Interacting With Computers, 12, 23-36. 1999.

THE VALIDITY OF RATIONAL CRITERIA FOR THE INTERPRETATION OF USER-HYPERTEXT INTERACTION

Andre Tricot
Brittany University Institute for Teachers' Education

and

Emmanuelle Puigserver, Dolly Berdugo and Mariama Diallo
University of Provence

Key words: Information retrieval, hypertext, interaction design, on-line interaction

Address for correspondence
Andre Tricot
Brittany University Institute for Teachers' Education
153 rue de Saint Malo
35 043 Rennes cedex
France
e mail : andre.tricot@bretagne.iufm.fr
phone: 33+ 2 99 54 44 02
fax: 33+ 2 99 54 44 20

Abstract

This study is about how to interpret users' interactions with multimedia systems, in particular their interaction with hypertext. It is generally admitted that our knowledge of how to interpret users' search paths in information systems is very limited. Furthermore, there are also controversies about the validity of the import of rational criteria, such as recall, precision and economy, from the domain of information retrieval to the interpretation of users' behavior. The purpose of the present study was to examine the relevance of these criteria to human behavior. The results of two experiments show that rational criteria such as precision and economy seem to be irrelevant criteria for the interpretation of users' search paths in terms of recall, except when the user's task is simple to achieve and constrained (i.e. a task where subjects have to find precisely a small number of relevant information units in a simple system). In the latter case precision and economy are positively correlated with recall. For less simple and unconstrained tasks, the performance of subjects seems to be much more influenced by strategic considerations. Furthermore, it was shown that task constraints lead to more precise performance. Apparently subjects spend more effort under these circumstances to search precisely than in conditions without pressure.

Introduction

The general topic of this study is the evaluation of user's performance in a hypermedia environment. In the field of interactive multimedia environments, user's performance evaluation is not easy because the same user's behavior is interpreted as adequate performance in the opinion of some authors, and as inadequate performance in the opinion of others. The main contradiction concerns the interpretation of user-system on-line interaction. Some criteria such as recall (i.e. when the user finds information that she/he was searching) are admitted by everyone as a good performance criterion. But other criteria, like precision (i.e. when the user does not open nodes containing information that she/he was not searching) or economy (i.e. the user opens only relevant nodes - only once -, and no irrelevant nodes) are:

- positively related to performance in the opinion of some authors (Edwards & Hardman, 1989; Foss, 1988; Rouet, 1990)
- negatively related to performance in the opinion of others (Bernstein, Joyce & Levine, 1992; Tricot & Coste, 1995)
- not associated with performance at all, in the opinion of others again (Bernstein, 1993).

Rational criteria such as recall, precision and economy come from the domain of information retrieval (Su, 1994; Buckland & Gey, 1994) , where these criteria are used to evaluate retrieval algorithms or indexing techniques. Our claim is that their application to the interpretation of users' performance is not straightforward or may even be irrelevant. The relevance of these criteria and the direction of the correlation between them depends, for instance, on the kind of task that the user has to achieve. In a previous experiment (Tricot & Coste, 1995), we have shown that for a complex task (answering 25 questions, involving problem solving and information retrieval) in a complex system (more than 1300 nodes), subjects were only successful on the task when they had opened more than four times the

relevant nodes (relevant means here relevant to answer the corresponding question). However, opening more than four times a relevant node was not always associated with a correct answer to the corresponding question.

In this study, we will examine whether the relevance of rational criteria to the domain of evaluation of user's performance depends on the complexity of tasks, for instance, the number of relevant nodes in the hypertext that users have to open to perform the task successfully.

Rational analysis and the interpretation of behavior

Using rational criteria to analyze human behavior involves a very common methodological issue in cognitive psychology and in human-computer studies (Anderson, 1990). A task can be described a priori in a rational way, even if the subject does not have a rational model in his/her mind: the subject interprets the task, builds a cognitive representation of it, and this representation will guide her/his activities (Figure 1).

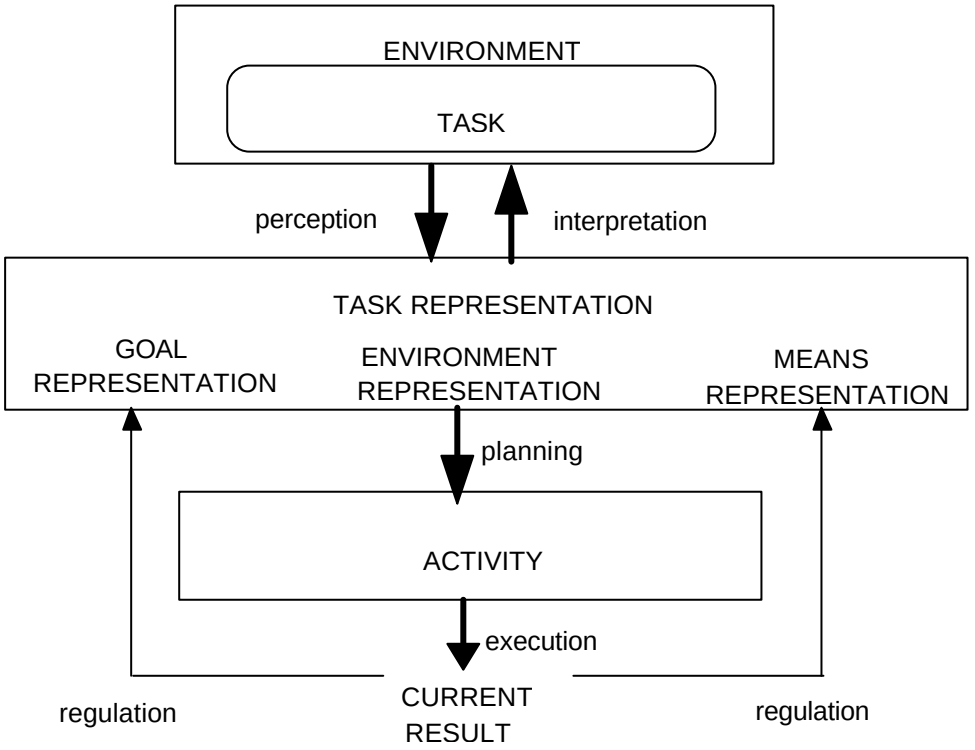

Figure 1. *Relations between task, task representation and activity*

We have proposed elsewhere a framework for the description of information usage tasks, in which a task is described in terms of a goal, several means to reach this goal and an environment that contains this goal (Rouet & Tricot, 1996). Each part of the task description can be done at three levels: the rational task level, the subject's cognitive representation task level and the subject's actual activity level (behavior). One way of interpreting subject's activity is to study the differences between a rational model of the activity (i.e. the "means" part of the rational task model) and the subject's actual activity. Then it is possible to infer in what way the

subject's cognitive task representation is different from the rational task model : representations differ on the points which correspond to differences between the rational model of activity and the subject's activity. But this interpretation is only possible when the relevance of the rational task model has been previously established. For example, Wright and Lickorish (1994) showed that a rational task model such as GOMS (Card, Moran & Newell, 1983) lacked relevance for information seeking activities, because the way users choose a strategy depends on their task representation, and this task representation depends on the users' perception of the system interface and its tools. This perception and representation are changing as the user discovers the system. Opposed to this, GOMS is only able to predict at the initial state of task performance what a user can do with the system interface and its tools.

This problem has important consequences for the domain of interactive multimedia design and the evaluation of human behavior. Because as long as we don't have adequate criteria to evaluate user's behavior, we will not be able to evaluate multimedia systems seriously. That is, we will keep on being able to evaluate these systems merely by referring to users' satisfaction feelings, or to technical specifications. As in the SuperBook project (Landauer, Egan, Remde, Lesk, Lochbaum & Ketchum, 1993) or in the famous Perseus project (Marchionini & Crane, 1994) we believe that user centered evaluation is crucial and should be part of the design process (see e.g. Nanard & Nanard, 1995).

Rational criteria in information retrieval and users' performance evaluation

In information retrieval, rational criteria are distinguished such as recall -that is, retrieving relevant information - and precision - that is, not-retrieving irrelevant information (Salton & McGill, 1983). A relevant piece of information is generally called a target. Both criteria (recall and precision) are related to the principle of economy, i.e. finding as many targets as possible with the lowest costs. Cost can be measured, for example, by time spent to find targets, or in terms of doing unnecessary moves - doublures - or irrelevant moves.

Finding several times the same information (redundancy) is not a real problem in the field of information retrieval, because it is easy to automatically eliminate doublures. But in the domain of hypermedia usage, redundancy has been considered as a problem for the user. It points to uneconomical behavior. The famous 'looping' phenomenon coined by Foss (1988), describing the fact that the user opens more than three times the same node, is frequently interpreted as if the user is getting lost, does not reach his/her goals, or doesn't understand information that he/she is processing. Thus, it is interesting to examine the fate of elimination of redundancy, and we will include in this study redundancy as a third rational criterion, and we will call it "economy".

In the domain of information retrieval (see for example Salton & McGill, 1983), perfect search algorithms or indexing techniques do not exist. The evaluation of search algorithms or indexing techniques is done by referring to the recall / precision ratio, such as represented in figure 2 (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). The function in this figure means that there is no simple threshold that can indicate precisely what is a good information retrieval system, but only a recall / precision ratio: It is easy to achieve a high recall at the cost of precision. And also, it is easy to achieve high precision at the cost of recall. The problem for an effective information retrieval system is to find the right balance between recall and precision.

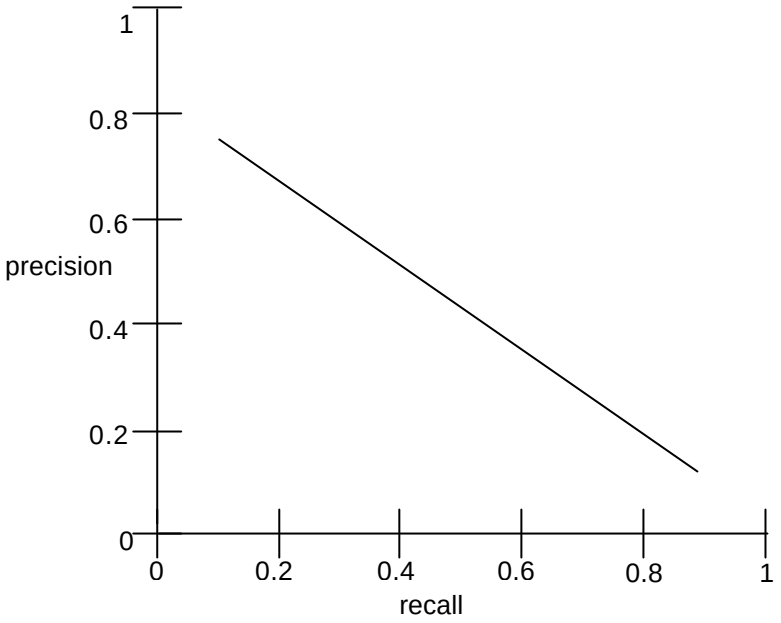

Figure 2. *The classical precision-recall curve*

Recall and precision are influenced by many factors. A first kind of factor concerns the characteristics of information. For example, the prevalence of synonyms in a database tends to decrease the recall performance of retrieval systems (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). Polysemy (the phenomenon of multiple meanings of words) is one factor underlying poor precision. Other factors concern the characteristics of users (see Marchionini, 1995 for a review). For example, in a study on activities of professional on-line searchers, Fidel (1984) found two search styles: 'operational' and 'conceptual'. An operational searcher devotes considerable efforts to manipulate the system and conducts high-precision searches. A conceptual searcher devotes more efforts to the concepts and terminology and develops subsets of results that are then combined in various ways to yield high recall. Saracevic and Kantor (1988) showed that both precision and recall are positively correlated to the information seeker's knowledge level about the database content.

In information retrieval, as we saw in figure 2, general performance of information retrieval systems can be described by a negative correlation between the two performance indicators recall and precision.

However, in other studies - concerning users and not systems - authors found a positive correlation between off-line performance (answering questions) and on-line performance (path economy) (Edwards & Hardman, 1989; Foss, 1988; Rouet, 1990). The quality of answers (off-line performance), for instance, was positively correlated with search efficiency (on-line performance). In a previous experiment (Tricot & Coste, 1995), we showed that like in the field of information retrieval, two rational criteria (recall and economy) were negatively correlated. But we also suggested that we had to distinguish between "getting lost" (bad on-line performance) and "not achieving a goal"(bad off-line performance) and that the relation between these two kinds of performance depends on the task that the user had to achieve. Therefore, the best way to study on-line rational criteria in the field of user's behavior evaluation is to study the correlation between off-line user's performance (e.g. problem solving, questions answering) and on-line user's behavior (in terms of recall, precision, etc.).

In the present study, however, we would like to examine a much more simple question: concerning the behavior of users, are recall, precision and economy negatively correlated as in the domain of information retrieval, or, does the direction of the correlation between recall, precision and economy depend on the kind of task that the user has to achieve? In the study previously mentioned (Tricot & Coste, 1995) a result was obtained opposed to Foss' interpretation. In the Tricot and Coste study, we asked subjects to complete a complex task. Thus it could be that task complexity explains the discrepancy between Tricot and Coste's study and Foss' study. In the present study we examine more systematically whether the relation between recall, precision and economy depends on the complexity of the user task. We operationalize task complexity as the variation in number of targets. More specifically, the number of information nodes in the hypertext that is relevant to the task at hand.

Experiment 1: variation of task difficulty

Method

Subjects
Thirty students of the University of Provence participated in the experiment. None of the subjects took part in clinical psychology or computer science programs in order to control for prior-knowledge effects.

Materials
We presented to subjects a hypertext on a computer with an HyperCard stack describing the profession of psychiatrist, containing 60 cards (including 4 "tools cards": table of contents, alphabetical lexicon, welcome card and headings). Each of the 56 "content cards" contains a 80-words descriptive text. The instruction for the subjects was presented on a sheet of paper. It was mentioned that they had to answer questions (or make inquiries) "with the help of the electronic document describing the profession of psychiatrist". Then they started to read the hypertext.

Tasks
In Task 1, subjects had to answer the three following questions: "What is a neurologist?" "What is obsession?" "What are the moral qualities of a psychiatrist?". Three cards were relevant to answer these questions, one for each question.
In Task 2, subjects had to answer the following question: "Can you explain the difference between neurosis and psychoses using an example?". Nine cards were relevant to answer this question: two definitions and seven examples.
In Task 3, subjects were asked "to make inquiries about the profession of psychiatrist, with the intention of informing young people about this profession". Every "content card" (all 56) was relevant.
We suppose that task 1 to task 3 can be characterized by an increasing difficulty for subjects.

Design

Each experimental task was performed by 10 subjects. Subjects were randomly assigned to conditions (tasks). The independent variable Condition (type of task) was included as a between-subjects variable.
We did not analyze here the subjects' off-line performance. We only examined the way they accessed information, and we used three kinds of measures (i.e. the three rational criteria mentioned above):

- Recall: number of different relevant cards opened by the subject divided by the total number of relevant cards in the system (Swets, 1969). The task is correctly achieved if recall = 1; this applies for each condition.
- Precision (Swets, 1969): number of different irrelevant cards not opened by the subject divided by the number of irrelevant cards in the system.
- Economy[1]: number of different relevant cards opened by the subject divided by the total number of cards opened by the subject. Opening the same cards more than once leads to an increase of number of cards opened, and consequently, to a decrease of economy.

Results

The general results on the dependent measures are shown in Table 1.

means	Task 1 (3 targets)	Task 2 (9 targets)	Task 3 (56 targets)
number of opened cards	40.6	42.2	105.2
recall	1.	.60	.78
precision	.73	.69	-
economy	.12	.23	.46

Table 1. (*mean) subjects' performance as a function of Tasks in Experiment 1*

The average number of cards opened was relatively large. Thus the task of locating information was not trivial, which could mean that the subjects had trouble locating targets or, alternatively, it could also mean that they simply explored the database while performing the task. The number of cards opened varied as a function of condition, that is, the number of information nodes relevant to the question ($F(2,27) = 8.54$, $p < .005$). However, the difference between the number of opened cards in Task 2 and in Task 1 was not significant ($F(1,18) = 0.01$, ns).

There was a significant effect of condition on recall ($F(2,27) = 6.49$, $p < .01$). Pairwise comparisons showed that the mean recall was significantly ($p<.05$) higher in condition 1 than in condition 3, and also significantly higher than in condition 2. These results suggest that the recall performance does not linearly depends on the number of targets. The subjects' recall

[1] It is classic in information retrieval to measure recall as the ratio of relevant document retrieved to the total number of relevant document in the database (that is what we do) and precision as the ratio of relevant document retrieved to the total number of document retrieved (see e.g. Marchionini, 1995, p. 198). We used two different measures of precision (precision and economy) in order to be able to take into account doublures, which are considered as a problem in the field of hypertext (see Foss, 1988) but not in the field of information retrieval.

performance is better when they have to open 56 targets than when they have to open 9 targets, but it is still better when they only have to open 3 targets.

There is no precision index on Task 3 because there is no irrelevant card in this condition. There was no significant task effect on precision for Task 1 and 2 ($F(1,18) = 0.03$, ns).

Finally, there was a significant condition effect on economy ($F(2,27) = 18.4$, $p < .001$). It appeared on the basis of pairwise comparisons that subjects' paths were significantly ($p<.05$) more economical in Task 3 than in the other two tasks. The performance on Task 1 and Task 2 did not differ significantly ($p>.05$).

Summarizing these results: the effects of the three tasks on the three criteria are not congruent.

Correlations

The relations between recall, precision and economy were examined separately for the three tasks.

Task 1 (see Figure 3): (Correct) Recall is independent from precision, because all subjects succeeded (the recall performance does not vary). It is interesting to note that though all subjects have the same recall performance (1), the heterogeneity of precision performance is yet very high (s.d. = 0.34). The correlation between precision and economy was very high ($r= .90$, $p<.001$).

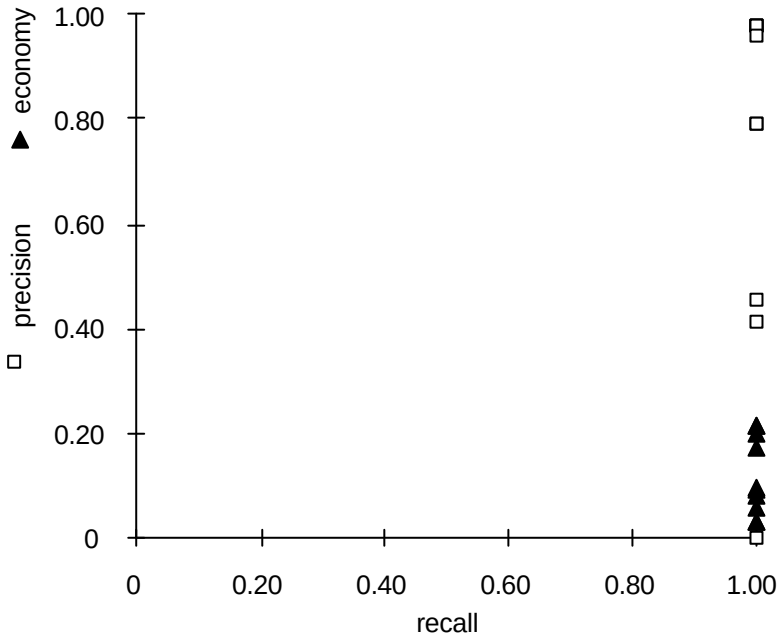

Figure 3. *Scatter diagrams between recall and precision, and between recall and economy for Task 1*

Task 2 (see Figure 4): (Correct) Recall is independent from precision ($r = -.31$; ns) and also of economy ($r =.17$; ns). The correlation between economy and precision is here very high too ($r= .68$, $p<.04$). Note that the recall performance is now heterogeneous (s.d. = 0.25), like the precision performance (s.d. = 0.39). Thus, the recall performance suggests that task 1 is rather easy to complete, and that task 2 is not easy for every one. However, in both cases, precision and economy are completely independent from recall and quite heterogeneous. It seems as if precision and economy are completely unrelated to recall.

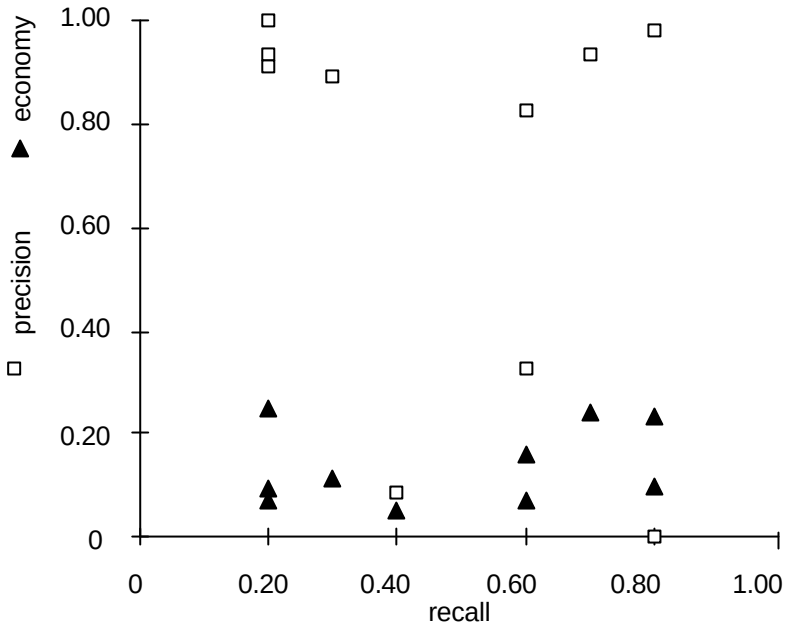

Figure 4. *Scatter diagrams between recall and precision, and between recall and economy for Task 2*

Task 3 (see Figure 5): (Correct) Recall is independent from economy (r =.35; ns). Note that 7 subjects out of 10 have a good recall performance, and 7 subjects out of 10 have quite the same economy performance (around 0.35 on average), but only 4 subjects out of 10 have both good recall and economy performance, near 0.35. Once again, it seems that the same task invokes rather different behavior of subjects, which is difficult to describe with these quantitative criteria.

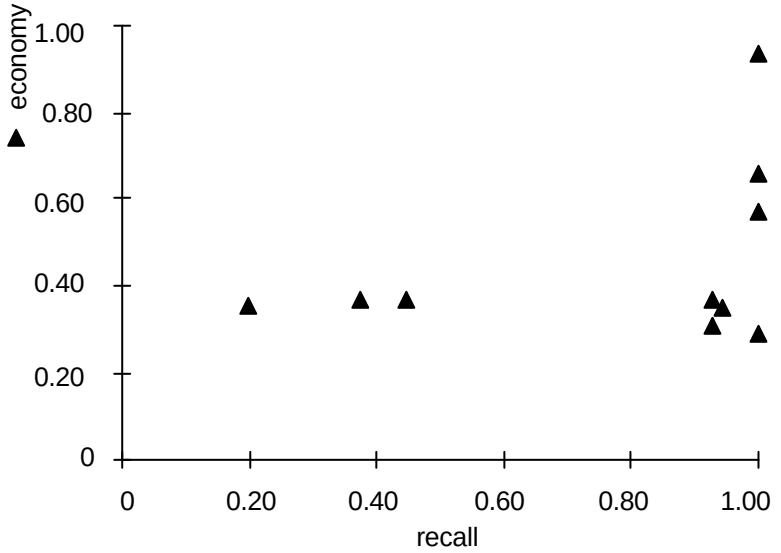

Figure 5. *Scatter diagrams between recall and economy for Task 3*

Discussion

The main question of this experiment was: Does the relevance of rational criteria depend on task difficulty, that is, the number of information nodes relevant to the question?

If we focus on subjects' recall performance, we can say that Task 1 is easier than Task 3. At the same time Task 3 is easier than Task 2. Therefore, this performance does not linearly depend on the number of targets. Precision is quite similar for Task 1 and Task 2. The path economy is better for Task 3 than for Task 1 and Task 2. Consequently, the effects of tasks on the three criteria are not congruent. Also concerning the correlations we saw that recall does not correlate with precision nor with economy (though the latter two correlate).

Therefore we can conclude that: there is no negative correlation nor a positive correlation between recall and the other two rational criteria. Objective task difficulty does not seem to have a simple relation with the three performance criteria. Finally, many subjects do not seem to care for precision or economy. These conclusions lead us to two subsequent, alternative hypotheses:

a) precision and economy are in general irrelevant to this domain of users' behavior;
b) the task representation of users does not under all circumstances include criteria as precision and economy: users do not think that they always have to achieve the task with efficiency; they don't set the goal of accuracy or efficiency automatically.

The second experiment was designed in order to examine these alternative hypotheses.

Experiment 2: giving constraints to subjects

Introduction

In a second experiment, we explicitly asked the subjects to be efficient. We thought that this could be an adequate way to test hypothesis (b) mentioned above. If the goal subjects set for solving a task is important, then we should expect that a stringent instruction to perform efficiently would enhance the relevance of the rational criteria, and consequently would increase the correlations between criteria. In this experiment, conducted with the same materials in the same task conditions, we have introduced a constraint in the instructions: we asked a second group of 20 subjects to reach their goal "by opening as few cards as possible". Only Task 1 and Task 2 were used, because Task 3 is not really interesting for our purpose (with Task 3 we cannot determine precision performance because every card is relevant in this condition).

Method

The method (subjects, hypertext, etc.) and procedure were similar to that of Experiment 1. Ten subjects received task 1 and 10 subjects received task 2, both with a 'constraint'

instruction. Subjects were randomly assigned to conditions (tasks). The data were analyzed in a between-subjects two-by-two factorial design: Condition (Task 1 vs. Task 2) and Constraints ('Free' vs. 'Constraint').

Results

The means on the dependent measures are presented for direct comparison, next to the results of the corresponding condition of the previous experiment in Table 2, here labeled as the 'free' and 'constraint' condition respectively.

There were two significant main effects on recall performance: Task (3,36)= 52.1, p<.0001, and Constraint F(3,36)= 3.76, p<.06. Task 1 is achieved better than Task 2, and the 'free' instruction results in a higher recall than the 'constraint' instruction, but this effect is not significant at a conventional level. The interaction between both variables is not significant (p>.05).

	Task 1 (3 targets)		Task 2 (9 targets)	
means	*free*	*constraint*	*free*	*constraint*
number of opened cards	40.6	16.1	42.2	28.0
recall	1.	.97	.60	.37
precision	.73	.95	.69	.86
economy	.12	.17	.23	.14

Table 2. *(mean) subjects' performance as a function of Tasks in Experiment 1(free) and Experiment 2 (constraint)*

Furthermore , there is a significant effect of constraint on the precision (F(3,36) = 3.88, p<.05). Subjects are more precise in the constraint conditions than in the free conditions. The precision performance is higher under influence of constraints and this holds for both tasks. The effect of the factor Task nor the interaction effect were significant (p>.05).

Finally, there is a significant interaction effect between task and constraint on the economy performance (F(3,36) = 4.75, p<.04). The main effects were not significant (p>.05). These results indicate that subjects are more economical in Task 2 than in Task 1 in the free condition, but they are more economical in Task 1 than in Task 2 in the constraint condition.

Concerning the correlations, we found the following correlations in the constraint condition:

Task 1: The correlations between recall and precision (r = .61, p=.06) and also recall and economy (r = .58, p=.07) are (almost) significant. The correlation between economy and precision is still very high (r= .75, p<.02).

Task 2: recall is still independent from precision (r = .17, ns) but correlated with economy (r = .63, p=.05). Economy is independent from precision (r= .53, ns).

Summarizing , the main results are that task constraints may show positive effects on task performance: though there is with regard to recall no significant effect of constraints, the precision under constraints increases significantly. Also the correlations between recall on the

one hand, and precision and economy at the other increase, though only for the rather simple task (Task 1).

General Discussion

One interesting, first result of Experiment 2 is that constraint does not significantly decrease recall performance. Concerning precision even the opposite is found. These results indicate that subjects can be more precise or rational when asked. Therefore, we may conclude that rational criteria such as precision and economy are (only) relevant when the subject knows the task has to be achieved with efficiency (for example when users have no time to loose, or when information search is expensive, etc.). The results give us also some indications about the relations between recall, precision and economy: apparently the rather simple tasks under conditions of constraints (Experiment 2) show positive correlations, and also constraints lead to an increase in correlation. Thus we find here support of the second hypothesis mentioned above, and at the same time no support of the first one.

The main interest of using rational criteria is that they enable us to compare different tasks and different performance when off-line performance is not measurable or comparable. In the domain of interactive multimedia systems, it is possible to conduct experiments in different contexts with formally comparable results. We mean that when it is difficult to compare off-line performance such as learning, performance. This comparison is based on dependent variables calculated on the basis of number of targets, number of nodes in the database and number of opened nodes. These variables are recall, precision and economy, and they are imported from the domain of information retrieval (the evaluation comprehension, etc. in different systems, it is possible to use and compare recall and precision of retrieval algorithms, evaluation of indexing techniques, etc.).

Task characteristics (i.e., the number of targets) have an effect on users' performance, but the general relation between performance and the number of targets is not easy to describe. Here we found a relation between performance and the proportion of targets that looks like a shaped function. The proportion means the number of targets divided by the number of nodes in the database. It seems that when the proportion of targets is very low (Task 1) or very high (Task 3), the tasks are easier to achieve than when the proportion of targets is medium (Task 2). When the proportion of targets is very low, the task is very easy to do because the subjects have just to match questions and the words in the alphabetical lexicon to find the relevant target. When the proportion of targets is very high, the task is very easy to do because the subjects can just employ a "page turning" strategy, that is, using systematically the « next page » button.

The constraint effect on the precision performance indicates that in the free condition some subjects don't create a task representation in accordance to rational criteria: it seems that subjects don't think that they have to access information efficiently. This is what Waterworth and Chignell (1991) called user's orientation. In the free condition, the representations built up during task 1 and task 2 seem to be based on the combined goal of locating precise information as well as exploring the database. In the free condition, subjects demonstrate a more economical behavior when they have to find nine targets than when they have to find three targets; in the constraint condition we observed the opposite effect, here we found that subjects show a more precise behavior. Thus task effects on users' behavior arise as a function of:

- the number of targets
- the proportion of targets in the database
- the user's interpretation of constraints (in terms of time, efficiency, etc.)

Some subjects seem to create a task representation mainly in terms of contents to process (comprehension) and others in terms of targets to locate. We have indeed observed that some subjects, when they have found targets, also explored the system with curiosity trying to comprehend the information.

If there are task effects, they don't enable us to understand all behavior of subjects. In the free condition, for example, it is very difficult to understand the users' behavior on Task 1, but very easy to understand what happened on Task 3: subjects can achieve Task 3 when they adopt a "page turning" strategy (7 subjects out of 10).

These results show limits and interest of the use of rational variables to evaluate the usage of interactive multimedia systems. They allow us to describe task-characteristical effects from a formal point of view: number or proportions of nodes. Other characteristics such as database structure or subjects' expertise have effects on these rational variables (de Vries & Tricot, 1997).

Finally we want to discuss the relevance of recall, precision and economy in the framework of interactive multimedia systems usage. The rational variables don't allow us to interpret the level of success in terms of exploration, browsing, or discovering relevant information which is not searched. These rational variables only make sense if they are included in a more general analysis that takes into account variables describing the main activity, i.e. comprehension, learning, problem solving, etc. The efficiency of information seeking activities in interactive multimedia is not easily assimilated to the efficiency of information retrieval in a database. We mean that efficiency as it is coined in the information retrieval domain is not an adequate way to measure what people do, search and wait for in a hypertext environment. We have to take into account that there are a number of different kinds of documents, which require many different goals for adequate processing, from the most precise to the most fuzzy goal. If we want to evaluate different systems with the same criteria:

- we must have general and rational criteria such as mentioned above,
- we must be able to categorize tasks, and above all
- we must identify the links between tasks and criteria, that is, we must know for what tasks what criteria are relevant under what circumstances (constraints).

Conclusions

In this study we have shown that the precision criterion is independent of the recall criterion, except with a simple task in conditions of constraints. Therefore, we can conclude that application of this rational criterion (precision) to the analysis of users' behavior is probably irrelevant except when subjects search a relatively small number of targets and are constrained. Then, they become more precise even if their recall performance is not better or worse. But it is worthwhile to note that the relation between these two criteria is positive while it is negative in the field of information retrieval. The relevance of the precision criterion has probably to be studied not by linking it to the recall criterion but by studying it in relation with the effects of different constraints, such as time, number of nodes to open, etc.

We have also shown that the economy criterion (the elimination of redundancy) is frequently positively correlated with the precision criterion. It also appeared that under constraints economy was in general positively correlated with recall, more or less like precision. It seems that economy is a useful criterion when subjects perform their information processing under constraints.

Finally, we have also shown that task difficulty is dependent from the proportion of targets in the database. It is easier to find a small or high proportion of targets in a database than to find an intermediate proportion of targets in a database.

In complex environments, when subjects have to achieve complex goals, the rational approach has limits, and it seems obvious that in that case researchers should try to adopt a more psychological approach if they want to understand subjects' activity (What strategy do users choose ? How do they manage these strategies ? How is cognitive load involved in this management ?). It seems to us that the application of rational criteria from the domain of information retrieval to the evaluation of users' information seeking activity must be refined, in studies linking subjects' on-line and off-line performance.

Acknowledgments

The authors wish to thank Jean-Francois Rouet and Herre van Oostendorp for their help.

References

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Bernstein, M. (1993). Enactment in information farming. *Hypertext'93 Proceedings*, Seattle (pp. 242-249). ACM Press.

Bernstein, M., Joyce, M. & Levine, D. (1992). Contours of constructive hypertexts. In D. Lucarella, J. Nanard, M. Nanard & P. Paolini (Eds.), *ECHT'92, Proceedings of the 4th ACM Conference on Hypertext*, Milano (pp. 161-170). ACM Press.

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science, 45* (1), 12-19.

Card, S.K., Moran, T.P. et Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41* (6), 391-407.

Edwards, D.M. & Hardman, L. (1989). 'Lost in hyperspace': cognitive mapping navigation in a hypertext environment. In R. Mc Aleese (Ed.), *Hypertext: Theory into practice.* (pp. 105-125). Oxford: Intellect Ltd.

Fidel, R. (1984). On-line searching styles: A case-study-based model of searching behavior. *Journal of the American Society for Information Science, 35* (4), 211-221.

Foss, C.L. (1988). Effective browsing in hypertext systems. *RAIO Conference*, "User-oriented content-based text and image handling", Cambridge, MA, March 21-24.

Landauer, T., Egan, D., Remde, J., Lesk, M., Lochbaum, C., & Ketchum, D. (1993). Enhancing the usability of text through computer delivery and formative evaluation : the

SuperBook project. In C. Mc Knight, A. Dillon & J. Richardson (Eds.), *Hypertext. A psychological perspective.* (pp. 71-136). Chichester: Ellis Horwood.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge University Press.

Marchionini, G., & Crane, G. (1994). Evaluating hypermedia and learning: Methods and results from the Perseus Project. *ACM Transactions on Information Systems, 12* (1), 5-34.

Nanard, J., & Nanard, M. (1995). Hypertext design environments and the hypertext design process. *Communications of the ACM, 38* (8), special issue "Designing hypermedia applications", 49-56.

Rouet, J.-F. (1990). Interactive text processing by inexperienced (hyper-) readers. In A. Rizk, N. Streitz & J. André (Eds.), *Hypertext: Concepts, systems and applications*. (pp. 250-260). Proceedings of the European Conference on Hypertext, Versailles. Cambridge University Press.

Rouet, J.-F., & Tricot, A. (1996). Task and activity models in hypertext usage. In H. van Oostendorp & S. de Mul (Ed.), *Cognitive aspects of electronic text processing*. (pp. 239-264). Norwood, NJ: Ablex Publishing.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. II. Users, questions and effectiveness. *Journal of the American Society for Information Science, 39* (3), 177-196.

Su, L. (1994). The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science, 45* (3), 207-213.

Swets, J. (1969). Effectiveness of information retrieval methods. *American Documentation, 20*, 72-89.

Tricot, A., & Coste, J.-P. (1995). Evaluating complex learner-computer interaction: What criteria for what task ? *EARLI'95 Conference*. Nijmegen, Netherlands, August 26-31.

Vries, E. de, & Tricot, A. (1997). Local goals in multimedia use and the study of learner activities through interaction variables. *EARLI'97 Conference*, Athens, Greece, August 26-30,

Waterworth, J.A., & Chignell, M.H. (1991). A model for information exploration. *Hypermedia, 3* (1), 35-58.

Wright, P., & Lickorish, A. (1994). Menus and memory load: navigation strategies in interactive search tasks. *International Journal of Human-Computer Studies, 40*, 965-1008.