceptable (satisfaction)". Effectiveness, efficiency and satisfaction can be seen as critical criteria that influence usability of interfaces. To evaluate these criteria, they need to be decomposed into sub-criteria, and finally, into usability measures. According to characteristics of usability measures, thus, each criterion (e. g., effectiveness, efficiency, satisfaction) can be decomposed into sub-criteria (or usability measures).

Measures of effectiveness relate the accuracy and completeness with which users achieve specified tasks. Measures of efficiency relate the resources that have to be expended to achieve the given tasks. The resources may be mental or physical effort, which can be used to give a measure of human efficiency, or time, which can be used to give a measure of temporal efficiency. Satisfaction is a subjective response of users to interaction with the system. Subjective measures of satisfaction are produced by quantifying the strength of a user's subjectively expressed reactions, attitudes, or opinions.

The relative importance of components of usability depends on the context of use and the purposes for which usability is being described. Effectiveness and efficiency are usually a prime concern, but satisfaction may be even more important, for instance where usage is discretionary. The relative importance weights of usability measures (sub-criteria) for each of criteria can be achieved through pairwise comparisons of measures with respected to their criterion.

A method is needed that integrates user testing results, thus providing a composite measure of usability to facilitate direct comparisons between interface design options. We also use the AHP to assess the relative importance for each set of usability components (e. g., criteria and sub-criteria) with the objective of selecting the best interface. The priorities (i. e., relative importance weights) of usability components are based on experts' judgments because users have not enough sense to judge the relative importance of them. The priorities of alternatives with respect to each of measures (sub-criteria) are based on user testing data. The priorities of alternatives and weights for each set of criteria

and sub-criteria are synthesized into a composite score for each alternative.

It is necessary to unite the first phase and the second phase outcomes (i. e., the priorities or scores of each alternative) for more reliable analysis. The formula for union of the two phase outcomes is:

$$P_i = w \times P1_i + (1-w) \times P2_i, \quad (0 < w < 1), \tag{1}$$

where

P_i = the combined priority of the ith alternative

$P1_i$ = the normalized priority of the ith alternative selected in the first phase

$P2_i$ = the normalized priority of the ith alternative selected in the second phase

The decision of the constant w depends on the analyst's judgment and the context of use. The alternative with the highest overall rating is ranked the best choice, taking into account user testing data as well as experts' assessments.

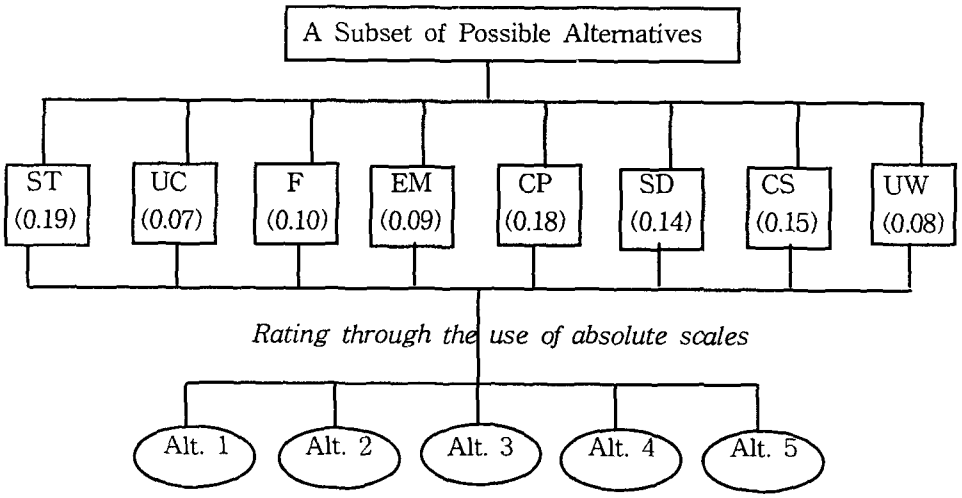
4. An Example

An application of the proposed approach involved a comparative assessment of several interface prototypes of an interactive system. For demonstration, five interface prototypes were designed to support a database system called Information System of Trip Event Cases (Park et al., 1996). This system running on an IBM-PC provides information obtained from the analysis of nuclear power plant trips, such as component failures or human errors that induced the trips, the sequence of unit events, and problems that contributed to the trips. The intended user population of this database system is general operators in nuclear power plant. The five alternatives, prototypes written in Multimedia ToolBook ™ for Windows, incorporated several types of user interface technology, with windows, screen layouts, system command modes (e. g., menu, button, and icon), colors, and information feedback.

Two hierarchical trees for selecting the best alternative are shown in Figure 1. The first phase started with five prototype alternatives. The first phase utilized absolute measurement AHP to identify a subset of alternatives from among the five alternatives. The evaluation team was composed of four members representing user interface design, software engineering and ergonomics.

From each set of pairwise comparisons, criteria weights and a consistency ratio were calculated using Saaty's eigenvector approach. The lowest level of the hierarchy represents the indicator categories comprising the scales for the criteria. For simplicity, three grade levels such as A, B, and C were provided for each criterion scale. The rating of all five prototype alternatives was performed with respect to qualitative criteria. For qualitative criteria, alternatives were rated through the use of absolute scales where grades have been assigned to alternatives according to how they fulfill the criteria.

Phase 1



ST:suitability for the task, UC:user control, F:flexibility,
EM:error management, CP:compatability, SD:self-descriptiveness,
CS:consistency, UW:user workload

Phase 2

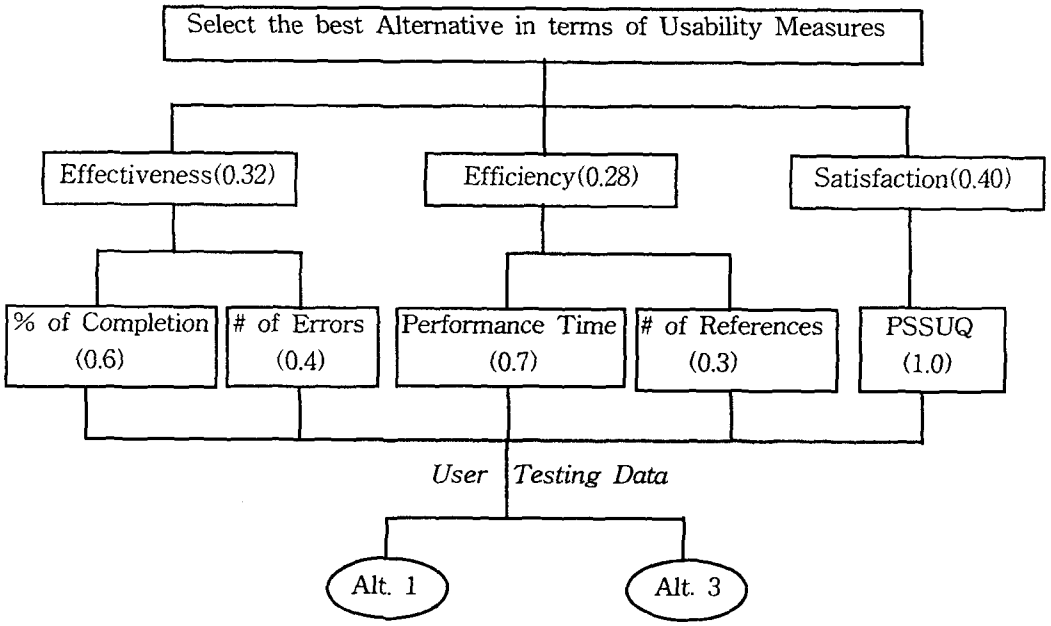


Figure 1. Overview of hierarchical trees in the proposed model

Consensus was reached on all ratings.

A composite score for each alternative was then developed by summing the product of the alternative ratings and their corresponding criteria weights. Table 1 provides an overview of criteria weights and alternative weights resulting in a rating of the proposed alternatives. The prototype alternative 3, with 24% relative priority, was favored. The analysis outcome in the first phase, however, indicates reasonably close outcomes especially between the alternative 3 and the alternative 1. These two alternatives were taken into the evaluation phase, which aims to evaluate them using objective measures.

Table 1. Usability criterion weight and alternative priority in the first phase

Criterion	Weight	Alt. 1	Alt. 2	Alt. 3	Alt. 4	Alt. 5
Suitability for the Task (ST)	0.19	0.25	0.48	0.34	0.23	0.31
User Control (UC)	0.07	0.56	0.29	0.36	0.48	0.23
Flexibility (F)	0.10	0.48	0.31	0.56	0.16	0.48
Error Management (EM)	0.09	0.23	0.16	0.16	0.24	0.29
ComPatability (CP)	0.18	0.29	0.23	0.23	0.25	0.16
Self-Descriptiveness (SD)	0.14	0.16	0.13	0.25	0.23	.0.20
ConSistency (CS)	0.15	0.48	0.25	0.56	0.29	0.23
User Workload (UW)	0.08	0.31	0.29	0.23	0.21	0.25
Rating	1.00	0.23*	0.19	0.24*	0.16	0.18

The two alternatives were assessed based on the evaluation criteria. This analysis used the percentage of successfully completing tasks and the number of errors as measures of effectiveness. Efficiency was quantified in terms of performance time and the number of references to help. Satisfaction was obtained by the Post-Study System Usability Questionnaire (PSSUQ) that is a kind of IBM questionnaires (Lewis, 1995). Because the PSSUQ has acceptable psychometric properties, usability practitioners can use the PSSUQ with confidence as a standardized measurement of satisfaction for user testing.

The two alternatives were tested by nine users. After a short tutorial on this system, each user performed a set of tasks on both prototypes. The mean performance times for the subjects were calculated to be 913 seconds and 876 seconds for the alternative 1 and the alternative 3, respectively. The mean number of tasks successfully completed for the subjects were determined to be 92 % and 96 % for the alternative 1 and the alternative 3, respectively.

Table 2. Usability measure weight and alternative priority in the second phase

Criterion and sub-criterion	Weight	Alternative 1		Alternative 3	
		Actual Measure	Relative weight	Actual Measure	Relative weight
Effectiveness	0.32				
percentage of completion	0.6 [0.192]	92 (%)	0.49	96 (%)	0.51
number of errors	0.4 [0.128]	4.2	0.46	3.6	0.54
Efficiency	0.28				
performance time	0.7 [0.196]	913 (sec)	0.49	876 (sec)	0.51
number of references	0.3 [0.084]	2.4	0.54	2.8	0.46
Satisfaction	0.40				
PSSUQ	1.0 [0.40]	3.81 (pts.)	0.45	3.17 (pts.)	0.55
Rating		0.47		0.53 *	

The summarized results are shown in Table 2. The weights of alternatives and weights for each set of criteria and sub-criteria were synthesized into a composite score for each alternative. The combined priority of each alternative was computed by equation (1). The constant w was determined to be 0.3. The overall rating for the alternative 1 was computed to be 0.476 ($=0.3\times0.49 + 0.7\times0.47$) while that of the alternative 3 was calculated to be 0.524($=0.3\times 0.51 + 0.7\times0.53$). The results from the two-phase model showed that the overall usability of the alternative 3 was better than the alternative 1.

5. Conclusions

This study presents a structured model for usability evaluation of user interface designs using usability criteria and measures. Usability evaluation usually has been conducted by human factors experts or by using user testing. Because each of two approaches has its advantage, we combine two approaches to evaluate user interfaces effectively. The proposed model consists of two phases: the prescreening phase (expert judgment-based approach) and the evaluation phase (user-based approach). It is particularly useful when multiple criteria and several alternative interfaces are considered and when usability evaluation must be performed under limited resources. The proposed model enables software developers or interface designers to efficiently evaluate interface designs through multiple criteria and measures.

References

1. 박경수, 장필식, 임치환, "The storage and implementation of retrieval logic for database on trip event analysis information in Korean NPPs", Report No. KAERI/CM-074-95, 한국원자력연구소, 1996.

2. 임치환, "A Two-phase Model for Usability Evaluation of Software User Interfaces", *'97 대한인간공학회 추계학술대회 논문집*, 무주리조트 호텔 티롤, 313-319, 1997.

3. Bevan, N., "Human-computer interaction standards", *Proceedings of the 6th International Conference on Human-Computer Interaction*, Yokohama, Japan, Elsevier, 885-890, 1995.

4. Lewis, J. R., "IBM usability satisfaction questionnaires: psychometric evaluation and instructions for use", *International Journal of Human-Computer Interaction*, 7(1): 57-78, 1995.

5. Mitta, D. A., "An application of the analytical hierarchy process: a rank-ordering of computer interfaces", *Human Factors*, 35(1): 141-157, 1993.

6. Mullens, M. A., Armacost, R. L. and Nippani, R., "A two stage approach to concept selection using the analytic hierarchy process", *International Journal of Industrial Engineering*, 2(3): 199-208, 1995.

7. Nielsen, J., *Usability Engineering*, Academic Press, London, 1993.

8. Ravden, S. J. and Johnson, G. I., *Evaluating Usability of Human-Computer Interface*,. Ellis Horwood, Chichester, 1989.

9. Saaty, T. L., "Decision making, scaling, and number crunching", *Decision Sciences*, 20: 404-409, 1989.

10. Scapin, D. L., "Organizing human factors knowledge for the evaluation and design of interfaces", *International Journal of Human-Computer Interaction*, 2(3): 203-229, 1990.

11. Shneiderman, B., *Designing the User Interface*, Addison-Wesley, 471-500, 1992.

12. Smith, S. L. and Mosier, J. N., *Guidelines for designing user interface software*, Report No. MTR-10090, Esd-TR-86-278, The Mitree Co., Beford, MA., 1986.

13. Stanney, K. and Mollaghasemi, M., "A composite measure of usability for human-computer interface designs", *Proceedings of the 6th International Conference on Human-Computer Interaction*, Yokohama, Japan, Elsevier, 387-392, 1995.

14. Whiteside, J., Bennett, J. and Holzblatt, K., *Usability engineering: our experience and evolution*, In: M. Helander (Eds.), Handbook of Human-Computer Interaction. Elsevier, 791-817, 1988.