Evaluating User Interfaces to Information Retrieval Systems: A Case Study on User Support

Giorgio Brajnik, Stefano Mizzaro, Carlo Tasso

{giorgio|mizzaro|tasso}@dimi.uniud.it
Dipartimento di Matematica e Informatica, University of Udine
Via delle Scienze, 206
Loc. Rizzi – 33100 Udine – ITALY

Abstract

Designing good user interfaces to information retrieval systems is a complex activity. The design space is large and evaluation methodologies that go beyond the classical precision and recall figures are not well established. In this paper we present an evaluation of an intelligent interface that covers also the user-system interaction and measures user's satisfaction. More specifically, we describe an experiment that evaluates: (i) the added value of the semi-automatic query reformulation implemented in a prototype system; (ii) the importance of technical, terminological, and strategic supports and (iii) the best way to provide them. The interpretation of results leads to guidelines for the design of user interfaces to information retrieval systems and to some observations on the evaluation issue.

1. Introduction

Information retrieval (IR) technologies are playing an increasingly important role in more and more applications that range from retrieval of textual information from databases, to navigation in a network of information sources.

They all make the fundamental assumption that the end-user is directly operating an *artificial intermediary*, which enables and actively supports the access to information sources. The user can, and indeed has to, explore the information space, examine selected information items and control the information search process.

In most cases neither the end-user nor the artificial intermediary possess enough knowledge to solve autonomously the information problem. End-users might not be able to cope with complex search environments and situations. Artificial intermediaries cannot assess relevance and utility of retrieved information items. Thus, end-users and artificial intermediaries have to *co-operate* in an interaction dialogue aimed at incrementally solve the user's information problem.

Designing good artificial intermediaries is, however, difficult. First, the space of design alternatives is large and complex. Conceptual models of the functionalities of information intermediaries (e.g. MONSTRAT and MEDIATOR models [Belkin, Brooks and Daniels, 1987; Ingwersen, 1992]) show that a rich set of interrelated functions need to be implemented. Other studies (e.g. [Bates, 1990]) suggest that when users interact with artificial intermediaries the boundary between them might dynamically change (in abstraction level and in machine involvement).

Second, the criteria to be used to assess the quality of artificial intermediaries are not sufficiently well understood and established. Evaluation methodologies adopted for information retrieval systems tend to address a restricted formulation of the problem, often focusing on precision and recall figures only and forgetting to include end-users. Users' information seeking behavior and users' interaction with the artificial intermediary are crucial factors that need to be considered in the evaluation.

In this paper we discuss and evaluate different types of support that an intermediary may provide to its users. We describe an experiment performed for determining importance, effectiveness and best modality for providing technical (i.e. concerned with the system), terminological or strategic help.

The adoption of a taxonomy of user supports and a set of observation variables that covers not only search effectiveness but also user satisfaction enable a deep evaluation of user behavior and the identification of general design guidelines for user interfaces to information retrieval systems.

Obtained results show that all three kinds of support are important and needed in user interfaces to information retrieval systems. Different modalities, though, should be adopted for delivering them.

2. User interfaces to IR systems

A User Interface to an IR System (UIIRS) [Ingwersen, 1992; Marchionini, 1992; Gauch and Smith, 1993] is a front-end program which interacts with the user and controls an underlying information retrieval system accessing information resources. Its main goal is to empower the user with the capability to operate effectively without the need of

SIGIR'96, Zurich, Switzerland©1996 ACM 0-89791-792-8/9608.$3.50

an experienced human intermediary.

The absence from the scenario of a human intermediary has the consequence that users explore autonomously the information space; they don't need to make explicit their problems, and they don't have to cope with possibly difficult interpersonal communication. On the other hand, users might miss the conceptual help and specialized skills that human intermediaries can provide. Users are often overwhelmed by complex IR problems, whose effect is usually mitigated by relying on human intermediaries: the vocabulary problem [Furnas et al., 1987], the anomalous state of knowledge [Belkin, Oddy and Brooks, 1982a; 1982b], the label effect, and the explicitation of visceral and muddled needs [Ingwersen, 1992]. Users might also need technical help on commands, controls and displayed information of the program they are using. Therefore, effective UIIRS have to provide a wide range of technical and conceptual support to users. Furthermore, as suggested by several studies (e.g. [Bates, 1989; 1990; Ingwersen, 1992]) UIIRS should support unstructured user information-seeking behavior, with a dynamically changing man-machine boundary. Last, but not least, UIIRS should implement the functionalities described in conceptual models of intermediaries like MEDIATOR [Ingwersen, 1992] and, ideally, they should be able to engage in complex explanatory activities as analyzed by [Belkin, 1988].

It is thus not surprising that designing UIIRS is a difficult activity. The number and complexity of conceptual and technical issues can easily become overwhelming unless the designer identifies specific levels of analysis and defines specific options to be analyzed, implemented and evaluated (see, for example, [Belkin and Marchetti, 1990]). In this paper we focus on different kinds of support that the UIIRS can provide to the user.

2.1. Supporting users of IR systems

As an evaluation framework, we propose a simple taxonomy (called *support space*) of possible kinds of help that a UIIRS could provide to users during the interaction, possibly adopting different modalities.

According to the *nature* of the support, we distinguish among:

- *technical help*, enabling the user to interact with the UIIRS in an effective and satisfactory way. For example by highlighting the role of a certain control option.
- *conceptual help*, supporting the user in overcoming problems related to the information seeking process. Conceptual help can be distinguished in:
  - *terminological help*, to enrich the vocabulary the user adopts when formulating the problem. For example, to suggests lists of synonymous terms;
  - *strategic help*, to improve the user's effectiveness in conducting a search session. For example, to overcome adverse situations, such as when zero items are retrieved by a query.

Help is provided through a of dialogue, i.e. a specific interaction engaging the user and the UIIRS. There are different *modalities* that can be adopted during such dialogues:

- *contextual* vs. *generic*, according to the context the help depends on. If it depends on the specific user behavior and user situation (e.g. suggesting a synonymous term of a term just added to the query by the user) then it is contextual. It is generic if it refers to general aspects or guidelines (e.g. suggesting that reducing the number of AND-ed concepts is a way to enlarge the number of retrieved items).
- *prompted* vs. *unprompted*, according to the agent (user or UIIRS) that starts the help dialogue. It is prompted if it follows an explicit user request (e.g. user asks for synonymous terms). Unprompted help is given automatically by the UIIRS when certain situations are detected (e.g. users actions are inconsistent with the query, such as reducing synonyms for a query that retrieved zero items).
- *user-* vs. *system-controlled*, according to the agent that controls the evolution of the dialogue. For example, a user-controlled help dialogue may be an interaction where the UIIRS offers to the user a browser on a network of related terms, and the user decides how to navigate it. A system-controlled dialogue would present to the user a list of terms, ask for confirmation or selection, and move on.

2.2. Evaluation of user interfaces

Well-founded studies of behaviors of users interacting with UIIRS could shed light on many issues that are yet to be thoroughly investigated. First, results from experiments could be used to better understand the role of UIIRS and to identify important issues in man-machine interaction aimed at information retrieval and usage. Secondly, diagnostic evaluations are essential in developing design guidelines and in determining most promising design options.

Evaluation of UIIRS shares the difficulties of evaluating effectiveness of information retrieval systems (e.g. dealing with the evasive concept of relevance; coping with the core IR problems mentioned previously; variability of many factors, like systems, databases, intermediaries, etc.) [Tague-Sutcliffe, 1992; Robertson and Hancock-Beaulieu, 1992; Saracevic et al, 1988; Saracevic and Kantor, 1988a; 1988b]. Moreover, it puts in the foreground users, their information problems, and their interaction with the UIIRS. The consequence is the need to deal not only with quantitative performance levels (such as precision and recall), but also with qualitative aspects that are more complex to define, acquire, and codify, like users' perception of obtained results and of the information seeking process.

Furthermore, diagnostic evaluations, whose purpose is to determine relative strengths of alternative methods to interact with the user (i.e. different functionalities or different interaction styles), need to characterize and measure effects on cognitive aspects that may be difficult to bring into focus and to isolate from undesired environmental influences.

Evaluation of user satisfaction with respect to information retrieval systems have been pursued in some research effort [Dalrymple, 1990; Dalrymple and Zweizig, 1992; Bailey and Pearson, 1983; Doll and Torkzadeh, 1988; 1989]. However, often the data acquisition techniques adopted seem not to be the most appropriate. Direct questions (such as "was the system easy to use?") typically lead to positively biased answers; interviews need complex methods for codifying the rich set of information gathered from subjects answers; and "think aloud" techniques don't

guarantee that the verbalization adheres to the real behavior [Nisbett and Wilson, 1977]. Semantic differentials, Likert scales, line length methods, when used appropriately, seem more accurate and effective.

3. The FIRE system

FIRE (*Flexible Information Retrieval Environment*) [Brajnik, Mizzaro and Tasso, 1995] is a prototype of a UIIRS. Its main goal is to emulate some of the functions of a human intermediary by interacting directly with end-users and by supporting them during query reformulation. These capabilities are accomplished through the use of explicit representations of knowledge of intermediary skills and subject domains.

FIRE enables the user to enter an information problem, to retrieve documents, and to read and classify them. When not satisfied with the results, the user can start a semi-automatic query reformulation process. On the basis of the current representation of the information problem, the query reformulation capability selects from an appropriate knowledge base a set of alternative or additional terms, proposes them to the user for confirmation, modifies the problem description and performs searches in the database. In this way, FIRE provides terminological support in a contextual, prompted, and system-controlled modality.

FIRE includes the following four main subsystems:

- The *Information Retrieval Subsystem*, devoted to storing and accessing the documents of the databases, implemented with of a boolean system. Currently, this subsystem manages three bibliographic databases whose size ranges from 5,000 to 20,000 document descriptors.
- The *Information Retrieval Expert Subsystem*, a knowledge-based subsystem devoted to propose relevant terms to the user during the semi-automatic query reformulation. It exploits two knowledge bases:
  - The *Domain Specific Knowledge Base*, devoted to representing the terminological knowledge of the domain considered in the currently accessed database. Terminological knowledge is organized in a semantic network, whose arcs represent relationships utilized in a typical thesaurus, such as 'broader term', 'narrower term', 'related term', and nodes represent terms of the domain, together with additional information, such as posting count and a 'controlled/non controlled' flag. This knowledge base, together with stemming algorithms, is the source of the terms proposed to the user.
  - The domain-independent *Expert Knowledge Base*, that provides the criteria for selecting the terms to be offered to the user during the automatic query reformulation. It is constituted by: *tactics* [Bates, 1979], i. e. elementary operations utilized to appropriately modify a single aspect of a query; *plans*, i. e. sequences of one or more tactics exploited for performing more extensive modifications to a query; and *preference criteria*, used to identify the order to be used for processing the various terms of the query during the reformulation.
- The *User Interface*, the graphical user interface of FIRE, that allows the user (i) to provide to the system a representation of the information need, (ii) to display the content of the documents, (iii) to select terms (among those proposed by the system) to be inserted in the query for reformulation, and (iv) to classify retrieved documents into (built-in or user-defined) categories such as 'useful', 'not useful', 'relevant', 'not relevant'.

Figure 1 presents two windows of the user interface. The background window is the main window, in which the user enters a boolean query and constraints on the search, i.e. the number of desired documents and the search objectives (namely, high-recall or high-precision). The boolean query is constituted by *facets* (i.e. disjunctions of terms) which are logically AND-ed. Each term is shown with related information: its posting count, a controlled-term flag, a (user specified) degree of interest. Buttons are available to modify facets, terms and their attributes. The buttons on top of the window are used to start the reformulation process, to directly search the database, and to classify the retrieved documents. On the right hand side of the main window, mouse sensitive titles (and possibly contents) of retrieved documents are shown. The foreground window of Figure 1 shows a list of terms obtained from 'NEURAL NETWORKS' via a similarity search over the controlled terms of the Domain Specific Knowledge Base; they are the results of the application of a specific tactic. Terms accepted by the user (shown in boldface) are automatically added to the query in the appropriate facet. This kind of terminological support is system-controlled, but the user can get the control of the interaction by clicking on the "Suspend" button, and perform some other activity, like to read retrieved documents or to manually modify the query.

4. Evaluating user supports

The experiment has been performed to investigate the following issues: (i) determining the added value of the automatic query reformulation capability of FIRE (hereinafter AR), and (ii) evaluating the importance of different kinds of support along with the modalities for providing them. The former objective is motivated by the need of an effective method to determine improvements in the development of a UIIRS. The latter objective stems from the need to evaluate different design options concerning not only the terminological support but also the more general conceptual support.

4.1. Experimental design

The experiment was designed and executed together with a team of psychologists. According to [Robertson and Hancock-Beaulieu, 1992], it can be classified as:

- A *laboratory* experiment (as opposed to an operational one), in which induced information needs were used and the bibliographic database (20,000 items regarding artificial intelligence derived from INSPEC) is smaller than real collections.
- A *diagnostic* experiment (as opposed to a black-box one), aimed at obtaining useful information to guide the design of UIIRS, and not only at evaluating system performance.
- An experiment in which both *quantitative* and

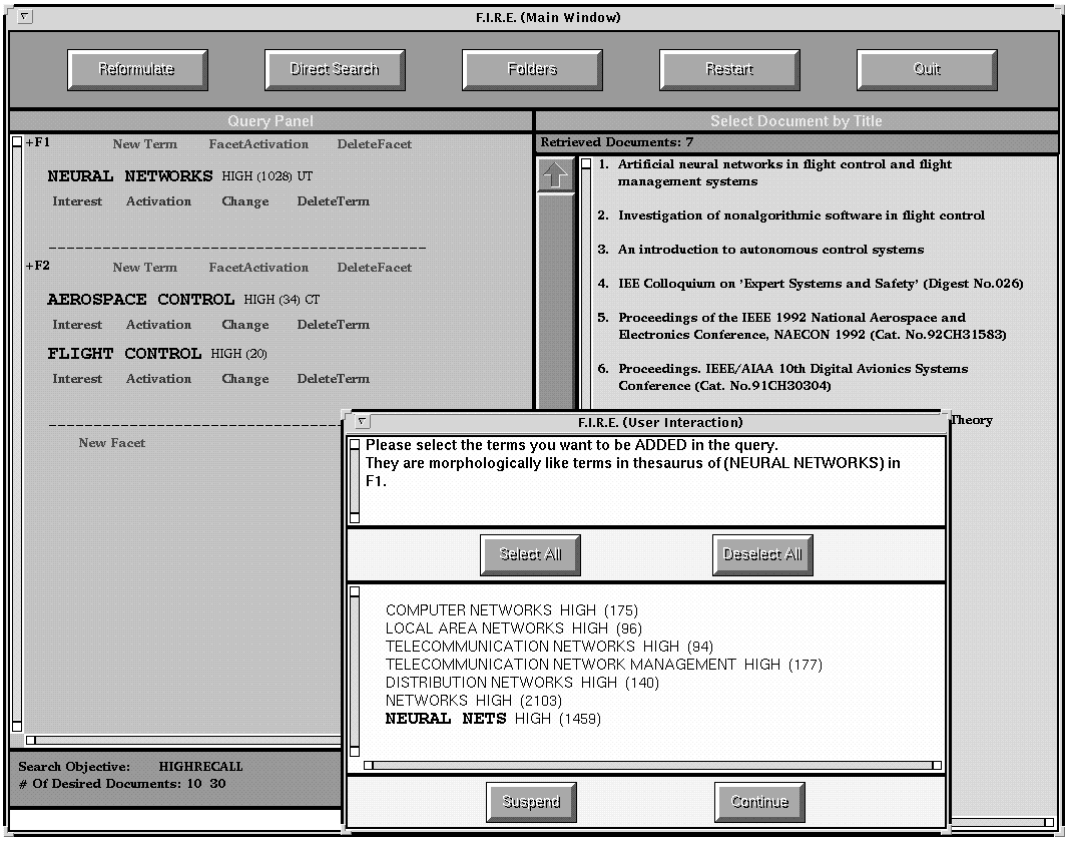

Figure 1. FIRE Main Window.

qualitative measurement methods have been used, in order to take into account search performance, user satisfaction and user behavior (what users ask, how often, which strategies they use, etc.).

4.1.1. Subjects groups

The subjects involved in the experiment were forty-five 3rd and 4th grade computer science university students. They had limited knowledge of IR systems, of IR techniques, and of the specific domain of the collection; none had been previously exposed to FIRE. After some training, each subject performed two search sessions: the first using FIRE where AR had been disabled, and the second using the complete system. Subjects were randomly divided into three groups on the basis of the kind of support received in the first session, as illustrated in Table 1, where:

- *TH* refers to a support provided through a printed thesaurus (obtained from the Domain Specific Knowledge Base) that the user could manually consult;
- *TE* refers to a human expert providing only terminological help in a contextual prompted modality;
- *SE* refers to a human expert providing only strategic help in a generic prompted modality;
- *FIRE–* stands for FIRE without AR, and *FIRE+* for the complete system.

In the two sessions, subjects were given (in random order) two information problems (labelled A and B, see Figures 2 and 3). Each problem includes a *topic*, which identifies the area of the search (e.g., "The role of human factors in interfaces for monitoring systems") and a *task*, which defines the criteria to be used for selecting useful documents (e.g., "prepare a seminar in two days"). The information problems were designed in order not to retrieve too many documents: in the database there were less than 20 useful documents for each of the two information problems.

	Group TH	Group TE	Group SE
Session I	TH + FIRE–	TE + FIRE–	SE + FIRE–
Session II	FIRE+	FIRE+	FIRE+

Table 1: Organization of the subject groups for the experiment.

4.1.2. Independent and dependent variables

Two *independent variables* are used to create the different scenarios: the *system type* (with or without AR) and the *support type* (TH, TE or SE).

The *dependent variables* are chosen to provide three categories of information: about user satisfaction, performance and user behavior.

User satisfaction is measured through questionnaires in which the user evaluates: the quality of the retrieved information; the effectiveness of the system; the

Imagine you are working for a firm dealing with nuclear reactors. The R&D lab needs literature information on the topic:

Fault diagnosis of cooling systems in fission reactors

Your manager asks you to get documentation on existing applications or validated prototypes, in order to see how the problem has been faced in realistic settings.
Based on your previous experience, you guess that the database contains no more than twenty documents treating the topic.

Figure 2. Problem A.

Imagine you are a PhD student. Your advisor asks you to prepare in two days a seminar for a first year course. The title of the seminar is:

The role of human factors in interfaces for monitoring systems

Your advisor suggests that there should be some fifteen papers dealing with the topic. You have no knowledge on this topic and decide to use FIRE for finding a bibliography. Since you only have a short time, you will read only the most general and introductory documents, discarding conference proceedings and other too specific documents.

Figure 3. Problem B.

complexity of the information problems; the received support; the influence of the reformulation in understanding and applying strategic knowledge; the relative strengths of the sources of support (thesaurus, human expert, or AR).

Performance indexes include search effectiveness, evaluated in terms of precision and recall with respect to topicality and utility. They are computed on the basis of two sets of documents. The first set was built a-priori by experimenters, who selected documents considered relevant (useful) for the specific information problem at hand. The second set of documents was determined a-posteriori by all the topicality (utility) judgements expressed by users while performing the experiment. It includes all the documents that received the highest ranks by the users.

Information about the *user behavior*, gathered through automatic logging and via audio-video recording, include: overall number of searches made by the system and directly by the user; overall number of reformulations; number of facets and terms of each query; search objectives defined by the user; number of terms suggested by the system; number of terms accepted by the user; number of significant commands given to the system by the user; duration of the session; the time used by the user to provide the first expression of the information need; overall number of documents found by the system; number and titles of the documents read by the user; user questions and actions.

4.2.3. Forms and procedures

Three questionnaires have been designed for the experiment. The first questionnaire (Q1) has been used to record (via closed questions and Likert-type scales) gender, age, IR experience with/without computers, English language knowledge, and the attitude towards the use of computers in IR. The second questionnaire (Q2) to obtain (via semantic differentials and Likert scales) the user satisfaction with respect to the system, the result of the search, and the kind of available help. In the third questionnaire (Q3) the user provides his evaluation (via line length items) of the overall search activity and the preferred kind of help.

The experiment has been divided into three phases:

1. *Training*: each subject was introduced to the experiment via an informal presentation, and was asked to fill in questionnaire Q1, to read the instructions about the system without reformulation, to observe the experimenter doing a sample search and then to perform a training search on his own (the sample search made by the experimenter and the training search were the same for all users).
2. *Testing*: the subject was told what kind of help was made available and was given a first information problem. S/he had half an hour to perform the search at the end of which questionnaire Q2 had to be filled in. Subsequently, the subject was given the information problem for the second session and the instructions for using the system with AR. After half an hour, questionnaire Q2 had to be filled in again.
3. *Post-Testing*: the subject had to fill in questionnaire Q3.

5. Experimental results

We present now the most important results emerging from the analysis of questionnaires, logs, and videotaped material.[1]

5.1 General results

Statistical analysis of data deriving from Q1 shows that the sample of user population is homogeneous, as are the three experimental groups.

The two different methods for evaluating search effectiveness (i.e. on the basis of reference documents determined a-priori or a-posteriori) show a high correlation. Similarly, there is a strong correlation between performance evaluated on the basis of topicality and utility. We believe this is caused by the database used in the experiment (limited size and homogeneous topics). The data reported in the following refer to the a-priori judgments about utility.

On the other hand, different levels of complexity of the information problems emerge. Independently from the session, problem B systematically leads to poorer performance figures than A. A possible explanation of this result is that problem B is more difficult to conceptualize (i.e. identifying appropriate facets) than problem A. While in A users typically identify three facets, in B they identify only two, which are apparently not sufficient to achieve a good performance level. In fact, there is a significant difference (detected by the analysis of variance: $F=4.4; p<.05$) in the mean number of facets between the two problems.

1More detailed observations have been derived from the experimental data, mostly regarding usability of FIRE and its user interface. We concentrate here only on the most general and important aspects of the interaction.

Surprisingly, there is no significant correlation between user perception of the complexity of information problems (obtained from Q2 via three items of the semantic differential) and performance indexes. Furthermore, in session 1 there is no difference between the groups in the perceived difficulty of finding keywords, of including them in the appropriate facet, and of identifying good search strategies. Users perceive a higher difficulty in identifying good strategies for problem B ($F=3.1, p=.08$). In addition, the frequency of help requests is not affected by the problem.

User satisfaction has been determined via Q2 on the basis of four variables: satisfaction for the retrieved information, for the used system, for the search process and for the received help. Very often these four variables are strongly correlated; in such cases we refer to satisfaction in general. During the first session the information problem does not affect user satisfaction, as there is no correlation between performance level and satisfaction. During session 2, however, there is a strong dependence of satisfaction upon the problem: subjects who solved problem A are more satisfied than those solving B (correlation between performance and satisfaction on the system is $0.35, p<.05$; between performance and satisfaction on retrieved information is $0.5, p<.05$). Comparison of satisfaction over the two sessions shows that for problem A, FIRE+ leads to higher satisfaction for the retrieved information and for the search process ($F=7.9, p<.01$; $F=5.3, p<.05$ respectively). For problem B, even though the performance level does not change, user satisfaction decreases ($F>7.3, p<.01$).

In terms of the preferred system for future searches, subjects that used FIRE+ for problem A, prefer it ($F=6.4$, $p<.05$). Independently from the problem, the most autonomous users (which posed few requests) and those in TE believe that FIRE+ gave a determinant help in solving the problem ($F=3.4$, $p<.05$).

5.2 Added value of automatic reformulation

AR leads to a slight (statistically not significant) improvement of performance, with different effects on different groups: TE seems to improve the performance, while SE seem to worsen it.

Tables 2 and 3 report median and interquartile range of the distribution of precision and recall for each information problem. Means are represented in graphical form in Figures 4 and 5.

A possible interpretation is a learning effect: during the first session users in the TE group "learn" how to use effectively the terminological help, which is later exploited using FIRE+. Members of group SE, on the other hand, have not been exposed to any terminological help and are left on their own.

In session 1 there is no correlation between the average number of terms per query and performance. Such a correlation becomes positive, though, in session 2 for problem A ($r=0.5, p<.05$). This suggests that when users conceptualized appropriately the problem (as occurred for problem A), then AR had a synergetic effect. In more difficult problems (like B), where users may fail to achieve a good conceptualization, AR does not help them.

Prec.		Session I		Session II	
		Median	IQR	Median	IQR
A	TH	43	38.5	30	32.8
	TE	31	29.7	40	22
	SE	46.5	31.5	33	20.5
	all	40	34.5	35	31.5
B	TH	17	4.2	33	35
	TE	10	17	17	14.8
	SE	17	20	0	5
	all	17	20	17	35

Table 2: Distribution of precision (percentage).

Rec.		Session I		Session II	
		Median	IQR	Median	IQR
A	TH	50	41.5	33.5	41.5
	TE	33.5	25.2	34	25.5
	SE	42	24.5	34	25
	all	34	33	34	33
B	TH	17	4.2	17	17
	TE	17	17	17	4.2
	SE	17	17	0	4.2
	all	17	17	17	17

Table 3: Distribution of recall (percentage).

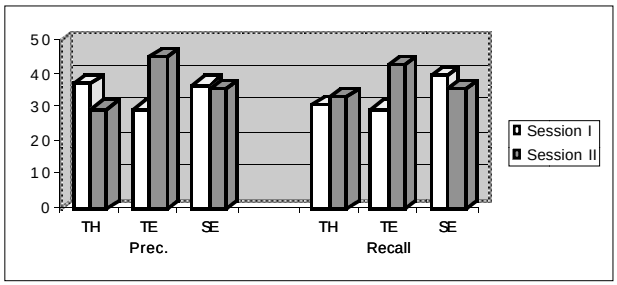

Figure 4. Graphical representation of performance (mean) for information problem A.

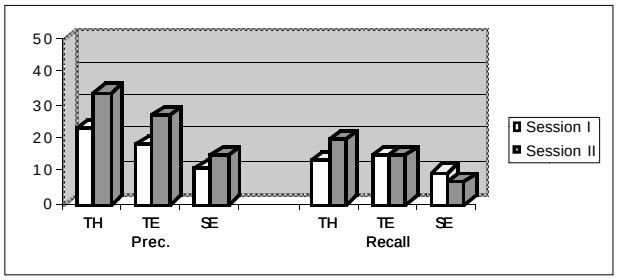

Figure 5. Graphical representation of performance (mean) for information problem B.

5.3 Importance of and modalities for kinds of supports

Technical support is either requested and contextual or it is implicitly needed.

1. 40% of the total number of requests (115) concern technical help. All are contextual.
2. There are many situations where users don't utilize appropriately FIRE commands and controls. Often such situations could be automatically detected and an appropriate help could be automatically given. For example, sometimes they submitted the same query twice consecutively, without changing anything.
3. Users tend not to utilize all the commands/controls that are available. For example, users seldom changed the search objective from its default value (high recall).
4. Independently from the achieved level of performance, users had problems in using the graphically restricted boolean operators. And, as discussed above, these are unperceived problems.

Terminological support is explicitly requested in contextual form. In fact, the majority of help requests (58% of 115) concern specific terms to include in the query.

Strategic help is not requested but it is needed. (In the following discussion, unless specified, no significant difference emerges between experimental groups.)

1. Strategic help is seldom requested (only in 3 cases out of 115). Users tend not to perceive strategic problems in the information seeking process (in session 1, the number of terminological requests posed by groups TE and SE is significantly larger than strategic requests χ^2=13.63; p<.001).
2. Users get stuck in adverse situations. These situations can be classified as: "no-hits", with queries retrieving no item; "too-many-hits", retrieving a number of items much greater than the number of relevant documents; "anomalous queries", like queries including overlapping facets.
3. Enlarging the set of retrieved items is particularly difficult for less successful users. Table 5 shows that "no-hits" situations are much more common for the least successful users (defined as the bottom five subjects ranked on the basis of search effectiveness) than for best ones (the top five ones). "no-hits" situations are also much more common, for worst subjects, than "too-many-hits" situations.
4. Less successful users (defined as above) are unable to diagnose their behavior and show limited flexibility in changing it. They adopt a small set of actions for changing the query (e.g. like truncating a term) and keep applying those actions even if they were ineffective.
5. Both kinds of users perform similar numbers of incoherent or ambiguous actions. An incoherent action is a modification of the query that leads to an increase (respectively, decrease) in the number of retrieved results when the query already retrieved too many (respectively, too few) documents. As seen from Table 6, for both kinds of users the proportion of coherent actions is similar to the sum of incoherent and ambiguous (i.e. that cannot be easily classified as coherent or not) actions. The table, however, shows also that less successful users do a lot more query modifications.

	no-hits	too-many-hits
Best	7	5
Worst	33	9

Table 5: Number of "no-hits" and "too-many-hits" situations for the most successful and least successful subjects.

	+	–	?	Totals
Best	19	6	6	31
Worst	35	14	18	67
Totals	54	20	24	98

Table 6: Number of coherent (+), incoherent (-) and ambiguous (?) query modifications performed by the most successful and least successful subjects.

User-controlled interaction is requested and preferred.

1. Users prefer to retain control during AR. Users often suspended AR (45% of the times) in order to get the control and follow a different route. Furthermore, there is a positive correlation between user satisfaction for the received help and the number of suspensions operations (r=0.47,p<.05).
2. 78% of the users interleaved reformulations, searches and classifications; the remaining ones examined and classified retrieved items at the very end of the sessions. This suggests that users do not want "a system that takes a request in natural language, goes off and searches the information store, and returns to the user the ideal best retrieved set of documents" ([Bates, 1990], p. 575).

6. Discussion and conclusions

This evaluation of FIRE provides useful information for understanding some of the issues underlying the design of effective UIIRS. The rich set of measured variables, the support space and the specific interaction model implemented by the user interface of FIRE constitute an explicit model of the information seeking process carried on through a UIIRS. Such a model supports not only an input/output evaluation of effectiveness of a UIIRS, but also diagnostic evaluations of the sensitivity of performance indexes to different modalities for providing different kinds of support.

More specifically, the following conclusions can be derived:

- Terminological support is important; it is explicitly requested in a contextual form. The automatic reformulation performed by FIRE, while positively affecting users' satisfaction on simpler problems, does not significantly improve the performance.
- Strategic support is likely to characterize good UIIRSs. It should be unprompted, provided under user control, specifically oriented towards conceptualizing the query, enlarging its content and diagnosing ineffective or inconsistent user behaviors.

- Technical support is also needed to augment usability of UIIRSs and improve effectiveness of search. It should be either prompted or unprompted and contextual.
- User-controlled interaction is preferred. The UIIRS should support users in interleaving different activities, in exploring the space of information items (relationships between terms, terms and documents, document content, etc.), and in developing an adequate search strategy.

The interaction model and the experiment we presented neglect important issues. We did not investigate the level of abstraction and involvement of UIIRS as outlined by [Bates, 1990]. These aspects would provide additional information very useful to extend our design guidelines. Secondly, we did not emphasize aspects related with the information presented to the user and the way it is presented. Information spaces could be represented by rich network of relationships between information items; interactive exploration of the space could be based upon a filtering activity performed by the UIIRS to reduce what is presented to the user. Users would keep control of the exploration, but the UIIRS would limit the amount of information presented to them overcoming typical overload effects. We believe that this would be a key factor in providing effective and satisfactory terminological help.

Experiments like the one we just presented provide useful data to better understand general issues and specific problems of a UIIRS. On the basis of the work done, the following hints on the evaluation of UIIRS can be derived:

- A crucial role in the evaluation is played by subjective information and logged or recorded user behavior. The use of semantic differentials, Likert scales and line length techniques has proven to be an effective and accurate method for acquiring, validating and analyzing such important data.
- Experiments of this type are very complex and time and resource consuming: the system being tested cannot be a simple prototype, it must be improved for obtaining reliability, robustness, and efficiency; the experiment itself took about one month of three-person full-time work; and the huge amount of data derived from it (still under analysis) requires sophisticated statistical processing. An inter-disciplinary team is required: computer specialists, psychologists (for designing questionnaires and performing the experiment both without biasing the subjects) and statisticians (for statistical elaborations).
- Is it worth enough to do efforts like this? We would say yes. A lot of guidelines on FIRE and on the UIIRSs in general have been derived. Some of these consequences were foreseeable and foreseen before the experiment (for instance, the AR of FIRE proposes too many terms to the user), but other ones were unexpected (for instance, strategic support should be provided unpromptly). A less rich experiment could have missen to provide this unforeseen information.
- The different level of difficulty of the two information problems, neither observed after the pilot test, nor fully understood after the experiment, is a problem that should be taken into account in similar experiments.

Acknowledgements

We would like to thank Alessandro Bortuzzo, Antonella De Angeli, Danilo Fum and Irina Stultus for helping us in designing and performing the experiment; Pier Giorgio Marchetti that kindly gave us the subset of INSPEC database used in the experiment; and the 45 students involved in the evaluation. Finally, we thank Nick Belkin for very useful comments on a draft of this paper.

References

[Bailey and Pearson, 1983] J. E. Bailey and S. W. Pearson, Development of a tool for measuring and analyzing computer user satisfaction. *Management Science*, 29(5), 1983, pp. 530-545.

[Bates, 1979] M.J. Bates, Information Search Tactics. *Journal of the American Society for Information Science*, July 1979, 1979, pp. 205-214.

[Bates, 1989] M.J. Bates, The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5), 1989, pp. 407–424.

[Bates, 1990] M.J. Bates, Where should the person stop and the information interface start? *Information Processing and Management* 26(5), 1990, pp. 575-591.

[Belkin, 1988] N.J. Belkin, On the nature and function of explanation in intelligent information retrieval. *Proceedings of the ACM SIGIR*, 1988, pp. 135–145.

[Belkin and Marchetti, 1990] N.J. Belkin and P. G. Marchetti, Determining the functionality and features of an intelligent interface to an information retrieval system, *Proceedings of the ACM SIGIR*, 1990, pp. 151–178.

[Belkin, Brooks and Daniels, 1987] N. Belkin, H. Brooks, and P. Daniels, Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies*, 27, 1987, pp. 127–144.

[Belkin, Oddy and Brooks, 1982a] N.J. Belkin, R.N. Oddy, and H. M. Brooks, ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2), 1982, pp. 61–71.

[Belkin, Oddy and Brooks, 1982b] N.J. Belkin, R.N. Oddy, and H.M. Brooks, ASK for Information Retrieval: Part II. Results of a Design Study. *Journal of Documentation*, 38(3), 1982, pp. 145–164.

[Brajnik, Mizzaro and Tasso, 1995] G. Brajnik, S. Mizzaro, and C. Tasso, Interfacce intelligenti a banche di dati bibliografici. In D. Saccà (editor) *Sistemi evoluti per basi di dati*, Franco Angeli, Milano, 1995, pp. 95–128.

[Dalrymple and Zweizig, 1992] P. Dalrymple and D. L., Zweizig, Users' experience of Information Retrieval

systems: an exploration of the relationship between search experience and affective measure. *Library and Information Science Research*, 14, 1992, pp. 167-181.

[Dalrymple, 1990] P.W. Dalrymple, Retrieval by reformulation in two library catalogs: towards a cognitive model of searching behaviour. *Journal of the American Society for Information Science*, 41(4), 1990, pp. 272-281.

[Doll and Torkzadeh, 1988] J.W. Doll and G. Torkzadeh, The measurement of end-user computing satisfaction. *MIS Quarterly*, 12, 1988, pp. 259-274.

[Doll and Torkzadeh, 1989] J.W. Doll and G. Torkzadeh, A discrepancy model of end-user computing involvement. *Management Science*, 35(10), 1989, pp. 1151–1171.

[Furnas et al., 1987] G.W. Furnas, T.K. Landauer, L.M. Gomez, and S.T. Dumais, The Vocabulary Problem in Human-System Communications. *Communications of the ACM*, 30(11), 1987, pp. 964-971.

[Gauch and Smith, 1993] S. Gauch and J.B. Smith, An expert system for automatic query reformulation. *Journal of the American Society for Information Science*, 44(3), 1993, pp. 124–136.

[Ingwersen, 1992] P. Ingwersen, *Information Retrieval Interaction*, Taylor Graham, London, 1992.

[Marchionini, 1992] G. Marchionini, Interfaces for end-users information seeking. *Journal of the American Society for Information Science*, 43(2), 1992, pp. 156–163.

[Nisbett and Wilson, 1977] R.E. Nisbett and T. Wilson, Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 1977, pp. 231–259.

[Robertson and Hancock-Beaulieu, 1992] S.E. Robertson and M. M. Hancock-Beaulieu, On the evaluation of IR systems. *Information Processing and Management*, 28(4), 1992, pp. 457–466

[Saracevic et al., 1988] T. Saracevic, P. Kantor, A. Chamis and D. Trivison, A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, 39(3), 1988, pp. 161–176.

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

[Saracevic and Kantor, 1988b] T. Saracevic and P. Kantor, A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science*, 39(3), 1988, pp. 197–216.

[Tague-Sutcliffe, 1992] J. Tague-Sutcliffe, The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28(4), 1992, pp. 467–490.