Evaluation methodologies and user involvement in user modeling and adaptive systems

(Article begins on next page)

UNIVERSITA' DEGLI STUDI DI TORINO

FACOLTA' DI LETTERE E FILOSOFIA

DOTTORATO IN SCIENZE E PROGETTO DELLA COMUNICAZIONE

XV CICLO

CRISTINA GENA

EVALUATION METHODOLOGIES AND USER INVOLVEMENT IN USER MODELING AND ADAPTIVE SYSTEMS

RELATORE
PROF. LUCA CONSOLE

Dedicated to my parents,

Mirna & Cesare

Introduction

This thesis faces with the issue of evaluation and user involvement in development and testing of user modeling and adaptive systems. Quantitative and qualitative testing methodologies are here described, moving from the common and most used techniques to those ones less explored but indubitably fruitful to the improvement of such systems.

Indeed, in the last few years the international user modeling and adaptive systems community has underlined the importance of evaluation for a more user-centered approach to these systems. The final goal is moving the application field of user-adapted systems from the research labs to real field usage. Also in HCI and information retrieval communities, which are directly connected with the user modeling and adaptive systems community, the role of evaluation and testing with real users is largely carried out in every design phase and significant results are reported. On the contrary, in user-adapted systems community evaluation and user involvement are not so frequent. Indeed, the goal of the community is now reaching more rigorous levels and making the testing a common practice in the development of such systems.

My interest in evaluation started during the first stages of my research activity at the Intelligent User Interface Group of the Computer Science Department of the Turin University, where I carried out my first empirical evaluations of user-adapted systems. Then, after the permanence at University of California at Irvine, where I participated in the design and the running of an empirical evaluation concerning information visualization systems, I focused my research activity and my thesis on the problem of evaluation in user-adapted systems and its peculiarity and future directions.

The thesis is organized as follows:

o Chapter 1 describes methodologies commonly exploited in empirical evaluation and user-centered approach and the inferential statistics suggested to correctly analyse the data. Then, I face with the sampling problem and the methodologies exploited in the evaluation of the selection process. Finally, I illustrate the state of the art and the future directions of user modeling and user adapted systems.

o Chapter 2 describes qualitative research and evaluation methodologies deriving from social science such as ethnography and Grounded Theory and from experimental psychology, such as observational studies. Connected to this last point is the description of the information visualization experiment carried out at the University of California, Irvine. Finally, I illustrate the Dourish's theory of *embodied interaction* and its possible applications in user-adapted systems.

o Chapter 3 starts with a proposal of evaluation methodologies according to the three different tasks in which the development of an adaptive system can be divided and then describes four different evaluations I carried out during these last two years.

1. Evaluations in user modeling and adaptive systems

The empirical evaluation of user modeling[1] and adaptive system[2] is a fundamental stage in their development and it should become a common practice. As the application field of these systems moves from the research lab to real field usage, the evaluation of the real user-system interaction becomes crucial.

The adaptations generated by user modeling techniques often pretend to improve the user-system interaction. Since most of the times the exploitation of such techniques makes the system more complex, slower and buggy, it should be evaluated whether the adaptivity really improves the system and whether the user really prefers the adaptive version of the system. Moreover, a user of adaptive system has more expectations and therefore she is more frustrated when the application does not work as she expected. Empirical evaluations have shown that the users have often problems with the adaptive features of a system and thus they avoid using them. Therefore, the challenge of adaptive systems becomes demonstrating that their exploitation can improve the interaction by testing the utility of the adaptations choices.

Another good reason for a deeper user involvement is the closeness with human computer interaction techniques focused on user involvement. As HCI systems, also the UM and adaptive systems should adopt a user centered approach because the users are both the main source of information and the main target of the application. In fact, some researcher [Benyon, 1993] proposed the exploitation of adaptive systems as solution to usability problems.

As in regular interactive systems evaluation, also in case of user modeling and user adapted systems, the evaluation should occur throughout the entire design life cycle and provide feedback for design modifications. In particular, the first evaluation of the system should ideally be performed before any implementation of the system in order to avoid expensive design mistakes.

1.1. HCI-oriented evaluation methodologies

The methodologies for evaluating adaptive system are generally borrowed by the methodologies used in HCI and by those ones exploited for the evaluation of information retrieval systems.

---

[1] A **user model** contains the system's assumptions about all aspects of the user that are deemed relevant for tailoring the dialog behavior of the system to the user. A user modeling component in an interactive system draws assumptions about the user based on the interaction with her, stores them in an appropriate representation system, infers additional assumptions from initial ones, maintains the consistency in the set of current assumptions, and supplies other system components with assumptions about the user (Kobsa, 1994).

[2] A system is called **adaptive** if it is able to change its own characteristic automatically according to user's needs (Opperman, 1994).

"By adaptive hypermedia system we mean all hypertext and hypermedia system which reflects some features of the user in the user model and apply this model to adapt various visible aspect of the system to the user. (Brusilovsky, 1996)".

One of the possible classifications of methodologies of evaluations in the HCI is [Burattini and Cordeschi, 2001; Dix et al., 1998; Preece et al., 1994]:

- o Collection of user's opinion
- o User observation and monitoring
- o Predictive evaluation → based on experts evaluation
- o Formative evaluation → aimed at checking the first choices and getting clues for revising the design
- o Summative evaluation concerned with the testing of the final system with effective users performing real tasks in their environment. Summative evaluation includes methods as empirical evaluation and interpretative evaluation.

As interpretative evaluation will be deeply analyzed in Chapter 2, in the current section I analyze all the methodologies listed above. In section 2 I describe the evaluation of the selection process in adaptive systems, in section 3 I deal with the sampling problem in social science, in section 4 I make an excursus through current state of the art in the evaluation of user-adapted systems, and finally in Section 5 I will discuss some future direction.

1.1.1. Collection of user's opinion

The collection of user's opinion, also known as *query technique,* is a method that can be used to elicit details of the user's point of view of a system. These techniques embody the philosophy of "asking the user" [Dix et al., 1998] and can reveal issues not considered by the designer. They are simple and cheap and can offer supplementary information to other methods.

User's opinion can be collected by means of:

- o Interviews, which are used to collect self-reported experiences, opinions, preferences and behavioral motivations. When used in conjunction with observation they can be useful to clarify events.

   Interviews can be:

   - ▪ structured → where the same set of questions is made in the same sequence to every interviewee. The structured interview is easy to carry out and easy to analyze.
   - ▪ semi-structured → where the interviewer has to follow a script reporting the topics of the interview. The questions and their order are non-fixed. The semi-structured interview is more adaptable to the context and to the unexpected answers.

- unstructured → where the form and the content are not fixed and can vary from subject to subject. The unstructured interview is totally flexible and the only constraint is to talk about the topics relevant to the goal of the interview.

o <u>Questionnaires</u> → where both the questions and the answers are fixed. The questionnaire can be filled by the user or read by the interviewer (e.g., face to face interviews, telephonic interviews, etc). However, in the latter case the influence of the interviewer in the questions completion has to be strictly controlled in order to avoid possible interferences.

The questionnaire is less flexible than the interview but it is timesaving since allows collecting opinions from more users in a shorter period of time. In the self-filled questionnaire where the interviewer is not present (such as postal questionnaire, Web-site forms, etc) is difficult to control the correctness of the sampling process (a lot of subjects do not reply, the random selection is not guaranteed, etc).

The questions included in a questionnaire can be [Dix et al., 1998]:

- general → to collect general information such as age, sex, occupation, etc;
- open ended → to ask the user to provide his own opinion;
- scalar → to ask the user a judgment on a numeric scale (e.g., Likert scale);
- multi-choice → to offer a choice of explicit responses;
- ranked → to order the items in a list useful to indicate a user's preference.

o <u>Focus group</u> → a focus group is a structured discussion about specific topics moderated by a trained group leader [Dumas & Redish, 1999]. A typical focus group session includes from 8 to 12 people and lasts for two hours. The people are carefully chosen to represent the potential users of the product. If the potential audience is large, several focus group sessions can be performed. Focus groups are excellent ways to probe users' attitudes, beliefs and desires. They do not provide information about what users actually do with software.

o <u>Having user evaluating items</u> → this methodology is often used to acquire user interest profile in collaborative[3] and feature-based[4] filtering. On the one hand the system can use the user's rating to generate recommendation based on the

---

[3] **Collaborative filtering systems** adapt themselves to the individual user on the basis of her "interest neighbors", users who show similar interaction behavior, similar interest, etc. For details see [Malone et al., 1987], [Sharnadand & Maes, 1995].
[4] **Feature-based systems** acquire models of user interest by analyzing the features of the objects in which the user has expressed an interest. Such a profile can be regarded as a user model that allows for personalized recommendation or other kind of filtering (Kobsa et al., 2001, p. 27).

choices of similar neighbors or to produce feature-based recommendation. On the other hand, users can evaluate the items suggested by the system and then the collected data can be used to evaluate the effectiveness of the selection process (see Sections 3.3 and 3.5 for an application of this approach).

An alternative to the above solutions can be the exploitation of existing surveys about the target population such as psycho-graphic surveys (see the Italian survey called Eurisko described in Sections 3.3 and 3.5), Web-users research such as the Italian Audinet, past evaluation of the system or evaluation of similar systems, etc.

1.1.2. User observation

This family of methods is based on user observation. They can be carried out with or without predetermined tasks. The drawbacks of user observation are the possible interferences that can be

- originating from the user → the expectation leads to improvement (Hawthorne effect);
- originating from the observer → the experimenter can influence the user's performance (Rosenthal effect, Pygmalion and Golem effects).

o <u>Think aloud protocols</u> → having users thinking out loud when they are performing a task is called having the user *giving a verbal* or *think-aloud protocols* [Dumas & Redish, 1999]. Since the early 1980s these protocols have been used very successfully in understanding user's problems. People are asked to think out loud while they work so the evaluators can hear and record their reactions. This method *i)* is simple to carry out, *ii)* can lead to spontaneous interpretation of events and *iii)* can be associated to other types of evaluations (e.g., a usability test). However, is not natural since people do not normally think out loud while they work. In addition, studies showed that participants could take longer to perform tasks ant they can be less flexible at solving problems. This method is also applied in co-operative evaluation (see 2.1.2.2).

o <u>Observing users in context</u> → these methods are aimed at learning from the direct or indirect users' observation in the context of real usage by interviewing and observing them at work, when they do their own works. These methods are most reliable and precise but they are also very expensive. They are related to the ethnographic investigations that will be discussed in detail in 2.1.

o <u>Logging use</u> → can be considered a kind of indirect observation and consists, for instance, in the analysis of Web server log files that register the actions of every users of the Web site. This information can be used to improve the Web site design and to analyze possible usage patterns extracted by the application of data mining[5] techniques to the collected data.

1.1.2.1. <u>**Task analysis and cognitive and socio-technical models**</u>

Task analysis methods are based on the analytic de-composition and the re-composition of user's actions and user's cognitive processes to complete tasks. "*Task analysis* is the process of analyzing the way people perform their job: the thing they do, the things they act on and the things they need to know. (…) Task analysis is about existing system and procedures; its main tools are those of observation in various forms [Dix et al, 1993, pp. 260-261]".

In most tasks, analysis tasks are decomposed in sub-tasks. *Hierarchical Task Analysis* (*HTA*) [Diaper, 1989] uses this approach and decomposes tasks in a hierarchy of tasks and subtasks, and exploits plans to describe order and conditions of subtasks. For instance, if a subject wants to send an e-mail she as to perform a set of tasks in sequential order: 1) open the e-mail client, 2) create a new e-mail, 3) write the e-mail. Task 2 can be decomposed in 2a) click on the File menu, 2b) click on New-Email Messages and so on.

Task analysis can be also *knowledge- based* (all objects and actions involved are listed and taxonomies about them are constructed) or *entity-relationship-based* (including the actions performed and subjects performing them).

The source of a data collection can be direct or indirect observation, documentation, interviews, etc…

The understanding of the internal cognitive process as a person performs a task is also the purpose of goal-oriented cognitive models. In HCI these models "claim to have some representation of users as they interact with an interface; that is they model some aspect of the user's understanding, knowledge, intentions or processing [Dix et al, 1993, pp. 230]". In particular *performance models* "not only describe what the necessary behavior sequences are but usually describe both what the user needs to know and how this is employed in actual task execution. [Simon in Dix et al, 1993, p. 230]". These models reflect the analogy between computation and cognitive psychology that can have some disadvantages (see the Suchman's discussion in 2.5.2.1) and some advantages (for instance, it makes the human-system matching easier).

An example of a goal-oriented cognitive model is the *GOMS* model (acronym of Goals, Operators, Methods and Selection [Card et al, 1983]). A typical GOMS analysis

---

[5] **Data mining** is a methodology for the extractions of not arbitrary knowledge from data.

decomposes tasks into a sequence of unit tasks, all of which can be further decomposed into basic operators. The original GOMS model has served as basis for much cognitive modeling research in HCI. It was good to describe how expert perform routine tasks.

KLM (Keystroke Level Model) is a low-level GOMS [Card et al, 1983]. KLM makes predictions about user performance at very low level, as simple command sequences, taking no more than 20 seconds.

Socio-technical models, instead, consider social and technical issues and recognize that technology is a part of wider organizational environment. For instance, the USTM/CUSTOM [Macaulay et al., 1990; Kirby, 1991] model focuses on establishing stakeholder requirements. A stakeholder is defined as anyone who is affected by the success or the failure of the system (e.g., who uses the systems, who receive output from it or provide input, etc).

Both goal-oriented cognitive models and socio-technical models can be considered as *generative* models since they typically contribute during the interface design process and can provides a means of combining design specification and evaluation into the same framework [Dix et al., 1998]. Moreover, they could make predictions about the behavior of different kind of users and therefore used in the construction of their user models.

1.1.3. Predictive evaluation

These kinds of method are aimed at making predictions, based on experts' evaluation, about the performance of the interactive systems and preventing errors without performing experimental evaluations.

- o Heuristic evaluation → heuristic is a guideline or a general principle or rule of thumb that can guide a design decision or be used to criticize existent decisions. *Heuristic evaluation* [Nielsen and Molich, 1990] describes a method in which a small set of evaluators examine a user interface and look for problems that violate some of the general principle of good interface design. Nielsen and Molich [1990] recommended using heuristic evaluation with three to five evaluators. The evaluators can be either interface/usability experts or non-experts. Expert in usability engineering are much better at finding problems then software engineers. Usability experts who have experience with the technology they are evaluating are the best.

If the evaluators are not properly usability experts, their task is to verify that HCI principle[6] and HCI guidelines[7] (Nielsen and Mack, 1994; Sheiderman, 1992) have been correctly applied.

In any case to help evaluators at finding usability problems Nielsen and Morich [1990] suggest a list of heuristic that are relates to HCI principle and guidelines:

- Simple and natural dialog
- Speak the user's language
- Minimize the memory load
- Be consistent
- Provide feedback
- Provide clearly marked exits
- Provide shortcuts
- Good error messages
- Prevent errors

o Domain expert appraisals → in some case the presence of domain experts can be beneficial. For instance, in UM and user-adapted systems a domain expert can help defining the dimension of the user model and domain-relevant features. They can also contribute toward the evaluation of correctness on inferences mechanism. For instance, a user-adapted system that suggests TV programs, such as an Electronic Program Guide, can benefit of audience TV experts working in TV advertising that may illustrate habits, behaviors and preferences of homogeneous groups of TV viewers.

This method, as well as cognitive walkthrough [see 1.1.3.1], scenario-based design and prototypes can be used to develop *parallel design*: exploring different design alternatives before settling on a single proposal to be further developed. Parallel design is very suitable for systems that have a user model since in this way designers can propose different solutions (what to model) and different interaction strategies (what the user can control). Also the *design rationale[8]* and *design space analysis[9]* can be helpful in context of exploring and reasoning among different design alternatives. For details about design

---

[6] A **principle** is a very broad statement that is usually based on research about how people learn and work. For instance, *be consistent in your choices of words, formats, graphics and procedures.* (Dumas & Redish, 1999)

[7] **Guidelines** are more specific goals that HCI specialists distill from the principles for different users, different environment and different technologies. . For instance, be *consistent in the way you have users leave every menu.* (Dumas & Redish, 1999)

[8] *Design rationale* "is the information that explains why a computer systems is the way it is, including its structural or architectural description and its functional or behavioral description [Dix et al, 1993, p. 212]".

[9] *Design space analysis* is an "approach to design that encourages the designer to explore alternative design solution" [Preece et al. 1994].

rationale see [Lee and Lai, 1991; Rittel and Webber, 1973], while for design space analysis see [Bellotti and MacLean].

1.1.3.1. Cognitive walkthroughs

Walkthroughs are a peer-group review of a technical product. They can be used to review specifications, design or programming code. During a walkthroughs the team of people who are developing a software "walk through" the specifications or the programming code one step at time looking for errors or inconsistencies. For more details see [Dumas & Redish, 1999].

The cognitive walkthrough [Lewis & Polson, 1990] is a variation of the walkthrough designed to evaluate the usability of the user interface. The focus of the cognitive walkthrough is learning though exploration, since "experience shows that many users prefer to learn how to use a system by exploring its functionality hands on, and not after sufficient training or examination of a user's manual [Dix et al, 1998, p. 409]."

On the basis of usage scenarios, step-by-step tasks are selected and then performed. Assuming that a user learns about an interface by the exploration, the evaluator has to answer a set of questions about each of the decisions the users must make as they use the interface (for instance, the ease in identifying the consequences of an actions, the evaluation of progress toward a goal, etc). Every non-corrected answer is inserted in a list of detected problems.

Even if this method can results useful, other methods, such as the heuristic evaluation are more effective at finding usability problems [Dumas & Redish, 1999, p. 68].

Walkthrough can also be performed after the task (*post-task walkthrough*) [Dix et al., 1998]. The subjects are asked to reflect back after the event and comment to their actions.

1.1.4. Formative evaluation

Formative evaluation is aimed at checking the first design choices and getting clues for revising the design. Formative evaluation can be performed without a running system.

- o Mock-ups → mock-ups are models of the final artifacts. They are usually exploited for verifying the perceptive and physical features of the final artifacts such as weight, dimension, etc. They are related to the *look & feel* (the sensorial experience of the user) of the final product.
- o Wizard of Oz simulation → is a lab simulation where the experimenter (*the wizard*) is hidden and acts on behalf of system. The user interacts with the emulated system without being aware of the trick.

o   <u>Scenario- based design</u> → this is a method used to describe existing activities or to foresee new activities. A scenario is aimed at illustrating a usage situation by showing step-by-step the user's actions. It can be represented by textual descriptions, images, videos and it can be employed in different design phases. This method can be used to organize the data mined during the observation, to imagine the features of a new system, to the parallel design of alternative prototypical solution, during the empirical evaluation.

o   <u>Prototypes</u> → prototypes, which are artifacts that simulate or animate some but not all features of the intended system [Dix et al., 1998], can be divided in two main categories: *static, paper-based prototypes* and *interactive, software-based prototypes.*

Static, paper-based prototypes are generally screen images (*screenplay*) on paper of what an interfaces looks like. A screenplay can be effective at solving problems such as developing a menu hierarchy the user can understand. If the paper prototypes describe the sequence of actions illustrating the system usage in a narrative way, the prototype is a *storyboard*. The storyboard can also be implemented on a PC as sequence of different screenshots.

Interactive, software-based prototypes can be realized with specific software that make possible to simulate the look and feel of a software user interface. The software prototypes can be: *horizontal* when they contain a shallow layer of the whole surface of the user interface; *vertical* when they include a small number of deep paths through the interface, but do not include any part of the remaining paths; *scenario-based* when they fully implement some important tasks that cut through the functionality of the prototype. There are now plenty of prototyping tools allowing the rapid development of simulation prototypes. These simulations are aimed at providing a quick development process for a very wide range of small but interactive applications (for instance, HyperCard for Macintosh).

Software prototypes can be further divided in [Dix et al., 1998]:

> o   *throw-away,* when the prototype is built and tested and the design knowledge from this exercise is used to built the final product, but the prototype is discarded;
>
> o   *incremental,* the final product is built as separate components and each subsequent release include one more component. For each component a prototype is realized;
>
> o   *evolutionary,* here the prototype is not discarded and serves as basis for the next iteration of design;

Usability test of prototypes are becoming very common because they allow designers to make changes before is too late. A study [Nielsen, 1990] showed that the interactive prototypes are more effective at detecting global problems then paper based ones. However, prototyping tools are best used to explore alternative concepts and they cannot be considered as a finished products (e.g., they do not have the response time of the real software and thus the estimate of user's reaction can be not precise).

1.1.5. Summative methods: the empirical evaluation

Before introducing the main issue of this section, the empirical evaluation, I want to say something about *usability engineering* and *usability testing* since these are summative methodology that can be applied also in case of user-adapted interfaces.

Usability engineering techniques derives from *software engineering.* Software engineering (starting in the 1960s) is a large subdiscipline within computer science that addresses the management and technical issues of the development of software systems [Dix et al., 1998] and lent some principle to the evaluations carried out in HCI. Indeed, one of the cornerstones of software engineering is the *software life cycle* that describes the activities that take place from the initial concept formation for a software system up until its eventual phasing out and replacement. One of the key point derived from software life cycle is that issue affecting evaluation should be relevant within all the activities of the cycle. However, the difference between the evaluation phase (validation and verification, which are aimed respectively at checking *i)* the correctness and the completeness of applied algorithms and of programming language rules and *ii)* the satisfaction of costumer's requirements) of software life cycle and that one of interactive system is that the software life cycle does not support the user's perspective as the HCI evaluation may do. All of the requirements for an interactive system cannot be determined from the start and so it is necessary to observe and evaluate users to determine how they interact with the system. Moreover, usability specifications (derived from shared HCI principle and guidelines) can be incorporated in the requirement specification of software life cycle to include the usability since the early stage of the design. The international ISO standard 9241 [http://www.system-concepts.com/stds/status.html] on *Ergonomic requirements for office work with visual display terminals* also recommends the use of usability specifications as a means of requirements specifications. This is one of the basic of *usability engineering* [Nielsen, 1994; Whiteside et al, 1988] approach. In relation to the software life cycle, one of the most important features of usability engineering is the "the inclusion of usability specification, forming part of the requirement specification,

that concentrates features of the user-system interaction which contribute to the usability of the product [Dix et al., 1998, p. 199]".

Whiteside, Bennett and Holtzblatt [1998] provide a list of measurement criteria (**usability metrics**) that can be used to determine the measuring method for a usability attribute:

- o time to complete a task,
- o per cent of task completed,
- o per cent of task completed per unit time,
- o ratio of successes to failures,
- o time spent in errors,
- o per cent or number of errors,
- o per cent or number of competitors better than it,
- o number of commands used,
- o frequency of help and documentation use,
- o percent of favorable/unfavorable user comments,
- o number of repetitions of failed commands,
- o number of runs of successes and of failures,
- o number of times interface misleads the user,
- o number of good and bad features recalled by users,
- o number of available commands not invoked,
- o number of regressive behaviors,
- o number of users preferring a specific system,
- o number of times users need to work around a problem,
- o number of times the user is disrupted from a work task,
- o number of times user loses control of the system,
- o number of times user expresses frustration or satisfaction.

The ultimate test of product usability is based on measurements of users' experience with it. So, the problem with usability metrics is that they rely on measurements of very specific user actions in very specific situations that are not available at the early stages of design. Moreover, usability engineering provides a means of satisfying usability specification and not usability.

For Dumas and Redish [1999, p. 4] usability is "an attribute of every product, just like functionality. Functionality refers to what the product can do. Testing functionality means making sure that the products works according to specifications. Usability refers to how people work with the product. Testing usability means making sure that people can find and work with the functions to meet their needs."

Usability test is characterized by the following features [Dumas and Redish, 1999, p. 22]:

- o   the primary goal is to improve the usability of a product,
- o   participants represent real users,
- o   participants do real tasks,
- o   users' performances are recorded,
- o   data are analyzed and, as consequence, changes will be recommended.

For more details about usability testing see [Dumas and Redish, 1999; Rubin, 1994].

Usability testing and controlled experiments (empirical evaluation) share some aspects. The main difference is that usability test has the specific goal of testing usability, while controlled experiment can be based on other (and more complex) hypotheses. However, before starting an empirical evaluation of a system it should be better to test its usability, in order to avoid that usability problems affect the final experimental results. For this purpose, Kobsa [2002] in his comparison of usability tests and "Social Science" studies sketches these similarities:

- o   both often performed in a laboratory,
- o   participants sampled who are representatives of the population of interest,
- o   participants are frequently trained (to control variables, etc.),
- o   objective and subjective measures taken,
- o   data are analyzed and report is written,

and differences

- o   different goals (improving interface vs. truth of hypotheses),
- o   multicausal interface problems vs. isolating few variables,
- o   often, descriptive statistics only is used in usability tests, and inferential statistics in social science studies,
- o   more weight given to reports/observations in usability tests.

The empirical evaluation, also known as controlled experiments, refers to the appraisal of a theory by observation in experiments. This method of evaluation derived from cognitive and experimental psychology. The origin of this derivation has to be searched in the need of developing a relationship between psychology and computer science. In particular, techniques and models of cognitive psychology have been applied to the problem of understanding what goes when people work with computers and how those understanding can be reflected back into the design of those systems. Dix et al. [1998], sustain that since the absence of a predictive psychological theory, in order to test certain usability property of their design designers must observe how

actual users interact with the developed product and measure their performance. The problem is that the tasks a user will perform are only known when the user keep interacting with the system for real tasks and moreover, some of the task a user performs with a system were never explicitly intended as tasks by its designers (see the Suchman's discussion in 2.5.2.1).

The general idea underlying the empirical evaluation is that by changing an element in a controlled environment its effects on user behavior can be measured. The most important criteria to follow in every experiment are

- o participants have to be credible: they have to be real users of the application under evaluation;
- o experimental tasks have to be credible: the subjects have to perform tasks usually performed when they are using the application;
- o participants have to be observed during the experiment and recorded during their performance. The recording tools can be: paper and pencil, user block notes, audio, video, computer logging software, automatic protocol analysis tools.

Empirical evaluation takes place in a lab environment. Well equipped laboratory may contain sophisticated audio/video recording facilities, two-way mirrors, and instrumented computers. On the one hand, the lack of context, and the unnatural situation create an artificial situation, far from the place where the real action takes place. On the other hand, there are some situations where the laboratory observation is the only option, for instance if the location is dangerous and sometimes the experimenters may want deliberately manipulate the context in order to create unexplored situations [Dix et al. 1998].

Since the empirical evaluation attempts to define cause-and-effects relationships, the most efficient means to establish this kind of relationship between certain events in a given environment and selected forms of behavior is the *experiment*. The basic notion of experiment is [Keppel et al., 1992, p. 6]:

> *At least two groups of subjects are treated exactly alike in all ways except one – the treatment of interest. Any difference observed in the behavior of the two groups of subjects are then attributed to, or said caused by, the differences in the specific treatment conditions.*

According to [Keppel et al., 1992], the schematic process of an experiment can be summarized in these steps:

o *Identify the issue or question of interest.* Most research starts with some question that the researcher has.

o *Review the relevant theories[10] and research* (*review based evaluation*) to find out what others have done and said in trying to answer that question and to see what research methods are typically employed within a particular paradigm.

o *Develop research hypothesis* as succinct statement of the purpose of the study.

o *Identify the independent and dependent variables.* The **independent variable** comprises the range of treatment conditions under the control (manipulated or varied) of the experimenter (for instance, an application with or without user model). Independent variable can assume a number of different values: each value that is used in an experiment is known as *level* of the variable. For example, if we are comparing an adaptive system with the non-adapted version, the various adaptation techniques tested can be the level of the independent variable adapted system. For an example see 3.4.

More complex experiments may have more than one independent variable. For instance in 3.4.8 I will describe an experiment were two variables are taken into account: *i)* adaptation and *ii)* the presence of a team of presentation agents.

The **dependent variable** (also known as response measure) comprises the behavior observed after the manipulation of independent variable and then measured (for instance, the task completion time, the number of error, etc).

o *Conduct the experiment,* which consists primarily of collecting data using a particular experimental design[11]. In this phase subjects are assigned to the different treatment conditions. In the simplest procedure called **between-subjects design** an *experimental group* of subjects is assigned to the treatment, while another group of subjects, called *control group*, is assigned to a condition consisting of absence of a specific experimental treatment. For instance, one group of subjects could be asked to complete tasks using an application with user model (experimental group) and another group without (control group). Usually, this procedure is aimed at detecting difference between the two experimental conditions (e.g., is it more successful the application designed with user model or that one without?). There may be more than two groups, of

---

[10] A **theory** is a set of proposition used to describe or explain a phenomenon. The purpose of a theory is to summarize, organize and explain a variety of specific facts into a logical framework, as well as to generate new knowledge and identity gaps in the current state of the knowledge. A good theory must be able to make predictions, not just explain what has already happened.
[11] In a statistic book, **experimental design** is referred to a general plan for conducting an experiment. The most common design for an experiment in which two or more independent variable are manipulated is called the factorial design.

course, depending on the number of independent variables and the number of levels each variable can assume.

At the other extreme is the **within-subjects design** in which each subjects serves in all of treatment conditions (e.g., subjects completing tasks using both the application with UM and that one without). In between are designs in which the subjects are serving in some but not all of the treatment conditions (**partial, or mixed, within-subjects factorial design**).

Within-subjects design can cause undesired learning effects that have to be balanced.

o *Use descriptive statistic to describe the data.* These statistics (such as the mean, the variance, the standard deviation) are designed to describe or summarize a set of data.

o *Use inferential statistic to evaluate the statistical hypothesis.* These statistics are designed to make inferences about larger populations [see section 1.1.5.1].

o *Draw conclusion regarding the research hypothesis.*

o *Prepare a formal report for publication or presentation.*

Several issues regarding the population studied in the experiment must be considered during the planning phase. Three important decisions must be taken at this point:

o the nature of subjects,

o selecting the subjects,

o choosing the number of subjects.

The nature of subjects. In many cases, the primary factor is the availability. Students, friends and relatives are largely available, but the type of subjects should strictly depend from the nature of the research question. In case of software applications, for instance, the most appropriate participants are those ones chosen between the actual users of the evaluated application. If they are not available, other subjects can be chosen to be of similar age and level of education as the intended user group.

Selecting the subjects (also known as sampling, see 1.3). Even if the *population*[12] of the experiment is a limited population, it is often difficult if not impossible to collect data from all members of that population. Psychologists use to make inferences about populations based on information collected from a sample of that population. Therefore, the goal of sampling is to collect data from a representative sample drawn

---

[12] The **population** is here defined as the total number of possible units or elements that can be included in a study.

from a larger population to make inferences about that population. The ability to make generalization about a population from a sample is what is referred to as *external validity*. The external validity can be threatened by *population* and *ecology.* Population affects variability when experiment population differs from the target population (for instance student are used for testing whereas the actual users have different abilities). Ecological threats include incorrectly describing independent variables, incorrectly describing and measuring dependent variables, multiple treatments interference, etc.

Choosing the number of subjects (see section 1.3). An experiment needs as many subjects as are necessary to provide a relatively sensitive[13] test of research hypothesis. One way to increase the sensitivity in an experiment is to increase the number of subjects.

In an ideal experiment only the independent variable should vary from condition to condition. In the reality, other factors are found to vary along with the treatment differences. These unwanted factors are called *confounding variables* (or *nuisance variables*) and they usually pose serious problems if they influence the behavior under study since it become hard to distinguish between the effects of the manipulated variable and the effects due to confounding variables. The presence of confounding variables usually ruins an experiment. For instance, in an experiment comparing an application with and without UM (*with and without UM evaluation design*), if the UM application is always tried in the morning and that one without UM always in the afternoon, the different time of the day may influence the subject's performance because of the fatigue, different network loading times, etc. Other potential problems with times, locations, or other environmental conditions can influence the dependent variables: there may be more distracting noise at certain times; computer may be slower at certain times; the experimenter may bias the participants by words, body language, appearance and so on.

Experiments can be performed blind or double-blind. In blind experiments, participants do not know if the software is adaptive and so is "supposed to be better" (to avoid the placebo effect). In the double-blind experiment, the experimenter is also not aware, and so cannot inadvertently influence participants.

One way to control the potential source of confounding is holding them *constant,* so that they have the same influence on each of the treatment conditions (for instance the testing environment, the location of the experiment, the instructions given to the participants may be controlled by holding them physically constant).

Unfortunately, not all the potential variable can be handled in this way (for instance, reading speed, intelligence, etc). For the other remaining nuisance variable,

their effect can be neutralized by **randomly assign**[14] subjects to the different treatment conditions. Other rules of thumb to avoid the influence of nuisance variable are the following (Chin, 2002, pp. 185,186):

o   randomly assign time slots to participants

o   test room should not have windows or other distractions.

o   participants should be isolated as much as possible.

o   the computer area should be prepared ergonomically for different sized participants.

o   if it is needed, using the network avoiding the high load times.

o   prepare uniform instructions to participants, preferably in a written or taped form.

o   experimenters should not know if the experimental condition has or not a user model, for instance. each experimenter should run equal number of each treatment condition to avoid inadvertent bias and she should minimize interactions with participants.

o   be prepared to discard participant data if the participant requires interaction with experimenter during the experiment or if the data are compromised for some other reasons (computer crashes, too slower performances, etc).

o   follow typical local rules and law about human experimentation (such as consent form, etc)

o   planning enough time: experiments typically take months to run.

o   do run a pilot test[15] before the main study.

o   brainstorm about possible nuisance variable.

1.1.5.1.   **Inferential statistic**

Chance factors are present in every experiment. The uncontrolled source of variability assumed to occur randomly during an experiment is called **experimental error**, while the effects of the differential treatments are known as **treatment effects.** The **between-groups variability**[16] reflects the effects of the treatments plus chance factors:

between-groups variability = (treatment effects) + (experimental error)

---

[13] The **sensitivity** is the ability to detect difference when they are present..

[14] **Random** means that each subject has an equal chance of being assigned to any of the treatment variables.

[15] **Pilot test** is nothing more then a dry run for the main test to follow. In a pilot test participants are treated exactly as in the main test: same procedures are followed, same data are collected. However, the so collected data are not analyzed or included with the rest of data collected during the main test. The most important objective of a pilot test is to "debug" the equipment, material, software and procedures used during the test. A pilot test should be scheduled a couple of days before the main test. For details see Dumas and Redish (1999).

[16] In the **between-subjects design** where subjects are randomly assigned to only one of the treatment conditions, the **between-groups variability** represents the difference among the treatment means.

The other measure of variability that takes into consideration the variation of subjects treated alike is the **within-group variability**[17] that provide an estimate of experimental error:

within-groups variability = experimental error

To determine whether differences among the treatments are due to chance, these two measure of variability are used to form a useful index called **treatment index,** which is obtained by dividing the between-groups variability by the within-group variability:

$$treatment\ index = \frac{between\text{-}groups\ variability}{within\text{-}groups\ variability}$$

since

$$treatment\ index = \frac{(treatment\ effects) + (experimental\ error)}{experimental\ error}$$

if the expected value is equal to 1 there are not treatment effects (and therefore the **null hypothesis**[18] is true), otherwise if the expected value is greater than 1 treatment effects are present (and therefore the null hypothesis is rejected and an **alternative hypothesis**[19] can assumed to be true).

The treatment index is also known as **F ratio** and is calculated by the following formula, which is known as **ANOVA** (analysis of variance) [20]:

$$F = \frac{MS_A}{MS_{s/A}} \qquad\qquad [1,\ 1]$$

where

$$MS_A = \frac{SS_A}{df_A} \qquad\qquad [1,\ 2]$$

and

$$MS_{s/A} = \frac{SS_{s/A}}{df_{s/A}} \qquad\qquad [1,\ 3]$$

MS_A = the mean square of between group deviation[21]

---

[17] The **within-group variability** takes into consideration the variation of subjects treated alike.
[18] The **null hypothesis** is the statistical hypothesis evaluated by hypothesis testing. Usually represented as the absence of a relationship in the population.
[19] The **alternative hypothesis is** the hypothesis that is accepted when the null hypothesis is rejected.
[20] The **variance** is the average of the sum of the squared deviations from the mean.

SS_A = the sum of square[22] of between group deviation

df_A = the degree of freedom[23] of between group deviation

$MS_{s/A}$ = the mean square of within group deviation

$SS_{s/A}$ = the sum of square of within group deviation

$df_{s/A}$ = the degree of freedom[24] of within group deviation

If the obtained value of **F** is equal or exceeds *a critical value* the null hypothesis is rejected, otherwise is retained. The critical value of F is calculated by means of a table of *critical value,* constructed for all possible combinations of *a* (value of the number of the treatment conditions) and *n* (*n* is the value of the number of participants) in a relatively small space.

For instance, in an experimental condition with 2 treatment conditions (a system with and without user model) and 10 participants to each condition (*a*=2 and *n*=10) the degrees of freedom associated with the numerator term of the F ratio[18] (MS_A) are df_A = 1, while the degrees of freedom associated with the denominator term ($MS_{s/A}$) are $df_{s/A}$ = 18. The other piece of information needed to find the critical value of F is the *significance level a*[25]. The accepted significance levels are 0,05 (*a* =5%, "significant") and 0,01(*a* =1%, "very significant"). Turning to the F table the first step will be to find the intersection of the column listing df_{num} = 1 df_{denom} = 18 for the two critical values of *F*. For instance for *a* = 0,05 (the commonly accepted level of significance in HCI) the corresponding value of *Fa* is 4,41. The decision rule concerning this experiment will be: *if the obtained value of **F** is equal or exceeds Fa=4,41 the null hypothesis is rejected, otherwise is retained.* Therefore if in the example the obtained value of *F* are

F(1, 18)= 67,42 p = 0,05

the null hypothesis will be rejected.

Another test that analyzes the results of two-groups studies and the differences between two means is the **t test**[26]. The *t* test is algebraically equivalent to the *F* test and it also consist of a ratio:

---

[21] In al the presented formulas the deviation is meant as the deviation from the mean score of the referred group.

[22] In the calculation of variance, the **sum of square** is the sum of squared deviations of scores from their mean.

[23] The **degrees of freedom** are the number of independent pieces of information available in the estimation of population parameter. In the **numerator** term of the F ratio the *df* is equal to the number of the treatment conditions minus 1, *df=a.*

[24] In the **denominator** term of the F ratio the **degrees of freedom** are calculated as (*a*)(*n-i*) where *a* is the number of the treatment conditions and *n* the number of subjects.

[25] The **significance level** is the probability (α) with which an experimenter is willing to reject the null hypothesis then in fact it is correct.

[26] The *t* test is a special case of the *F* test because the two tests are algebraically different: $F = (t)^2$ and $t = \sqrt{F}$.

$$t = \frac{Y_{A1} - Y_{A2}}{\sigma_{diff.}} \qquad [1, 4]$$

where Y_{A1} and Y_{A2} represent the two means being compared and the $\sigma_{diff.}$ is the estimate of the experimental error in the context of the t test and is called **standard error of the difference between two means** and may be calculated as follows:

$$\sigma_{diff.} = \sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}} \qquad [1, 5]$$

where s^2_1 and s^2_2 are the variance of the two groups and n_1 and n_2 are the corresponding sample sizes. The significance of a t is determined by a table of *critical values of the t Distribution* that lists critical values of t for two significance levels (a=0,05 and 0,01) and for various degrees of freedom[27]. If the obtained value of t, disregarding sign, is equal or exceeds *a critical value* the null hypothesis is rejected, otherwise is retained.

The methods described above can be applied when the variables to measure are *continuous:* they can take value, as time, for instance, or number of errors, etc...If variable are *discrete*, or *categorical* [28], they can take only a finite number of values, or levels, such as, for instance the color of a computer screen.

The common measure exploited to evaluate the significant values assumed by categorical data in a contingency table (to test the relationship between rows and columns) is the **Chi square (X^2) test statistic.** A contingency table is a two-dimensional matrix used with categorical data to classify subjects jointly on the basis of one variable (having values in columns) as a function of another variable (having values in rows).

In Chi Square test the observed frequencies with which different classes of response occur are compared with expected frequencies derived from theoretical or empirical considerations. The Chi square is calculated as follows

$$X^2 = \sum \frac{(f_0 - fe)^2}{fe} \qquad [1, 6]$$

where

f_0 are the observed frequencies

---

[27] The degree of freedom for the t test is df_{diff}=(n_1 − 1) + (n_2 − 1).

[28] The **categorical data** are data consisting of a classification of the behavior of subjects into a number of mutually exclusive response categories (e.g., preference for screen colors)

fe are the expected frequencies

For more details see [Keppel, 1991; Keppel et al., 1992].

1.1.5.2. Statistical errors

The **type I (or α error)** happens when then null hypothesis is rejected even though it is true (wrongly assume treatment effect). The α errors can be reduced by increasing the significance level (this is increase the type II error though).

The **type II (or β error)** happens when then null hypothesis is accepted even though it is false (fail to recognize treatment effect). The β errors can be reduced by increasing the sample size, by increasing the size of the treatment effect, by decreasing the amount of experimental error, by using a more sensitive experimental design (see below).

To keep **familywise type I error**[29] from exceeding some arbitrarily chosen level regardless of the comparison conducted correction techniques can be designed. For instance the **Scheffé test** guarantees that the probability of familywise type I error will not exceed 0,05, no matter how many comparison a researcher chooses to make. Other correction techniques are the **Tukey test** and the **Dunnet test**. For more details see [Keppel, 1991; Keppel et al., 1992].

Statistical theory tells that even if the treatment mean is the best estimation of the treatment population mean, this estimation is in error because of chance factors. Some estimate of sampling error can be obtained by using the standard deviation[30] of the group to estimate the variability of subsequent estimate of the mean obtained by repeated random sampling from the same treatment population. The standard deviation of this distribution (**sampling distribution of the mean**) is called the **standard error of the mean**. That is

$$\text{standard error of the mean } (\sigma_M) = \frac{s}{\sqrt{n}} \qquad [1, 7]$$

where

s is the standard deviation obtained from the sample data and the denominator is the squared root of the sample size *n*.

1.1.5.3. The sensitivity

The sensitivity of an experiment is given by the effect size and the power.

The **effect size (treatment magnitude)** measures the strength, or the magnitude, of the treatment effects in an experiment. It gives the magnitude of the change in the

---

[29] The **familywise type I error** refers to the probability of committing type I errors over a sets of statistical tests.
[30] The **standard deviation** is the square root of the variance.

dependent variable values due to changes in the independent variable as a percentage of the total variability. The treatment magnitude is calculated by the following ratio:

$$treatment\ magnitude = \frac{estimated\ population\ treatment\ effects}{(estimated\ population\ treatment\ effects + the\ estimated\ population\ error\ variance)}$$

The estimate of treatment magnitude may be expressed by the following formula:

$$\text{Estimated magnitude of treatments } (\omega^2{}_A) = \frac{\sigma^2{}_A}{\sigma^2{}_A + \sigma^2{}_{S/A}} \qquad [1, 8]$$

The expression $\omega^2{}_A$ is called **estimated omega squared** where
$\sigma^2{}_A$ is the estimated population treatment effects and it is equal to

$$\sigma^2{}_A = \frac{(df_A)(MS_A - MS_{S/A})}{(a)(n)} \qquad [1, 9]$$

and
$\sigma^2{}_{S/A}$ is the estimated population error variance and it is equal to [see 1, 3]

$$\sigma^2{}_{S/A} = MS_{S/A}$$

ω^2 is independent of the sample size. It is a measure of the proportion of the total variation that is explained (accounted for) by the treatment.

In behavioral sciences *small, medium, large* effects of ω^2 are respectively

$$0,01 \ / \ 0,06 \ / > 0,15.$$

If the effect size is smaller, then larger number of participants will be needed to accumulate the signal from the independent variable manipulations.

The **power of an experiment (1-β)** is the ability to recognize treatment effects. The power can be increased by reducing the treatment variability, by analyzing the control factors instead of randomizing them, by reducing the subject variability (by means of, for instance, within subjects designs or analysis of covariance[31]). The power can be used for estimating the sample size by the following formula:

$$n' = \varphi^2 \left(\frac{1 - \omega^2{}_A}{\omega^2{}_A} \right) \qquad [1, 10]$$

where

---

[31] The **covariance** is the measure the degree to which deviations vary together or covary. If the deviations of the X variable tend to be of the same size as the corresponding deviations of the Y variable, the covariance will be large. If the relationship is inconsistent the covariance will be small in value, while if a systematic relationship is absent the covariance will be zero. Common covariates include age, gender, socioeconomic status, ethnic background, education, learning styles, previous experience, prior knowledge and aptitudes.

n′ is the estimated sample size that will be determined by the formula

φ is a statistical quantity that can be obtained by means of a table of power functions for the analysis of variance and it strictly depends on the established value of the power. In social science the accepted value of the power is equal to 0,80, which means that the 80% of repeated experiments will give the same results.

Designing the experiments to have a high power rating not only ensures greater repeatability of results, but it makes more likely to find the desired effects. However, if the power is low, then it may just mean that there were not enough participants in the study rather than there was no difference.

The following factors are dependent on each other in the following way

power * effect size * sample size * significance level = *quality of experiment*

where * means something like "positive contribution" [Kobsa, 2002].

To improve the sensitivity of experiment (and thus reduce the number of participants needed) *within-subjects design* can be planned. Within-subjects design uses the same participant for multiple dependent variable conditions (for example, the same subject use both the user-adapted system and the system without adaptations).

1.1.5.4. **The factorial experiments**

The statistics we have considered so far share one characteristic: they involve *single factors experiments* since they manipulate only a single independent variable at time. When two or more variable are manipulated at the same time we are dealing with *factorial experiments.* Two independent are chosen because they have been shown to be important variable in single-factor experiments and studying their combined effects in a factorial experiment seems natural.

The essence of the factorial design is the joint manipulation of two or more independent variables. If the two variables combine to influence each other an *interaction* is present. The term *treatment combination* is used, instead of treatment conditions, in order to stress the fact that the distinguishing characteristics of the different treatment result from the combination of two independent variables.

	Level of Factor A		
Level of Factor b	a1	a2	a3
b1	*n*	*n*	*n*
b2	*n*	*n*	*n*

Table 1.1. An example of 3 x 2 factorial design (from [Keppel et al. 1992, p. 230])

Consider the graphic representation of a factorial design in Table 1.1. The independent variable, factor A, consist of three levels (a = 3), while the independent variable, factor B, consist of two level (b = 2). The design of this example is called 3 x 2 factorial design (read three by two). In the example there are 6 treatment combinations.

As stated above, interaction considers the joint influence of the two independent variables. More specifically, *an interaction is present when the effect of one of the independent variable on the dependent variable is not the same at all levels of the second independent variable* [Keppel et al. 1992, p. 244]).

Other two possible effects are possible: they are known as **main effects of factors A and factor B**. They disregard the joint influence and comprise the separate effects of each independent variable, averaged over the levels of the other. If the interaction is not significant, the interaction is turned to the main effects. The detailed analysis of mains effects essentially treats the results of the factorial experiment as equivalent to two separate single-factors experiments:

- o the A main effect being viewed as a single-factor experiment where factor A is manipulated (paying no attention to factor B),
- o the B main effect being viewed as a single-factor experiment where factor B is manipulated (paying no attention to factor A)

Therefore, to pinpoint the features of the independent variable that are responsible for the significance of the overall main effect further analyses are needed. The **main comparisons** are those analyses that examine portions of a main effect.

Otherwise, if an interaction is present we have to examine the way in which the two independent variables combine to influence the behavior under study. A common way to identify the factors that are acting in the interaction is to examine the pattern of results associated with one of the independent variable as the other variable is changed systematically. This pattern is called **simple effect**. Interaction exists if the simple effects are different: if the pattern of results for one of the independent variables is not the same at all levels of the other independent variable.

Pairwise comparisons are created when a treatment mean is compared with another. They are also calculated in factorial design for further analyzing significant simple effects. Again this analysis is conducted as an actual single-factor experiment. These test are called **simple comparisons**.

The calculation of ANOVA in factorial design is more complicated then single factor design. For more details about this formula and the other ones concerning main effects and analytical comparison see [Keppel et al, 1992].

1.1.5.5.   **The within-subjects designs**

The simplest type of experimental design is one in which subject are assigned randomly to the different conditions in the experiments and are given only one of treatments. This is called *completely randomized design* or *between-subjects design.* When the same subject serve all the treatment conditions rather than just one, the design is called a **repeated-measures** or a **within-subjects design**.

One reason to use within-subjects design is the scarcity of subjects. Another good reason is that this type of design minimizes the amount of experimental error (error variance) and therefore *increases power and sensitivity.*

However, in most experiments the primary source of variability is the subjects. Using *matched subjects* (matched on characteristic assumed to be relevant to the behavior under study) the effects of this variability can be reduced by holding factors constant. Two types of matching exist:

- o *forming a group of homogeneous subjects* then assigning them randomly to the different treatment conditions;
- o using a *blocking design*, which consists of smaller groups of subjects, each one matching closely a relevant characteristic. Then subject within each block are randomly assigned to the different treatment conditions. The final design is a factorial experiment.

The most typical method to reduce error variance is to use the same subject in all the treatment conditions. The result is a perfect matching across conditions reducing error variance. However, some problems are associated with this type of design as well. First of all, *i*) *the carryover effect* from previous conditions on the responses of subjects to the current treatments and the *ii*) *general practice effects* occurring as subjects' progress through the entire experiment.

Between-Subjects Design				Within-Subjects Design		
a1	a2	a3		a1	a2	a3
s1	s4	s7		s1	s1	s1
s2	s5	s8		s2	s2	s2
s3	s6	s9		s3	s3	s3

Table 1.2. A comparison of Between-Subjects and Within-Subjects Design (from [Keppel et al. 1992, p. 321])

While in between-subjects design differences between treatment means may be directly entirely due chance, since subjects are randomly assigned, in within-subjects design this variability is not present, because the same subjects serves in each

treatment conditions. The chance factors operating in the within-subjects design can be determined and then removed from the analysis by estimating the degree to which the subject is consistent from treatment to treatment and then remove this estimate from the individual treatment means. For the calculation of ANOVA in within-subjects design and analytical comparisons, see [Keppel et al, 1992].

1.1.5.6. **Partial (or mixed) within-subjects design**

The two features described above characterize contemporary experimental research:

- o *factorial design*, with two or more variable are manipulated in the same experiment and all the subjects involved receive all the treatment combinations [see A in Table 1.3];
- o *within-subjects design*, having subjects serving in more than one treatment combination [see B in Table 1.3].

A - Completely Randomized Between-Subjects Factorial

	b1	b2	b3
	s1	s4	s7
a1	s2	s5	s8
	s3	s6	s9

	b1	b2	b3
	s10	s13	s16
a2	s11	s14	s17
	s12	s15	s18

B - Pure, or Complete, Within-Subjects Factorial

	b1	b2	b3
	s1	s1	s1
a1	s2	s2	s2
	s3	s3	s3

	b1	b2	b3
	s1	s1	s1
a2	s2	s2	s2
	s3	s3	s3

C - Partial, or Mixed, Within-Subjects Factorial

	b1	b2	b3
	s1	s1	s1
a1	s2	s2	s2
	s3	s3	s3

	b1	b2	b3
	s4	s4	s4
a2	s5	s5	s5
	s6	s6	s6

Table 1.3. A comparison of Factorials Designs (from [Keppel et al. 1992, p. 361])

Point A in Table 1.3 shows a **completely randomized two-variable factorial design** where an equal number of subjects (*n*) are randomly assigned to each of the (*a*) (*b*) treatment combinations.

Point B in Table 1.3 shows a **two factors within-subjects design** all subjects (*n*) serve all the (*a*) (*b*) treatment combinations.

Point C in Table 1.3 shows a **partial, or mixed within-subjects factorial design,** which contains element of elements of both between-subjects and within-subjects design. This type of design can be referred to in the literature as a "2 X 3 factorial design with factor *A* represented as a between-subjects variable and factor *B* as a within-subjects variable [Keppel et al. 1992, p. 362]."

Table 1.3 easily shows that the three designs differ substantially for the number of subjects required: respectively 18, 3 and 6.

Within-subjects design is the one requiring fewer numbers of subjects. However, completely within-subjects design present problems such as the carryover effects and large time demands on a single subject that are reduces in mixed factorial. For instance, in the 2 x 3 shown in Table 1.3, the mixed factorial requires that subjects serve in three conditions, while the complete within-subjects requires that subjects serve in all six conditions. So, the above problems result reduced in mixed design.

For details for the calculation of ANOVA see [Keppel et al, 1992]. While for analytical comparison the rules are the same as for factorial design: if the A x B interaction is significant analyze simple effects and simple comparisons, otherwise focus on main effects and main comparisons. For an example see 3.4.

To sum up, the total variability of a mixed within-subjects factorial design is divided in

- o the variability extracted from any tow-variable factorial experiment: the main effect of factor A, the main effect of factor B and the A x B interaction,
- o the between-subject error, which estimates the extent to which chance factors are responsible for any differences observed among the different levels of factor A, the between-subjects factor,
- o the within-subject error, which estimates the extent to which chance factors are responsible for any differences observed within the same subjects.

1.1.5.7. **The statistical correlation**

In an experiment, variables are manipulated and the consequential changes in another variable are measured. In correlational studies, both variables are measured because there are no true independent variables. The variables being measured are characteristics naturally occurring in the subject, in spite of the study. Here, independent variable refers to the variable or characteristic whose influence on

another variable (dependent variable) is the object of study. For instance, is there a relationship between user's features and layout preferences?

Correlational data are plotted in a scatter plot in order to examine the relationship between the two variables of interest. This is a graphical representation of the relationship between individual scores on two measures (identified as X and Y). Each individual is represented by a single point on the graph so that the coordinates of the point (the X and Y values) match the individual's X score and Y score.

A statistic usually exploited to measure correlation is the **Pearson correlation coefficient** (**r**), also known as the **product-moment correlation coefficient** or the **linear correlation coefficient**. The Pearson correlation coefficient measures the strength of a relationship and it is the most common index of linear relationship[32] between two variables. It ranges from –1.0 and +1.0 (perfect negative and perfect positive correlations, respectively. The sign refers to the direction of a relationship). A value of zero represents the complete absence of a correlation. The *r* is calculated by the following formula:

$$r = \frac{covariance\ (X, Y)}{\sqrt{[variance\ (X)][\ variance\ (Y)]}}$$

$$r = \frac{\sum (Xi\text{-}X)(Yi\text{-}Y)}{\sqrt{\sum (Xi\text{-}X)^2 \sum (Yi\text{-}Y)^2}} \qquad [1, 11]$$

which is the ratio between the covariance of X and Y and the product of standard deviations of X and Y. The correlation coefficient may therefore be defined as the ration of the joint variation of X and Y relative to the variation of X and Y considered separately.

If the data involved in the relationship are ordinal, a non-parametric test such as Spearman *rho* Correlation is used. For details see Keppel et al. [1991]

Correlation tells whether there is a relationship between two variables. If we want to make use of correlational data to make prediction on a variable on the basis of another variable we can exploit a **regression equation.** For details see [Keppel et al., 1992].

---

[32] The linear relationship between two variables is depicted by a best-fit straight line called **linear regression line** which is characterized by two features: slop and intercept. The formula for the best-fit straight line is used to predict one variable from another.

1.1.5.8.   **Non-parametric statistic**

Statistical hypothesis tests can be broadly categorized into two types: parametric and non-parametric.

Parameters describe a mathematical function, as a frequency distribution. For example, a normal distribution can be described by its mean and standard deviation. The mean and standard deviation are therefore the parameters of the normal distribution. Parameters describe the population distribution, which we can only estimate, rather than the sample distribution, which we can measure.

Parametric tests (e. g. Anova, *t* test) make three assumptions about the individual score of the subjects present in a hypothetical treatment population:

1. normality - the scores should be normally distributed (**assumption of normality**)

2. homogeneity of variance - the scores from different populations should have the same variability (**assumption of equal variance**)

3. interval/ ratio - the scores should be measured at either the interval or ratio level (**assumption of independence**)

If any of these cannot be met, must turn to non-parametric tests.

Non-parametric tests are those that make no assumptions about the distribution of the data. They are therefore more robust when data do not have well-behaved distributions. They are generally used to investigate hypotheses about samples as a whole, rather than about properties such as means. Example of non-parametric tests are [Keppel, 1991]:

Type of design	Number of conditions	Test
Between-Subjects	Two	Mann-Whitney
	More than two	Kruskal-Wallis
Within-Subjects	Two	Wilcoxon
	More than two	Friedman

Table 1.4. Classification of non-parametric tests [Keppel, 1991]

For more details about non-parametric statistics see [Keppel, 1991].

1.1.6. Empirical evaluation methodologies in user-adapted systems

As discussed above, the key to good empirical evaluation is the proper design and execution of experiments so that the particular factors to be tested can be easily separated from other confounding factors. In UM system, for example, it can be tested

if the system with user model works better then the same system without user model by testing the different user interfaces. In this case the independent variable is the presence/absence of the user model while the dependent variable can include response variables or recorded measures such as the frequency of certain behaviors, qualities of a behavior in a particular situation, number of errors, error rate, time to complete a task, proportion/qualities of tasks achieved, interaction patterns, learning time/rate, user satisfaction [Chin, 2001]. Some dependent variables can only be measured indirectly such as cognitive load measured through blood pressure or pupil dilatation and eye-tracking [Marshall, 2001].

At the end of his overview on empirical evaluation of user-adapted system, David Chin proposed that authors pushing in UMUAI should report the following common measures [Chin, 2001]:

- o the number, the source and relevant background of participants,
- o the independent, dependent and covariant variables,
- o the analysis method,
- o the post-hoc probabilities,
- o the raw data (in a table or appendix) if not too voluminous,
- o the effect size and the power (which should be at least 0,8),
- o the non significant results (with corresponding effect size and power)

Moreover, Chin listed a set of further factors that could compromise the validity of the experiment:

- o the data can be contaminated,
- o there be unwarranted assumptions about scales for variables,
- o nuisance variables can be confounded with relevant variables,
- o previous training of participants should be taken into account,
- o a non-sufficient number of participants cannot provide the needed precision,
- o some experimental procedure can affect the observed conditions (e.g., the video cameras),
- o factors such as *history, maturation, testing, instrumentation, statistical regression[33], mortality* and *selection* can threaten the internal validity of the experiment[34] ,
- o threats to the external validity (see 1.1.5),
- o difficulties in the interpretation of the results (using unsuitable measures, improvement only in some subsets of participants).

---

[33] The **statistical regression** refers to the tendency for the means of extreme scores to drift back to the middle.
[34] The **internal validity** of an experiment refers to whether the independent variables made a difference in the study and, if so, whether the researcher is able to infer a cause and effect relationship.

The statistics described above are often reinforced by user's suggestions in order to know the deeper effects generated by the exploitation of adaptive techniques such as a reduction of the complexity of the interaction, a deeper knowledge of the system, an increased satisfaction, a reduction of the interaction anxiety, and so on.

Tasso and Omero [2002] proposed other measures to globally evaluate adaptive e-commerce Web sites:

- o the percentage of repeated visits to the Web site,
- o the average length of the visit,
- o the average time spent to read a page,
- o the conversion ratio (referred to the percentage of users who become buyers),
- o the average user expense,
- o the ROI (Return of Investments),
- o the usability,
- o the user satisfaction.

1.2. The selection process evaluation

1.2.1. Precision and recall

The main metrics are derived from the evaluation of information retrieval system since the problem is quite similar: from a set of contents a sub-set of user-relevant contents has to be extracted [Tasso and Omero, 2002]. The two fundamental measures are

- o *precision*, which is referred to degree of accuracy of the selection process. It is measured as the ratio between the user-relevant contents and the contents presented to the user,
- o *recall*, which is the ratio between the user-relevant contents and the contents present in the contents collection.

Hence, precision indicates how selective the system is, and recall indicates how thorough it is in finding valuable information. While the precision can always be calculated, the recall needs the exact number of the contents in the collection, and sometimes this number is unknown.

1.2.2. Training set and test set

To overcome the above problem, in the collaborative filtering system, for instance, the metric are calculated starting from the data of the real system usage (e.g., the log files, the purchases data of an e-commerce site) which are divided in

- o *training set* (usually the 80% of the available data)
- o *test set* (the remaining 20%)

The training set data are used to find and select the user's neighbors and then generate the recommendations and the remaining data are used to evaluate the accuracy of the recommendations. Therefore the precision is calculated as the ratio between the user-relevant contents and the number of recommendations, while the recall is calculated as the ratio between the user-relevant contents and the number of data present in the test set. In this way, the precision represents the number of right suggestions among the overall suggestions, while the recall is the ratio between the right suggestions and the suggestions that should be correct for the selected subjects. Both the precision and the recall tale values ranging between 0 and 1. Typically when the recall increase the precision can decrease.

The training set - test set methodology is borrowed by machine learning.

1.2.3. The evaluation of the ordering

When in the information filtering system a relevance measure is present and the content are order on the basis of this measure, the ordering is evaluated by human subjects.

1.2.4. Coverage

Sarwar et al. [Sarwar et al, 1998, p. 6] define coverage as a measure of the percentage of items for which a recommendation system can provide recommendations. A low coverage value indicates that the user must either forego a large number of items, or evaluate them based on criteria other than recommendations. A high coverage value indicates that the recommendation system provides assistance in selecting among most of the items.

A basic coverage metric is the percentage of items for which predictions are available. This metric is not well defined, however, since it may vary per user, depending on the user's ratings and neighborhoods. To address this problem, Sarwar et al. [Sarwar et al, 1998, p. 6] use a usage-centric coverage measure that asks the question: "Of the items evaluated by the user, what percentage of the time did the recommendation system contribute to the evaluation process? More formally, for every rating entered by each user, was the system able to make a recommendation for that item immediately prior to it being rated? We compute the percentage of recommendation-informed ratings over total ratings as our coverage metric".

1.2.5. Statistical accuracy metrics: MAE and RMSE

Statistical accuracy metrics [Good et al, 1999; Sarwar et al, 1998] evaluate the accuracy of a filtering system by comparing the numerical prediction values against user ratings for the items that have both predictions and ratings.

In particular, MAE and RMSE evaluate the distance between the system predictions and the user's opinion by means of rate vectors. A smaller value means more accurate system's prediction.

MAE (Shardanand and Maes, 1995) is calculated by the following formula:

$$MAE = \frac{\sum_{I=1}^{n} |u_i \quad r_i|}{n} \qquad [1, 12]$$

where

n is the number of the contents

u_i is the user's opinion on the content i

r_i is the system's prediction on the content i

The **RMSE** metric [Good et al, 1999; Sarwar et al, 1998] is calculated by taking into account the mean squared error. Therefore, the bigger errors are more weighted than the smaller ones. The RMSE is calculated by the following formula (using the above notations)

$$RMSE = \sqrt{\frac{\sum_{I=1}^{n} (u_i - r_i)^2}{n}} \qquad [1, 13]$$

Compared to the MAE that weights all the errors in the same way, the RMSE is based on the criterion that is better having smaller errors than few bigger errors.

Also the **correlation** [see 1.1.5.4] can be used as measure of linear agreement between the two vectors. A higher correlation value indicates more accurate recommendations.

Good et al. [1999], in a recommender systems evaluation, suggest that good value of MAE an RMSE should be near to 0.7, in a range of 0-5.

1.2.6. Decision support accuracy metrics: reversal rate and sensitivity measures

This kind of measures evaluate the capability of providing relevant contents such as the PRC sensitivity [Salton & McGill, 1983] or providing relevant contents and eliminating the non-relevant ones such as ROC, the reversal rating characteristic [Le & Lindgren, 1995].

Decision support accuracy metrics [Good et al, 1999; Sarwar et al, 1998] evaluate how effective a prediction engine is at helping user select high-quality items from the item set. These metrics are based on the observation that, the majority of users filtering is a binary operation: they will either view the item, or the will not.

Reversal rate [Good et al, 1999; Sarwar et al, 1998] measures the percentage of reversal recommendations (the contents the user does not like). On a five-point scale, it is commonly defined as the percentage of recommendations where the recommendation was off by 3 points or more. Low reversals refer to cases where the user strongly dislikes an item (i.e., gives a rating lower than a threshold *L*) and the system strongly recommends it with a high recommendation score (i.e., above a threshold *H*). High reversals are cases where the user strongly likes the item, but the system recommendation is poor (i.e., user rating > *H*, system recommendation < *L*).

Roc sensitivity [Good et al, 1999; Sarwar et al, 1998] is a measure of the diagnostic power of a filtering system. Operationally, it is the area under the Receiver Operating Characteristic (ROC) curve, a curve that plots the sensitivity and the specificity of the test. *Sensitivity* refers to the probability of a randomly selected good item being accepted by the filter. *Specificity* refers to the probability of a randomly selected bad item being rejected by the filter.

The ROC curve plots sensitivity (from 0 to 1) and 1 – specificity (from 0 to 1), obtaining a set of points by varying the recommendation score threshold above which the article is accepted. The area under the curve increases as the filter is able to retain more good items while accepting fewer bad items. For use as a metric, good and bad items have to be determined. For that task, the users own ratings are generally taken into account. Particularly important values are 1.0, the perfect filter, and 0.5, a random filter.

PRC sensitivity [Sarwar et al, 1998] is a measure of the degree to which the system presents relevant information. In fact, it is the area under the precision-recall curve. Sarwar et al [1998] suggest plotting a curve of different precision-recall pairs for different recommendation score thresholds, and taking the area under that curve as a metric of PRC sensitivity. A higher value is more accurate, and a low value is less accurate.

1.2.7. Utility metrics

The *utility* [Hanani et al., 2001] is a metric able to detect the *positive false* (the contents that have been suggested to the user but she does not like) and the *negative false* (the contents that have not been suggested to the user but she does probably like). The utility is calculated by the following formula:

$$utility = (A*R+)+(B*N+)+(C*R-)+(D*N-)$$

where *R+* is the number of relevant contents that the system proposes to the user, *N+* is the number of *positive false*, *R-* is the number of *negative false* and *N-* is the number of non-relevant content that the system does not propose to the user. *A, B, C, D* are multiplicative coefficient used to establish the relative benefit (when they are positive) or the relative cost (when they are negative). The concept of positive false is particularly important in case of personalized e-commerce Web site and they should be reduced since they are not related with user's purchases.

Another metric focused on the errors is the *error rate (td)* (Lewis, 1995) that is defined by the following formula (using the above notations):

$$td = \frac{(N+)+(R-)}{(R+)+(R-)+(N+)+(N-)} \qquad [1, 14]$$

and represents the ratio between the non-correctly classified contents and the whole contents.

1.2.8. The simulation

In the simulation the same data sets is exploited for more experimental sessions. For instance, in the information retrieval systems the users evaluate all the contents then the behavior of the system is simulated by means of the collected data. Different selection algorithms and different solutions are tested on the same sets of data in order to reach the optimal results. The same approach is exploited by the TREC (Text REtrieval Conference) experiment that is now considered the main benchmark for the information retrieval comparisons (Hull, 1998).

1.3. The sampling problem

If the research goal is to build and to test a theory regarding a phenomenon, it has to be possible to make inferences regarding the phenomenon as it exists in the **population**. Population may be defined as the total number of possible units or elements that can be included in a study. Since it is often difficult, if not impossible, to collect data from all member of the referred population, the inferences about population are based on information collected from a portion, or **sample**, of that population. Thus the goal of **sampling,** which can be defined as the process of selecting participants for a research project, is to collect data from a larger population to make inferences about that population.

The samplings are usually divided in *probabilistic* (each subject has a known and not null probability to be selected) and *non-probabilistic* sampling.

1.3.1. Probabilistic sampling

When all the subjects of the population have the same probability of being included on the sample it is a **simple random sampling**.

When the subjects are selected every fixed number of subjects the sample is **systematic** (for instance, if both the population size (*N*) and the sample size (*n*) are known a, number *k* equal to the ratio between the population size and the sample size can be chosen, *k=N/n*).

In **(probabilistic) blocking sampling** the population is divided in homogeneous layers related to the variable that has to be estimated (for instance age, profession, etc.). Within each layer the subjects are randomly selected. Then the sub-samples are unified to compose the final sample. If the percentage of subjects in every layers is equal to percentage of subjects of the population, the sampling is **proportional**, otherwise is **non-proportional**. In the latter case the non- proportional layer are *weighted* on the basis of their features to reproduce the distribution of the real population.

When the population list is not available an **area sampling** can be applied, as in the American Survey Center of the Michigan University. In this survey, for instance, the American nation is divided in 74 *primary areas*, then every area is divided in *location* (a big city or 4-5 medium cities) and every location is divided in *chunks* (small cities or neighbors). The chunks are divided in *segments* (street or building blocks with 4-16 housing units) and in every segment the *housing units* are finally selected. All the choices here described are carried out with probabilistic procedures (random, layered sampling).

Other probabilistic procedures are the **step (phase) sampling**, the **cluster sampling** and **complex samplings**. For more details see [Corbetta, 1999]

1.3.2. Non-probabilistic sampling

Even if the random sampling is the best way of having a representative sample, these strategies require a great deal of time and money, therefore much research in psychology is based on samples obtained through a nonrandom selection, such as **availability sampling** that is sample of convenience, based on subjects available to the researcher, often when the population source is not completely defined. Since the availability sampling may cause a loss of external validity, researchers try to generalize[35] their results by duplicating their experiments.

Other more structured non-probabilistic samplings are:

- o **Non-probabilistic blocking sampling** → is one of the most used samplings in market researches. The population is divided in sub-groups on the basis of the

---

[35] This is another meaning of the concept of external validity.

known distribution of one or more variable (age, instruction, job, gender, etc). The proportion of every sub-group in the population is then replicated in the sample. The difference from the layered sampling is that the experimenter can chose the subjects, as she likes, instead of using probabilistic procedures.

o **factorial design** → is often used in quantitative research (see 1.1.5.4). It can be probabilistic or not, depending on the selection of subjects. The factorial design is similar to the blocking sampling, but the selected sub-groups are non- proportional to the corresponding groups in the population. Every sub-group has the same number of participants. The particularity of this procedure is that the independent variables are chosen on the basis of their importance for the considered phenomenon instead of their proportion in the population. The aim of the factorial design is not describing the population, but detecting relationships within small sample groups.

o **rational choice sampling** → the subjects are selected exclusively on the basis of their features

o **balanced sampling** → is a type of rational choice sampling where the average of the selected variable is close to that of the population.

o **stream sampling** → the start is a small number of subjects that will also be used as information sources to select the other subjects. It is useful to study phenomena within clandestine or background populations.

o **telephonic sampling** → the subjects are selected by a computer from the national telephone lists or from a random generation of the telephone numbers. The problem concerns the subjects without telephone or with reserved number that are not present in the list.

In social science, a good sample is **representative** and **broad**. A sample is defined statistically representative when it reflects the population without distortions. If the sampling has been rigorously random the sample is statistically representative: probability sample can be representative, while non-probability sample cannot.

However, the random sampling can be affected by the coverage error and the non-answer error and the researcher has to minimize them by means of a rigorous probabilistic sampling.

The breadth of a sample corresponds to the number of subjects involved in the sample. The breadth of a sample is proportionally connected to the degree of desired approximation and the variability of the phenomenon and non-proportionally connected to the accepted level of survey error. If the sample is too small the error is too big. However, it is recommended to randomly select a smaller number of subjects than a bigger number of subjects chosen in a non-accurate way (for instance, by asking friends, colleagues, parents, etc).

Indeed, the sampling strategies depend on the goal of the research. In the case of **descriptive research**, where the goal is describing with increasing amount of thoroughness the distribution of two or more variables, the sample has to be strictly representative. Whereas, in case of **exploratory research**, where the goal is to investigate a topic on which little systematic information exists for providing ideas to further systematic research, the sample may be non representative.

The experimental evaluation is clearly an exploratory research and the sample cannot be too broad since the variability of confounding factors increases together with the sample size. For this reason, sometimes also the representative feature is not respected and the subjects do not represent the target population. While the first drawback can be justified by empirical reasons, the second one should be lessened by the exploitation of non-probabilistic sampling procedures. Moreover, in experimental evaluation the accuracy of selecting subjects should be strictly tied to the presence of

- o the rules described above to avoid confounding factors and the measures to detect their presence (analysis of variance, covariance, etc)
- o the measures reporting the existence and the strength of the treatment effects in the experiment (effect size, omega squared).

1.4. *The state of the art in evaluation of user- adapted systems*

During the past years the empirical research on UM and adaptive system has been very fragmented. From an analysis [Chin, 2001] of the papers appeared in the User Modeling and User Adapted Interaction [UMUAI] review during its first nine years of activity, only one third of the articles includes any type of evaluation. Moreover, some of them reported only preliminary evaluations. Excluding these last ones, only one forth of the UMUAI papers reported significant evaluation results.

Concerning the last user modeling [UM 01] and adaptive hypermedia conferences [AH 02], they reported respectively one third and one forth of papers with some kind of

evaluations. By excluding preliminary evaluations, these two estimates decrease to one third and one fifth.

In order to classify the adaptive systems and their evaluation, taking also into account the applied methodologies, Chin [2001] distinguishes five areas:

o *student modeling system* → are typically evaluated by comparing system with and without student models, by analyzing real data (such as log files) or by means of pre and post test evaluations;

o *systems that use machine learning methods to acquire user models* → are typically evaluated by using standard machine learning measures that compare the user model against a reserved set of test data that was not used for training (typically an 80/20% split for training/testing);

o *adaptive hypermedia and information filtering* → are typically evaluated by measures developed in library sciences such as recall and precision and similarity/relevance metrics;

o *plan recognition* → are typically evaluated by the percentage of actual plans recognized in a test corpus of plans, by the frequency and the accuracy of predicted next actions or by comparison with an expert human plan recognizer;

o *mixed-initiative interaction* → are typically evaluated by comparing system responses choices with human choices or by comparing the efficiency of the dialog needed to achieve an information;

o *user interfaces/help systems* → are typically evaluated by subjective user satisfaction, task completion speed, error rate, quality of task achievement.

Kristina Höök (Höök, 1997) argues that the evaluation of an interactive system is always a hard task, and when the system is adaptive the task becomes harder because it is required to be able to distinguish the adaptive features from the general usability problems. Since most adaptive system evaluations are comparison between the system with and without adaptations, the problem is clear: most of the times the non-adaptive version is not well designed for the tasks. This could happen when the adaptive behavior is the core part of the system and the non-adaptive version is therefore not completed. Moreover, the measures taken in consideration for the evaluation (task completion time, number of errors, number of viewed pages) sometimes do not fit the aims of the evaluation. For instance, during an evaluation of a recommender system is more important the relevance of the information provided than the time spent to find it. Furthermore, lots of applications are designed for a

long-time interaction and therefore it is hard to correctly evaluating them in a short and controlled test.

1.4.1. The layered evaluation

The layered evaluation is concerned with the difficulty to distinguish the different aspects of the adaptation and therefore to evaluate them without distinction. Brusilovsky [Brusilovsky et al., 2001] distinguish two high-level phases characterizing the adaptive systems:

- o *the interaction assessment phase,* aimed at reaching high-level conclusion concerning the aspects of user-computer interaction that are considered significant for the particular application (for example, it can be noticed that the user is unable to complete a task);
- o *the adaptation decision making phase,* aimed at selecting specific adaptations, based on the results of the assessment phase, in order to improve selected aspects of the interaction (for example it ca be decided to present a pop-up message helping the user complete a task).

Therefore, during the evaluation of adaptive systems these two phases should be distinguished instead of evaluating the adaptation as a whole because if the adaptive solutions do not improve the interaction is not evident whether one or both the above phases have been unsuccessful. For instance, the conclusion of the interaction assessment phase should be correct, but the adaptive choices not meaningful or the adaptation decision should be reasonable but based on incorrect assessment results.

To solve these problems Brusilovsky advocates a *layered evaluation framework* where the evaluation is decomposed into different layers corresponding to the high layers described above:

- o *the interaction assessment layer* where only the assessment phase is being evaluated. For instance a question here can be stated as "are the user's characteristics being successfully detected by the system and stored in the user model? "
- o *the adaptation decision making layer* where only the adaptation decision is being evaluated. For instance a question here can be stated as "are the adaptation decisions valid and meaningful, for selected assessment results?"

The effectiveness of such an approach has been demonstrated, for instance, by experimental results of adaptive educational system. Here has been noticed that to successfully select the right adaptation the previous knowledge of the students has to be taken into account. Indeed, experiments [see section 1.4.3] showed that more

knowledge the students have in a subject (*interaction assessment layer*) more improvement they gain by means of the adaptive annotation (*adaptation decision making layer*), while the students less experienced obtain the best results by means of the direct guidance.

Similarly Weibelzahl e Lauer [Weibelzahl e Lauer, 2001] proposed an evaluation framework of six steps:

- o evaluation of reliability and external validity of input data acquisition used to build the user model,
- o evaluation of the correctness of inference mechanism and accuracy of user properties,
- o appropriateness of adaptation decisions, which may concern how to adapt the interface, how to change the layout, what additional information should be provided, which command to offer, how to tailor the presentation, etc
- o change of system behavior when the system adapts (in which way does system behavior change in comparison to the normal division of labor?),
- o change of user behavior when the system adapts (does the user change her behavior when the system adapts in comparison to the normal division of labor? In which way?),
- o change and quality of total interaction. The main question here concerns usability. How is the interaction quality? Does it change? Is the user satisfied?

The last evaluation step can only be interpreted correctly if all the previous steps have been yet completed (this is especially important in the case of finding no difference between an adaptive and a non-adaptive system). In addiction, they assume that the adaptivity reduces the complexity of the interaction and therefore they also measure the user's behavior by means of measures of complexity to demonstrate the complexity reduction. For instance, the number of clicks to reach a goal can be used as measure since an easier interface provides shorter paths to the goals. However, in addiction to objective measures, a correct interpretation also requires subjective criteria, such as the user's preferences for one of two versions, or a standardized usability questionnaire.

Paramithys, Totter and Stephanidis [Paramithys et al., 2001] suggest a modular approach to the evaluation of adaptive user interface as well. They exploit a high-level model of adaptation made up of the following components:

- o interaction monitoring,
- o interpretation/inferences,
- o explicitly provided knowledge,
- o modeling,

o   adaptation decision making,

o   applying adaptations, transparent models & adaptation "rationale",

o   automatic adaptation assessment.

Then, they identified five evaluation modules (comprising one or more of the adaptation components listed above), which can be evaluated individually and in combinations on the basis of the evaluation goals:

o   *module A1* comprises *interaction monitoring, interpretation/inferences* and *modeling* and can be used to evaluate the *correctness* of the interpretations/inferences, the *comprehensiveness*, the *redundancy*, the *precision* and the *sensitivity* of the model. Methods as focus group, questionnaires, interviews, think aloud protocol, logging use and expert-based evaluation can be used;

o   *module A2* comprises *explicit provided knowledge* and *modeling* and can be used for the same purposes of module A1 and also for evaluating the *transparency* of the process and the possible *overhead* imposed to the user. The principal method used is the expert-based evaluation;

o   *module B* comprises the *adaptation decision making* and can be used to evaluate the *necessity,* the *appropriateness*, the *acceptance* of adaptation. The suggested methods are expert-based evaluation and user involvement.

o   *module C* comprises *applying adaptations* and can be used to evaluate the *timeliness,* the *obtrusiveness* of adaptation and the *user control* of adaptation. Summative evaluation methods are generally applied.

o   *module D1* comprises *modeling* and *transparent models* and can be used to evaluate the *completeness,* the *coherence*, the *rationality* of the presentation. End users and expert can be involved in the assessment of the model together with testing of real interaction.

o   *module D2* comprises *adaptation decision making* and *transparent adaptation rationale* and can be used to evaluate the *coherence* and the *causality* of the adaptation rationale. The evaluation methods are the same of module D1;

o   *module E* comprises *automatic adaptation assessment* and its goal is to ensure that the system shares the same views as the users with regards to the success or the failure of adaptations. The feasibility of this approach has still to be investigated.

The possible combinations of modules are the *following*:

o   *modules Ax and D1* → these modules capture the entire process of constructing models and presenting these models to the users;

o *modules B and C* → these modules capture the process of deciding upon and applying adaptations;

o *modules C and D2* → these modules capture, for examples, how predictable and how controllable the adaptive interface is;

o *modules Ax, B and C* → these modules capture the entire traditional adaptation cycle of an adaptive system.

In their approach they further clarified the following points:

o since the concept of the typical user of a system cannot be applied in adaptive system, in all cases where users are not directly involved in the evaluation, each individual evaluation task takes into account a particular user (having characteristics encoded in some type of user profile) in a particular context of use (conveyed in a way analogous to the user);

o the expert evaluations are here conducted by domain experts instead of usability expert as in the usual HCI evaluations.

Jameson [Jameson, 1999] in his overview of types of empirical studies distinguished between studies that do not require a running system and studies with a system. The former can benefit from results of previous research, early exploratory studies and knowledge acquisition from experts, while the latter requires controlled evaluations with users and experience with real-world use. All the empirical studies, besides, should address questions concerning:

o *the correctness of assumptions about users relied on by inferences techniques* → can the general assumptions about users be shown empirically to be correct?

o *the appropriateness of inference techniques used* → are the techniques used well suited to dealing with the inference tasks faced by the system?

o *the adequacy of available data* → is there typically enough data available about each user to enable the system to make useful inferences about her?

o *the adequacy of coverage* → does the system take into account enough of the relevant input data and user properties to be able to make a useful number of adaptation decisions?

o *the appropriateness of adaptation decisions* → do the adaptation decisions that the system makes on the basis of decision-relevant properties actually improve the quality of the user's interaction with the system?

1.4.2. Evaluation and usability problems in user-adapted systems

As pointed out by Höök [2002] intelligence user interfaces may violate many of usability principles developed for direct manipulation systems. The main problem of such a systems is that they may violate many good principles such as giving the user *control* over the system, making the system *predictable* so that it always gives the same response given the same input and making the system *transparent* so that the user can understand something of its inner working. In addiction, most adaptive interfaces developers are more concerned with inferences and building knowledge then interface design.

For Höök the usability of intelligent user interface sometimes can require a new way of addressing usability, different from the usability principles outlined for direct-manipulation systems. For instance, when Benyon [1993] proposed the adaptivity as solutions to usability problems, he discusses five analysis phases that need to be considered when designing adaptive systems:

- o <u>functional analysis</u>, aimed to establish the main functions of the system;
- o <u>data analysis</u>, concerned with understanding and representing the meaning and structure of data in the application;
- o <u>task knowledge analysis</u>, focused on the cognitive characteristics required of users by the system such as the assumed mental model, cognitive loading, the search strategy required, etc;
- o <u>user analysis</u> that determines the scope of the user population that the system is to respond to. It is concerned with obtaining attributes of users that are relevant for the application such as intellectual capability, cognitive process ability, etc. The target population will be analyzed and classified according to the aspects of the application derived from the point mentioned above;
- o <u>environment analysis</u> that covers the environment within which the system is to operate.

Oppermann [1994] suggested a design-evaluation-redesign approach. For Oppermann the adaptive features can be considered as the main part of a system and thus evaluated during every development phase. The problem is circular:

- o a problem solvable by the adaptivity has to be identified,
- o the user characteristics related to the hard problem have to be detected,
- o ways of inferring user's characteristic from the interaction have to be found,
- o adaptation techniques offering the right adaptive behavior have to be designed.

This process requires a bootstrapping method: first some initial adaptive behavior is implemented then tested with users, revised and tested again. The reason is that it

is hard to decide which particular adaptations should be linked to given users' actions. Furthermore, adaptations must be found to be of real use. The necessity of an iterative process is due to the fact that it should be hard to foresee the real behavior of users in a given situation and thus only by monitoring the users' activity some evidence can be shown. On the iterative evaluation point of view the design phases and their evaluation have to be repeated until some a good result is reached.

The Oppermann's iterative process is very close to the *user centered system design* proposed by Norman and Draper in the 1980s [Norman and Draper, 1986]. According to this methodology, which can be applied during the development of interactive software systems, the user is the central element of the overall activities characterizing the lifecycle of a system. The underlined assumption is that if the system designer always takes into account features, habits, preferences and behavior of the users she will be able to produce easier-to-use systems. During the different stage of the system lifecycle the designer has to exploit different techniques to gain a deep knowledge of the target users and this could happen in two ways:

- by involving the users during the different testing stages of the system (*consultative design*),
- by cooperating with the users during the different development stages of the system (*cooperative design*) and by involving them actively in system design (*participatory design*, originated in Scandinavia, see [Greenbaum and Kyng, 1991 and 2.1.2.2]). In this last case the subjects involved are member of the design team and collaborate actively in the design process since they are experts in the work context and in organizational processes. The participatory design methods include brainstorming, storyboarding, workshops, pencil and paper exercises.

Benefits of the user-centered approach are mainly related to time and cost saving, completeness of system functionality, repair efforts saving and user satisfaction (Nielsen, 1993). As pointed out by Dix et al. [1998], the iterative design is also a way to overcome the inherent problems of incomplete requirements specification since not all requirements for an interactive system can be determined from the start.

The iterative evaluation process requires an empirical knowledge of the users' behavior since the first development phases. In case of user modeling system, the choice of features relevant to the user model (such as personal features, goals, plans, domain knowledge, the context, etc..) can gain advantages by prior knowledge of the real system users, the context of use and domain experts. A deeper knowledge of the real user can offer a broader view of the application goals and prevent serious mistakes especially in case of innovative systems.

Petrelli et al. [1999][36] proposed the user-centered approach to user modeling as a way to move from designer questions to guidelines by making the best use of empirical data. They advocated the incremental system design as a way to satisfying more users. For instance, during the early stage of the development of a mobile device offering contextual information to users visiting a museum, they decided to revise some of their initial user model assumptions. For instance, they rejected the exploitation of stereotypes (the socio-demographic and personal data taken in consideration did not characterize the users' behavior) in favor of a more socially oriented (people does not like going alone to the museum) and context aware (museum visitor prefer watching paintings than interacting with a device) perspective.

1.4.3. Data mining for automatic adaptations

Often a system reflects the designer mental model instead of the real users mental model. This could happen when the user are not involved in the early design process, for instance, or if the user needs are not sufficiently taken in consideration. Analyzing real users data can be a solution to discover real user-system interaction. For instance, Web usage analysis is a long process of learning to see a Web site from the perspective of its users [see Spiliopoulou, 2000; Mobasher et al, 2000]. By analyzing Web server log data, for instance, pages that occur frequently together in the same order could be discovered. This may be a signal that many users navigate differently than originally anticipated when the site was designed. The usage mining process can involve the discovery of association rules, sequential patterns, page view clusters, user clusters, or any other pattern discovery methods.

After having reached some evidence confirmed by statistical analysis of collected data, the re-design process of the interface may be accomplished in two ways [Perkowitz and Etzioni, 2000]:

- o by <u>customization</u> → adapting the site's presentation to the needs of each individual visitor based on information about those individuals,
- o by <u>transformation</u> → improving the site's structure based on interactions with *all* visitors.

Between these two alternatives, a third solution could be adopted: personalizing a site according to different cluster of users' behavior (for instance occasional, regular, novice, expert user) emerged from the data mining process.

Perkowitz and Etzioni [2000] proposed a Web management assistant called IndexFinder that processes massive amount of data about site usage and suggests

---

[36] To demonstrate the increasing importance of the empirical evaluation in user modeling, I would to remind the this paper won the Best Paper Award at UM '99.

useful adaptations to the Web master. Given a Web site and Web server logs, the assistant, by applying data mining techniques, creates new "index pages" containing collections of links to related but currently unlinked pages. IndexFinder does most of its work automatically, leaving to the Web master the questions of whether and where the page should be added to the site and how should be titled.

Mobasher et al. [2000] applied automatic personalization based both on usage profiles clusters created as weighted collections of URIs through the analysis of Web logs and page view clusters based on how often they occur together across user sessions. The system then provides a list of recommended hypertext links to a user while browsing through a Web site.

1.5. Future directions

The recent conferences and publications in the field of user modeling and adaptive system have underlined the importance of evaluation in the development of these systems. If this trend is reinforced, the evaluation of user modeling and adaptive systems will contribute to the creation of a corpus of principle and guidelines. The key point is to carry out evaluations conducting to significant results that can be re-used in other research. Only by having many studies showing that certain adaptations work general principles can be extracted. Some reference guidelines are already emerged and can be taken into account in the development of adaptive systems.

For instance, Sears's and Shneiderman's (1994) evaluation on menu voices sorting[37] reported that the users were disoriented by the menu voices sorting based on usage frequency because of the waste of significance in the items grouping. A preferable solution could be the positioning of the first more used voices at the top of the list before all the other ordered items (as in the font list in MS Word).

In the field of adaptive educational system, the adaptive link annotation[38] has been repeatedly evaluated. Different evaluations [Eklund and Brusilovsky, 1998; Brusilovsky, Karagiannidis and Sampson, 2001] reported that even if students seem to understand and like adaptive navigation support features, it did not influence their performance on test. Afterward, other experiments [Weber and Specth, 1997; Specth and Kobsa, 1999] have shown that the adaptive link annotation is of use for students who have some previous experience that is relevant to the subject being learned from an adaptive system. In turn, novices seem to profit from more guided and restrictive

---

[37] The **sorting** is a simple techniques primarily employed for ranking items on the basis of their relevance to user's interest, goals, knowledge (Brusilovsky, 1996, Kobsa, 2001).
[38] **Adaptive link annotation** is a specific adaptive hypermedia technology whose aim is to help users find an appropriate path in a learning and information space by adapting link presentation to the goals, knowledge and other characteristics of an individual user. (Eklund and Brusilovsky, 1998).

methods such as enabling/disabling links or direct guidance with the adaptive "next" link.

When a researcher decides evaluating an (adaptive) system, results of previous evaluation should be taken into account. To assist researchers in this goal and to promote the construction of a guidelines corpus, Weibelzahl and Weber [2002] proposed the development of an online database[39] for studies of empirical evaluations of adaptive systems called Easy-D. The aim of this proposal is to serve as reference for researcher in the field of adaptive systems and guide for planning new evaluations that fulfill the following methodological requirement:

- evaluation layer: according to the framework proposed in [Weibelzahl e Lauer, 2001] a study can be assigned to different evaluation layers (evaluation of input data, evaluation of inference, evaluation of adaptation decision, evaluation of interaction);
- method of adaptation: in which way does the adaptation take place;
- method of evaluation: a short description of the evaluation, using on of the following categories
  - without running system
    - results of previous research
    - early exploratory studies
    - knowledge acquisition from experts
  - studies with a system
    - controlled evaluations with users
    - controlled evaluations with hypothetical (i.e., simulated) users
    - experience with real world use
- data type: brief description of the kind of analyzed data;
- criteria: which were the main criteria and which measures were used;
- criteria categories: one or more of the following categories apply if at least one of the criteria belong to it
  - efficiency
  - effectiveness
  - usability
- n: number of subjects, sample size;
- k: number of group conditions;
- randomization: is the assignment of subjects to groups randomized or quasi-experimental;
- statistical analysis: which statistical methods are used (e.g., MANOVA, ANOVA, correlation);

In addition to the creation a corpus, the empirical evaluation in adaptive system needs reaching more rigorous level in terms of subject sampling, statistical analysis, correctness in procedures and experiment settings.

The future research directions are wide and stimulating and could include:

o   the study of the correlation between the user behavior and the system behavior,

o   the analysis of possible correlations between psychological user features and interface preferences [see Fucs, 2001],

o   the application of qualitative methods of research (interpretative evaluation, grounded theory, embodied interaction, etc) to the evaluation of user-adapted systems,

o   the rigorous and complete application of user-centered design approach to the whole lifecycle of an adaptive system and the application of all the HCI method described in this chapter to the different design phases.

The last points will be discussed respectively in Chapter 2 and in Chapter 3.

---

[39] http://art7.ph-freiburg.de/easy-d/

2. Other evaluation methods and approaches

This Chapter is about methods of evaluation and approaches less usual then those ones explored in Chapter 1, particularly in the field of user modeling and user-adapted systems. The topics mainly concern the so-called qualitative methods (Sections 1 and 4) of research and observational methods (Section 2). Section 5 is devoted to the description of the *embodied interaction,* a new stance toward user-machine interaction.

2.1.   The qualitative methods of research

An immediate and meaningful definition of qualitative research comes from Anselm Strauss, one of developer of the *Grounded Theory*. For Strauss [Strauss and Corbin, 1998, pp 10-11].

"*by the term qualitative research we mean any type of research that produces findings not arrived by statistical procedures or other means of quantification. It can refer to research about persons' lives, lived experiences, behaviors, emotions, and feelings as well as about organizational functioning, social movements, cultural phenomena, and interactions between nations. (…) In speaking about qualitative analysis, we are referring not to the quantifying of qualitative data but rather to a nonmathematical process of interpretation, carried out for the purpose of discovering concepts and relationships in raw data and then organizing these into a theoretical explanatory scheme*".

2.1.1. The origins

The distinction between quantitative and qualitative research originate with the social sciences. To understand these two different points of view, we have to go back to the XIX century when sociology was born. On the basis of the dominant paradigm of that period, the positivism, (i) the existence of the social reality is external to the individuals and (ii) therefore it can be objectively observed and investigated (iii) by methods inherited from natural sciences [Corbetta, 1999]. Following the positivist approach the method of investigation has to be inductive (from the particulars to the universal): through empirical observation and identifications of similar patterns, universal laws can be discovered. These laws are based on cause-effect relationship and their existence is independent from the observers.

During the XX century the positivist point of view has been continuously revised and modified. The two main movements inheriting the positivism perspective are the *neo-positivism* (1930-1960) and the *post-positivism* (from the 1960ies). Their main critics to the original positivism concern [Corbetta, 1999]:

o *the existence of the external reality*: even if the reality is external to the individuals it cannot be perfectly known because (i) human knowledge is imprecise and (ii) its laws are probabilistic and not certain;

o *the objective investigation of the external reality* is not yet sustained because possible interferences can arise from the subjects analyzing the phenomena and from their cultural and social background. Therefore an objective knowledge can be acquired only in a probabilistic way;

o *the methodologies* are quite similar to the original positivist methodologies (observation through experiments, variables manipulation, quantitative interview, etc) but major attention is devoted to inferential statistic and mathematical measurements. The deductive method becomes the main method of investigation (from the theory to the empirical reality) and the positivists establish that the theory validation can be achieved only by falsification: if the empirical data do not falsify the starting hypothesis the theory is confirmed, otherwise is rejected. Moreover, quantitative methods of analysis are taken into consideration and the replication of the experiments is used as a confirmation, since it is assumed that repeated results are true with a higher probability.

While the positivism and its evolution in the neo-positivism and the post-positivism are the foundation of the quantitative research, the *interpretative approach* is fundamental for the qualitative method of analysis. The basic idea of the interpretative approach is that external reality is not merely observed by subjects, but it is also interpreted. Following the famous Dilthey[40] distinction [Dilthey, 1954], while the object of the natural sciences is external to the researcher, the object of spiritual sciences is not detached from the researcher and therefore the knowledge can be achieved only by a particular process: the comprehension (*Verstehen)*. The famous sociologist Max Weber [Rossi, 1958; 1984] transfers this distinction to the sociology without the individualism of Dilthey: even if we can only comprehend the object of social science, this comprehension has to be objective, with no evaluation. For Weber, social sciences are different from natural sciences, because the former are more individual-oriented. The social researcher has to interpret the human actions and try to understand the subjective meaning that the individuals give to their actions. However, how an objective knowledge can be achieved starting from the subjective actions of an individual? For Weber the solution is the *ideal type,* an abstraction generated by the empirical uniformity such as recurrent behaviors present in the social actions. Several movements, which will be described in this chapter, such as

---

[40] The German philosopher Wilhelm Dilthey was the first promoter of the autonomy of the human science. For Dilthey we explain the nature, while we comprehend the psychical life.

phenomenology, ethnomethodology, symbolic interactionism will inherit Weber's point of view. While Weber always carried out his research under a macro-sociology point of view, analyzing phenomena like economy, religion power, etc, from the 1930s until the 1960s, the sociologists of the famous Chicago School developed their research under a micro-sociology point of view. They sustain that the individuals' interpretations and their reciprocal interactions generate the society structures. Therefore the interaction between the individuals has to be studied in order to understand the society structures. As a consequence, they give particular attention to a new field for the sociological analysis: the every-day life of every single person.

The qualitative methods differ from the quantitative methods, [Corbetta, 1999] for:

- o *the existence of an external reality*: the objective universal reality does not exist anymore because every subject generates her personal reality. Therefore, since there are multiple interpretation perspectives the existence of multiple external realities is assumed;

- o *the objective investigation of the external reality* disappears. Social research is defined as an interpretative science looking for a meaning instead of an experimental science looking for (universal) laws [Geertz, 1973]. Therefore the goal of the research becomes the understanding of individual behavior;

- o *the methodology*: the interaction between the researcher and the object of the research becomes now fundamental. Since the goal of the researcher becomes understanding *what meanings the subjects give to their actions*, the methods of research are strictly qualitative and they can change over different studies.

	Quantitative research	Qualitative research
Research setting	o Theory and its formulation in an empirical model, research design, data survey, data analysis, return to the theory → **sequential steps** o **Deductive method** → from the theory to the observation. The empirical data have to sustain the theory o The (**definitive**) **concepts** are translated into variables o **Controlled experiments** by variable manipulation o The researcher and the subjects do not physically interact o The research subjects are passive	o The theory comes out from the data → **inductive method** o The study of related theories is avoided because of its influence o The relation between theory and research is open and interactive o Exploitation of **sensitizing concepts** that can orient the approach to empirical data o **Participant-observation** → mere observation of the reality without any kind of interference or manipulation o The researcher and the subjects physically interact and their interaction is the basis for the comprehension o The research subjects have an active role in the research
Data survey	o The **research design** is decided before and it is strictly structured o The **sample** has to be statistically representative o The data are collected in the same way o The collected data are **hard**: they have to be objective and standard	o The **research design** is open and changing during the survey o The **subjects** have to be representative under a sociological point of view. The number of subjects is not important o The collected data are not homogeneous o The collected data are **soft**: their importance is related to their richness and depth
Data analysis	o **Statistical** and **mathematical** analysis o Exploitation of **dependent** and **independent variables** o Goal of the analysis: explanation of the **variance** of independent variables o Analysis **variable-based**	o Analysis **case-based** o Case **classification** o Goal of the analysis: **comprehension** of the subjects under their point of view
Results	o The data are presented in **tables** (**relational perspective**) o Discovering of **causal relationship** between the variables o The results can be **generalized**	o The data are presented by **narration** (**narrative perspective**) o Discovering of **classifications** and **typologies** o New interpretation of the reality (after the classification process) o The case studies can be examined in depth

Table 2.1. A comparison between quantitative and qualitative methodologies (Corbetta, 1999).

2.1.2. The user-involved methodologies

2.1.2.1. **The participant-observation**

In social sciences, and in particular in the field-study research, *participant-observation* is a qualitative method of research that requires the direct involvement of the researcher with the object of study [Corbetta, 1999]. During participant-observation the researcher immerses her in a new world with the goal of exploring the members' point of view.

Following this research approach, *the researcher inserts herself in a direct way and for a relatively long period of time in a defined social group and in its natural environment living a personal interaction with the group members in order to describe their actions and to understand their motivations by identifying herself with them* [Corbetta, 1999, p. 368].

The two main underlying principles of the participant-observation sustain that *i)* a complete social knowledge can be achieved only by comprehending the point of view of the social actors by living *with* and *like* them; *ii)* the identification with the group members can be achieved only by living with them for a long period of time and by interacting constantly with them.

Consequently,

- o the observation can be carried out only by the researcher in the first person;
- o the interaction period has to be sufficiently long;
- o the environment of the interaction has to be the natural environment of the social group under observation and not an artificial lab environment;
- o the researcher is not a mere observer but she has to interact with the observed subjects and to participate to their life;
- o the research goal is the comprehension of the reality under the observed-subjects point of view.

The origin of the participant-observation comes from anthropology, and particularly from Bronislaw Malinowsky, the father of modern anthropology and the dominant figure in developing the role of ethnography[41] in anthropology. Malinowsky criticized the methods of the XIX century anthropology and introduced the principles of modern anthropology. For Malinowsky, the anthropologist has to understand the indigenous populations by reaching their point of view and their vision of the world. Instead of considering the indigenous populations wild and primitive and writing detached reports while sitting in an office, the anthropologist can comprehend the populations

---

[41] Observations in the "real world" - contextual understanding.

only by living and interacting with them for a long period of time, as Malinowsky did living for a long period of time in the New Guinea. Malinowsky established the ethnographic field work as the dominant paradigm for anthropological research. For more details, see Malinowsky [1922].

Sociology and anthropology are closely related, and in some cases almost overlapping. Clifford Geertz [1973] suggests that whereas sociology examines the emergence and maintenance of social structures and patterns of social interaction, anthropology studies the cultural webs of signification that give those structures and interactions meaning. In fact, Malinowsky's point of view was successfully applied to the study of modern society. In sociology, for instance, the Chicago School realized a set of studies about the deviance in the American society by applying Malinowsky's methodology (the researcher lives together with the subjects of the study and observes their interactions, etc). The Chicago School sociologists adapted ethnographic, participant-observer approaches and their main topics of interest were subcultures on the fringes of ordinary society such as alcoholics, jazz musicians, drug users, etc.

As the ethnography emphasizes the detailed understanding of a culture, through an intensive and long-term involvement - what the anthropologist Clifford Geertz calls *thick description* [Geertz, 1973] -, the participant-observation requires the immersion of the ethnographer in the culture in question:

"*the central element is to explore the member's own view of his or her life and culture. That implies the need to be able to describe not just what the members of that culture do but what they experienced in doing it; why it is done and how it fits into the fabric of their daily lives* [Dourish, 2001a, p. 59]."

In the participant-observation method the researcher can declare her role of observer or hidden her real goals. The main reason to hide the observer's role is that people behave in a different way when someone is observing them. However, in most cases it is necessary to reveal the observer role in order to complete the research goals (i.e., if the researcher has to ask explicit questions).

The action of the participating observer has to be selective: the object of the observation has to be decided by the theory, even if, for some qualitative paradigm all the existing theory has to be ignored in order to discover the theory during the research without any kind of influence. However, sometimes the richness of the empirical data could overload the researcher and some theoretical guide becomes necessary.

An important point in participant-observation is that the object and the interest of the research can change during the development of the research, and it is not decided *a priori.* Some possible objects of observation could be [Corbetta, 1999]

- o the physical context,

- o the social context,
- o the formal interactions,
- o the informal interactions,
- o the interpretations of the social actors,
- o ...

The observed data can be recorded by writing notes, by audio-videotaping, etc. and this process is particularly important because it is hard to remember all the observed things without any external helps. Moreover, our memory is extremely selective and thus the material could be remembered in an altered way.

The final and most important stage of the participant-observation is the analysis of the empirical data. At this point the participating observer has to mix two points of view of her analysis: the insider perspective, derived by her participant-observation, and the outsider perspective, needed to highlight the aspect still unknown to the observed subjects. Since in the data gathering methods of qualitative research there are non-standard procedures, a successful analysis depends on the personal capability of the researcher. The data analysis should be a continuous process starting during the observation.

The first step of the analysis is the already mentioned *thick description,* "a description enriched by meaning and interpretation in a cultural and historical context, in a network of social relationship" [Geertz, 1973]. In addiction, interviews and the documentary analyses can be exploited.

The second step regards the classification: the individuation of *types* and the construction of *typologies.* An easy way of classifying could be the discovering of temporal sequences. Another way to classify is to put in order different social objects, for instance, by grouping them in classes on the basis of similar o dissimilar features.

The third step concerns the discovering of typology dimensions in order to reach the *ideal types* (or the *cultural themes,* the main lines of a cultural paradigm) by singling out the main features of the types.

The participant-observation can be applied to the study of every human activity and to the study of every human group when the researcher is interested in the discovery of their inner points of view. In sociology the participant-observation has been applied to the study of the communities and the sub-cultures.

The main critics moved to the participant-observation are [Corbetta, 1999]:

- o *the subjectivity of the researcher,* even if for the qualitative researchers the subjective involvement is the only way to reach the comprehension,
- o *the non - generalization of the results*, due to the involvement of too few subjects,

     o   *the non - standardization of the exploited procedures.*

The exploitation of the participant-observation highlights some interesting findings: the daily habits and the acts of the everyday life are rich of hidden meanings. Thus, it becomes interesting to understand which meanings the subjects give to their actions. The importance of the meanings and the interpretations has been the key point of another sociological movement, the *symbolic interactionism.* Herbert Blumer, who originated this paradigm, sustains that people behave on the basis of the meaning they give to things and to other people [Blumer, 1969]. The meanings are learnt during the social interaction and the whole meanings create the culture and every subject interprets the social facts on the basis of her culture. Thus, the meaning can be discovered by observing the individuals and their actions. This movement towards the daily life was also carried on by Erving Goffman and his studies about the rituals of social interaction aimed at the discovering of models. For instance, Goffman [Goffman, 1959] exploits the famous theatre metaphor for human life: the life consists of different performances with actors and audiences, with "front-stage" and "back-stage" behaviors (i.e., people behave in a different way when they are working or they are out for dinner with friends).

On the side of the analysis of everyday life can be collocated also the theory of the ethnomethodology by Harold Garfinkel [Garfinkel, 1967]. Preece [Preece et al., 1994] defines ethnomethodology a method that assumes no *a priori* model of cognitive process when a person does something, but instead analyzes behavior by observing events in their natural context. Ethnomethodology refers to the analysis of commonsense methods that people exploit by making their common actions. People invoke these methods as practical solutions to practical problems to render the world sensible and interpretable in the course of their everyday actions.

For the ethnomethodologists, the daily actions are managed by precise and implicit rules that deal with human interaction. In order to discover these rules, the ethnomethodologists proposed to break these implicit laws. Hence the non-convectional ethnomethodology experiments, such as talking too close to an extraneous person, or drinking from another's glass during a party. And the obtained disoriented reactions of the involved subjects confirmed the existence of such rules. The application to ethnomethodology techniques to HCI design will be discussed in 2.5.2.2

2.1.2.2.   <u>Sociology in HCI and Interpretative Evaluation</u>

As outlined by Dourish in [2001, pp. 61-64] the first appearance of sociology in HCI was the social psychology. Social psychology is concerned with how an individual's

thought and emotions are affected by interactions with others. The social psychology applied to HCI is interested in how these interpersonal relations could be manifested in communications mediated by computers systems.

As cognitive psychology lends methods to understand the cognitive implications of particular forms of design, sociology lends methods to understand the settings in which computer systems would be deployed, and the ways in which they would both affect and be affected by those settings. Methods, like ethnography, could be used to gain detailed understandings of how work is conducted, with particular attention to the context of activity.

The use of ethnographic materials is common in system evaluation. Supporters of socially based study have found that ethnographic approaches can uncover requirements for a system design through the detailed observation of the work settings. In contrast, more traditional approaches, such as laboratory-based usability studies, tend to be disconnected from the lived detail of the work. From an ethnographic perspective the usability methods are meaningless since they are decontextualized and examined in the sterile confines of a laboratory. Ethnographers look for a more direct engagement and they desire to take a broader view of the relationship between technology and work by understanding how a software system features as part of a set of working practices.

On a more analytic level, ethnographic methods are used to analyze *work processes* and *work practices*. *Work processes* are formalized or regularized procedures by which work is conducted (procedures for authorizing payments, for ordering supplies, etc). Work processes are captured and codified in rulebooks, manuals, information systems, etc. In contrast ethnographers in HCI have frequently drawn attention to *work practice*, the informal but nonetheless routine mechanisms by which these processes are put into practice and managed in the face of everyday contingencies. The ways in which people may deviate from formalized procedures tend to reflect a better or more fruitful adaptation of the process to the specific circumstances in which the activity is carried out. This is particularly relevant for the development of information systems where designers presume that the formalized work processes constitute a perfect description of what actually goes on. They are encoded into software systems without accounting for which they will be put into practice.

One of the most famous ethnographic investigations carried out in computer science and in particular in the domain of CSCW (Computer Supported Collaborative Work) is the workplace study into air traffic control conducted by a multi-disciplinary group of sociologists and computer scientist and focused on the link between work and setting of the work . For more details see [Hughes et al, 1995].

Sustainers of field observation studies criticize laboratory studies since they do not occur in actual use. However, also in case of field studies the situation is not completely natural since the subject are likely to be influenced by the presence of evaluators and/or recording equipment.

Preece et al. [1994] classify the ethnographic investigations under the umbrella term *interpretative evaluation*. The interpretative evaluation can be best summed up as "spending time with users" and it is based on the assumption that small factors that go behind the visible behavior greatly influence outcomes. Since lab conditions are not real world conditions only observing users in natural settings can detect the presence of these factors.

The interpretative evaluation comes in these flavors:

- o <u>contextual inquiry,</u> which is a semi-structured interview covering whatever interesting aspects is recorded in order to be elaborated by both the interviewer and by the interviewee. Usually the attention is focused on the context where the action takes place [Holtzblatt e Beyer, 1998];

- o <u>cooperative and participative evaluation</u>. The cooperative evaluation includes methods where the user is encouraged to act as a collaborator in the evaluation to identify usability problems and their solutions. One cooperative method is the "think aloud protocol" [see 1.1.2] which allows the user to ask questions, comments and suggest appropriate alternatives of the evaluators, and the evaluators to prompt the user (Monk et al., 1993). The participative evaluation is more open and subject to more user control than cooperative evaluation. It is strictly tied to participatory design techniques (user involved in the design phase) and applied methods such as focus group (see 1.1.1 and [Greenbaum and Kyng, 1991]);

- o <u>ethnography.</u>

2.2. *Observing interaction: the sequential analysis*

This paragraph introduces a new kind of observation, the *systematic observation*. The researchers following this approach are not directly involved with the object of their study, as in participant-observation, but they observe the object of study in a detached way.

The systematic observation can be defined as a "*particular approach to quantifying behavior. This approach is typically concerned with naturally occurring behavior observed in a real context*" [Bakeman and Gottman, 1986, p. 4]. The aim is to define before various forms of behavior (behavioral codes) and then asks observers to record

whenever behavior corresponding to predefined codes occurs. A major concern is to train observers so that all of them will produce an essentially similar protocol, given that they have observed the same stream of behavior.

The observation can be analyzed by adding nonsequential or sequential techniques. Nonsequential systematic observation can be used, for instance, to answer questions about how individuals distribute their time among various activities, while sequential techniques can be used to answer questions as how behavior is sequenced in time and how behavior functions moment to moment. Thus, sequential methods result more suitable for the analysis of social interaction.

In nonsequential analysis the subjects are observed for given time slots during time intervals. For instance, subjects can be observed for one minute each day in different time. This method is called "the method of repeated short samples" or "time sampling".

In a famous nonsequential experiment, the Parten's study of children's play [Parten in Bakeman and Gottman, 1986], each child was observed for one minute each day. The order of observation was determined in advance and was varied systematically One the average, children were observed about 70 different times, and each time they were observed, their degree of social participation was characterized using one of these six codes: Unoccupied, Onlooker, Solitary, Parallel, Associative, Cooperative. Weights were assigned to each behavioral code and then multiplied for the percentages representing the amount of time the subjects devoted to each codes. Then, the resulting products were summed for each subject and then correlated with the child's age and IQ. This study typifies a sort of "time-budget" information, a kind of distribution of activities during time.

In sequential analysis, each subject is observed for a given period of time and then behavioral codes are assigned. For instance, every subject is taped for 100 minutes, and then observers view the tape and decide which of the behavioral codes best characterize each successive 15-second interval. For details and examples see [Bakeman and Gottman, 1986]

2.2.1. Developing a coding scheme

The first step in observational research is developing a *coding scheme (or schemes)*.

Developing coding scheme (making distinction, categorizing, developing taxonomies) is a common intellectual activity. If the scheme is well constructed a clearer view of the world should emerge. A coding scheme is usually an informal hypothesis and the entire study is the testing of that hypothesis. The most important thing before defining a coding scheme is beginning with a clear question. Bakeman and Gottman [Bakeman and Gottman, 1986] did not suggest determining all the

behavioral codes before collecting data of interest. Even if to avoid a hypothesis-generating approach and looking at everything, it is essential to look for consistency across the study and starting with a coding scheme. However, if unexpected behaviors emerge during the observation, the scheme has to be modified.

Schemes can be *physically* or *socially based*. The former ones classify behavior with clear and well understood roots in the organism physiology, while the latter ones require the observer to make some inferences about the individual observed and their use depend on social process.

It is important to keep the coding scheme simple and to have clearly distinct codes. Codes must be *mutually exclusive and exhaustive*. This means that only one code can be associated with a particular event (mutually exclusive) but there is some code for every event.

2.2.2. Recording behavioral sequences

Behavioral sequences can be analyzed by means of behavioral recording scheme. The main recording scheme proposed by Bakeman and Gottman are [1986]:

- ○ <u>event recording</u>: when observers are asked to code when events occur. If time information is required the observers are also asked to record the onset and offset time of events;

- ○ <u>interval recording</u>: when observers typically record at certain predetermined times. The period is usually divided into a number of relatively brief intervals (10-15 seconds or so). Observers then categorize which codable events are occurring during each interval;

- ○ <u>cross-classifying events:</u> observers does not simply classify events (on a single dimensions) but instead cross-classify them (on several dimensions);

- ○ <u>time sampling</u>: observing is intermittent, not continuous. Repeated noncontiguous brief periods of time are sampled and something about them is recorded.

Whether behavior is observed live or videotaped does not matter. The data can be recorded by using pencil and paper or electronic devices (observers press keys corresponding to the appropriate behavioral codes).

While event recording and cross-classifying events is more suitable for sequential analysis, interval recording and time sampling are more useful for nonsequential analysis.

2.2.3. Assessing observer agreement

In observational studies, especially with "socially" based coding schemes, becomes particularly important to evaluate observer. First of all, the data coded by the observer could be influenced by hers subjective vision of the world under analysis. To solve this basic problem, the hypothesis under investigation is usually kept unknown to the observer. Moreover, more than one observer is usually exploited and then their agreement and their performance are evaluated to assure the accurateness and the replicability of their procedures and to calibrate their (eventually) different points of view. Finally, when the observers are being trained, they are evaluated to provide feedback.

Computation of conformity statistic may help to demonstrate the conformity of the results. For instance, if conditional probabilities[42] are analyzed, it is sufficient to demonstrate that data derived from two different observers yield similar conditional probabilities.

For instance, one frequent index of observer agreement is the "percentage of agreement" [Bakeman and Gottman, 1986], whose the most general form is defined as follows

$$P_A = \frac{N_A}{N_A + N_D} \times 100 \qquad\qquad [2, 1]$$

where

  P_A refers to the percentage of agreement

  N_A refers to the number of agreements

  N_D refers to the number of disagreements

In any given application, the investigator would need to specify the recording unit used (events or intervals), which is after all the basis for determining agreement and disagreement, and exactly how agreement and disagreement are defined.

However, when training of observers is the primary consideration or an investigator has sequential concerns in mind, point-to-point agreement may be demanded. If point-to-point agreement is assured, it can be generally assumed that scores derived from raw sequential data (like conditional probabilities) will also agree.

An agreement statistic that does correct for chance is Cohen's kappa (Cohen, 1960), which is defined as follows

---

[42] **Conditional probability** is the probability with which a particular "target" event *B* occurred, relative to another "given" event. *A,* p(*B/A*).

$$k = \frac{P_0 - P_c}{1 - P_c} \qquad\qquad [2, 2]$$

where

P_0 is the proportion of agreement actually observed and is computed by summing up the tallies representing agreement and dividing by the total number of tallies,

P_c is the proportion expected by chance and is computed by summing up the chance agreement probability for each categories.

2.2.4. Representing observational data

Observational data can be represented as follows [Bakeman and Gottman, 1986]:

o event sequences codes the events, ordered as they occur;

o time sequences pairs each event in an event sequence with a number indicating how long that event lasted;

o time-frame data allows for events to co-occur. Each "frame" represents a time interval and indicates the code or the codes that were occurring during it;

o cross-classified events can be represented only in this way: each line represents an event, each column a major category.

Usually data are represented in a way that depends on the goals of the analysis.

2.2.5. How to analyze sequential data

The analysis of sequential data is concerned with how to derive useful descriptive scores from sequential data. The very basic, but useful statistics for describing sequential observational data are [Bakeman and Gottman, 1986]:

o rates (or frequencies): how often a particular event of interest occurred;

o simple probabilities (or percentages): what proportion of all events were of a particular kind (event based) or what proportion of time was devoted to a particular kind of event (time based);

o mean event durations are computed by dividing the amount of time coded for a particular kind of event by the number of times that event was coded.

o transitional probability is simply one kind of conditional probability see (footnote 42) that captures sequential aspect of observational data. It is distinguished from other conditional probabilities in that the target and the given events occur at different times. Often the world "lag" is used to indicate this displacement in time. For example, if data are represented as event sequences, given the event *A*, of the target event *B* occurring immediately after

(lag 1), occurring after an intervening event (lag 2), etc. These event-based transitional probabilities can be written p(B_{+1}/A_0), p(B_{+2}/A_0), etc. Similarly, if data are represented as time sequences (or as time-frame data), given event *A*, of the target event *B* occurring in the next time interval p(B_{t+1}/A_t), in the time interval after the next p(B_{t+2}/A_t), etc.

Of course, all the scores derived from observing behavioral sequences can be used as input for whatever inferential statistics.

2.2.6. Analyzing event sequences

Investigators use to represent the collected data as sequences of coded events. One approach to analyze the event sequences is to define particular sequences of some specified length, then categorize and tally all sequences of that length and report the frequencies and probabilities for those particular sequences - Bakeman and Gottman call this approach "absolute", [Bakeman and Gottman, 1986].

A **z-score binomial test** or a **Chi square test** (see 1.1.5.1) can be used to measure the extent to which an observed frequency (or probability) for a particular sequence exceeds its expected value (frequencies or probabilities are expected because they follow from some assumption made previously by the investigators..

The *z*-score is defined as follows[43]:

$$z = \frac{x - NP}{\sqrt{NPQ}} \qquad [2, 3]$$

where

x is the observed frequency

NP is the expected frequency (*N* is the total number of event sequences and *P* is the expected probability of event sequences)

$\sqrt{NPQ}$ is the estimated standard deviation for the differences between observed and expected frequencies (*Q* is equal to 1-*P*)

If the *z*-score is to be tested for significance, its computation should be based on sufficient data. The rule of thumb for how many events have to be coded in order to assign significance in a computed *z* score is described in [Bakeman and Gottman, 1986].

In theory, absolute methods apply to sequences of any length. In practice, the number of possible sequences increases dramatically as longer and longer sequences are considered and expected frequencies for a particular sequence may be too small to

---

[43] The square root of chi-square with one degree of freedom is equivalent to *z*.

justify assigning significance (or because the number of occurrences for a particular sequence may be so few). Another problem concerns the number of codes defined. In general, when there are more codes, the expected frequencies for particular sequences are likely to be smaller, and hence more data will be required.

Even when z-score computation are based on sufficient data, the problem of type I error - of claiming that sequences are "significant" when in fact they are not (see 1.1.5.1) - remains. In any case, interpretation of results should take into account that some of the apparently significant findings are in fact due simply to chance. In case of large number of tests some techniques should be used to control the "studywise" type I error rate. For more details see [Bakeman and Gottman, 1986].

2.2.7. Analyzing time sequences and cross-classified events.

When successive intervals have been coded, or when event times have been recorded the result is time-sequence data. Time-sequence data are useful when investigators want to know how time was distributed among various activities. In general all the analytic techniques that apply to event sequences can be applied to time sequences as well, bur there are some cautions.

Cross-classified categorical data can be analyzed with what is called "log-linear modeling approach" whose results can be expressed in familiar analysis of variance. For details both for time-sequence data and cross-classified categorical data see [Bakeman and Gottman, 1986].

2.3. Coding for a Collaborative Information Visualization Experiment

From January 2002 to March 2002 I was visiting student at the Department of Computer Science at the University of California, Irvine. In that period I worked with my sponsor, professor Alfred Kobsa, in an experiment aimed at evaluating two different information visualization systems (for details on information visualization systems see [Card et al., 1994; Chen, 1999]), InfoZoom and Spotfire – software helping user to do visual data mining using visual representations - in different experimental conditions. The final goal of the investigation was to examine collaborative and individual decision-making about data using two different information visualization systems. One of the aspects of the investigation was to observe the subjects' interaction and to code their behavior in order to find evidences to falsificate the experimental hypothesis. After having introduced the two systems exploited during the experiment and the experimental setting, I will describe the behavioral codes decided before starting the sequential analysis of the recorded experiments in order to have a practical example of the issues discussed in the previous Section.

2.3.1. InfoZoom

Infozoom [Spenke et al., 1996; Spenke and Beilken, 1997] is an information visualization system that has three types of presentation styles:

Figure2. 1. Infozoom: wide view of the Car data set used during the training.

o <u>Wide View</u> (Figure 2.1), for individual objects. The Wide Table view is the typical format of spreadsheets and printed tables.
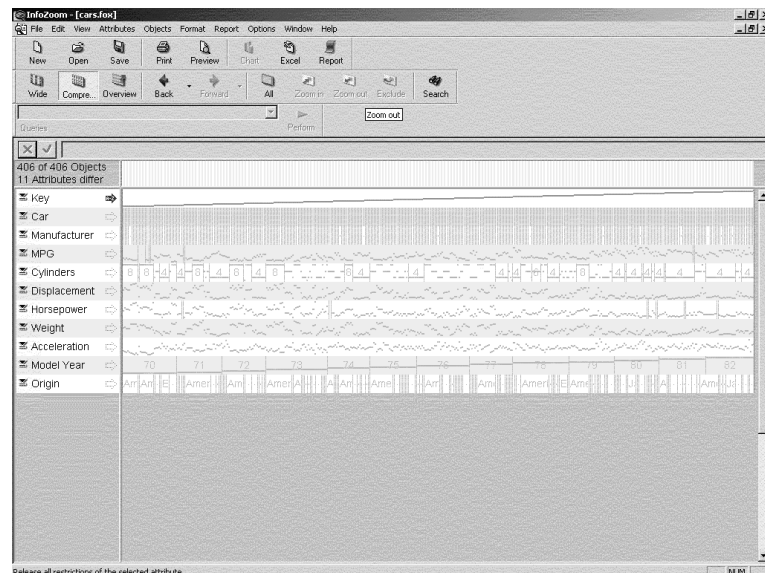
Figure 2.2. Infozoom: compressed view of the car data set used during the training (ordered by Model Year)

o <u>Compressed View</u> (Figure 2.2), to explore in a single view an entire data set. In the Compressed Table the view of the entire data sets is fitted to the width of the Infozoom window. Each row can be sorted in ascending or descending order and the values of other rows are being resorted accordingly.
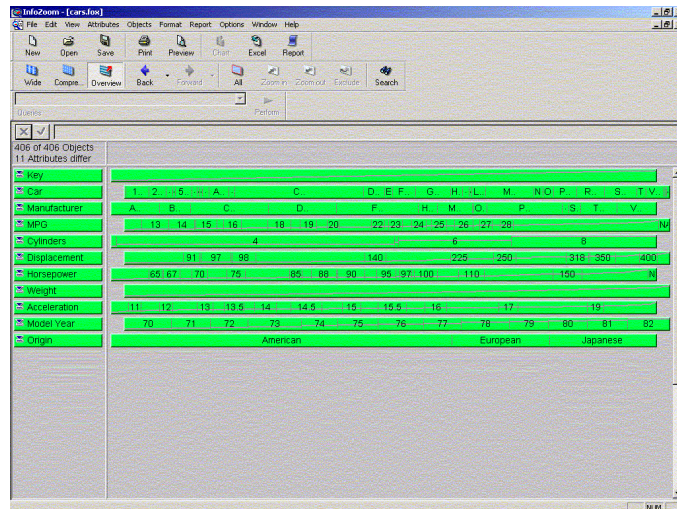
Figure 2.3. Infozoom: overview of the Car data set used during the training.

o <u>Overview</u> (Figure 2.3), to explore attribute categories. In this view, each row represents an attribute and its distribution and there is no correspondence between the different attributes' distributions plotted in different rows.

Within each presentation style, there are four important functions:
o <u>zooming in</u>: focuses on specific part of the database in relation to the rest of the database;
o <u>zooming out</u>: is the default setting that allows viewing all information present in the database;
o <u>sorting</u>: from least to greatest value and from least to greatest;
o <u>graphing</u>: plotting attributes of specific objects onto many graphs.

In InfoZoom is also possible deriving new attributes, such as average value, minimum or maximum value, etc.

2.3.2. *Spotfire*

Spotfire [Ahlberg and Wistrand, 1995] is an information visualization system made up of three components within the application [Figure 2.4]. Each one performs a different function for enabling the user to view data in different ways:
o <u>graphical view</u>: offers familiar graphical visualizations, such as scatterplot, histogram, pie chart, bar chart;

- o <u>query</u>: enables the user to define which cases to include/exclude by means of widgets such as sliders, checkboxes, radio buttons, etc;
- o <u>details on demand</u>: allows the user to click on any data point in the graphical view, and see its details.
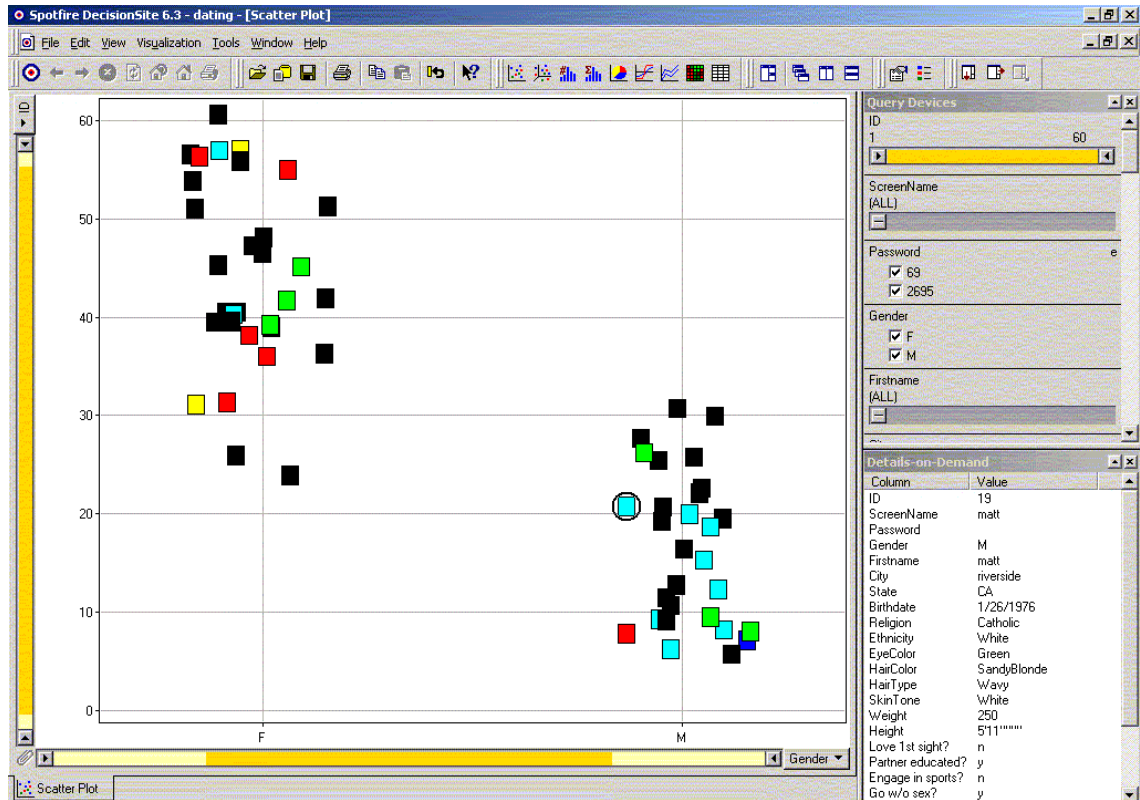
Figure 2.4. Spotfire: scatter plot's screenshot of the Dating data set used during the Focused Question Task of the Experiment (on the x axis Male and Female attributes are selected, while the Detail-on-Demand window shows a subject's details).

For each visualization, two variables can be selected to display x and y coordinates and other few additional variable can be selected through a dialog window.

2.3.3. The experiment

The leading experimental hypothesis was to investigate whether and how collaborative and individual decision differ when different information visualization system are used. The kind of collaboration under study was distinguished in remote and collocated and also the two information systems were characterized by different features: more specifically, a previous research [Kobsa, 2001] showed that InfoZoom is considered more transparent (since it invoke a an easy-to-understand system image in the users [Preece et al., 2002]) than Spotfire in term of visual representation and functionality.

One hundred undergraduate students with majors in Information and Computer Science or Engineering at the University of California, Irvine participated in the experiment. They received $25 for their participation and competed for a $100 prize for the best results in the discovery task of the experiment.

Two distinct types of tasks to test were chosen (see [Mark et al., 2002] for the complete experiment description):

- o <u>a focused question task</u>, where subjects were asked to use data from an online dating service to answer ten specific questions;
- o <u>an open-ended discovery task</u>, where subjects were given 40 minutes and instructed to discover as many findings as they could in a population survey data set.

The experiment used a two-factor between-subjects design. The factors were:

1) System - two factor levels:

    a) <u>InfoZoom</u>: subjects used the Infozoom system in both tasks;

    b) <u>Spotfire</u>: subjects used the Spotfire system in both tasks.

2) Interaction type - three factor levels:

    a) *Alone*: subjects sat at a workstation by themselves;

    b) *Remote:* subjects sat at workstations in adjacent rooms. They performed the task while interacting via Microsoft NetMeeting and a speaker phone. They used either InfoZoom or Spotfire as shared applications. Subjects did not see each other;

    c) *Shared Electronic Whiteboar*d: Two subjects worked side-by-side in front of a large 60" diagonal touch-sensitive electronic Whiteboard (Smart-Board), using either InfoZoom or Spotfire.

Subjects were randomly assigned to one of the combined conditions of System and Interaction Type.

In all conditions, subjects first received a 30-minute training on their visualization system. During the first 10 minutes of the training, subjects received a general demonstration of the main system functionality, followed by a 20-minute hands-on tutorial using six questions from a car statistic data set. After the training, the subjects were familiarized with the content of the two datasets to be used in the tasks. Subjects in the Remote and Electronic Whiteboard conditions additionally practiced using these systems. The first experimental task (focused questions) took 30 minutes and the subsequent free discovery task took 40 minutes. Subjects then filled in a short questionnaire and were interviewed about their experience with the system.

2.3.4. The experimental results

In the first task, the measure taken into account was the correctness of responses based on subjects' written answers on paper, while in the second task, the number, accuracy, and significance of findings, also based on their written descriptions.

The final showed that people who worked in groups were more correct in their answers for objective questions, based on searching a large dataset results (for more details see [Mark et al., 2002]). These results held for the more transparent system, InfoZoom, but not for Spotfire. In the second task, groups were more accurate in their results for a free data discovery task. Again, these results held for the more transparent system only. Subjects using this system also produced results that were higher in complexity but judged lower in importance. Groups and individuals did not differ. Mark, Kobsa and Gonzales [Mark et al., 2002] suggest that given the right visualization system, groups do better than individuals in finding more accurate results, but not necessarily increased or more meaningful results.

Though the results did not yield differences between Remote and Electronic Whiteboard conditions, it is important to keep in mind that in this study are only present quantitative measures. Strong differences in group processes, due to the different physical proximities of the participants, could be present. Furthermore, the Electronic Whiteboard provides an immersive and social experience that is absent in the remote condition.

2.3.4.1. **The proposed coding schemes**

To truly understand the difference conditions where the experimental interaction took place, Mark, Kobsa and Gonzales [Mark et al., 2002] suggest a qualitative study of the group processes using qualitative measures, e.g. video analysis of the recorded experiments.

Since I was there my coding task concerned to check the correctness of focused questions by analyzing the questionnaires. We have also discussed together about the aim of sequential analysis and the possible coding schemes.

My first proposal for coding scheme is describes below.

Infozoom codes. These codes are merely aimed at coding the possible actions the user can perform when solving the experimental tasks using Infozoom.

- zoom in
- zoom out
- exclude attributes
- overview
- compressed view

- wide view
- back (undo)
- all (including all attributes)
- insert derived attribute
- sorting a-z
- sorting z-a
- graphing
- query
- search
- excel (possibility to view data in excel)
- report
- help

SpotFire codes. These codes are merely aimed at coding the possible actions the user can perform when solving the experimental tasks with SpotFire.

- scatter plot 2d
- scatter plot 3d
- histogram
- bar chart
- pie chart
- line chart
- profile chart
- heat map
- excel table view
- query (which attributes does the subject select?)
- x-y-z axis (which attributes does the subject select?)
- zoom
- details on demand
- binning
- panning
- properties (request of properties window)

NetMeeting codes. In the Remote situation subjects using NetMeeting could perform the following coded actions:

- request control
- release control

SmartBoard codes. In Electronic Whiteboard situation subjects using the SmartBoard could perform the following coded actions:

- (using the) pen tray
- touching (the whiteboard)

This first group of codes is aimed at checking the actions performed when subjects interact with the programs. However, these kinds of codes cannot be considered meaningful for a social interaction analysis. Since this study was aimed at discovering the differences between collaborative and individual decision-making using two different information visualization systems in different interaction context, the codes have to be re-oriented at detecting these differences. Therefore, the following codes (and the questions that generate them) were discussed together with the other members of the Collaborative Information Visualization group and try to satisfy these purposes. After my departure, a sequential analysis of the recorded experiments had been planned.

Interaction codes
- cooperation - non cooperation (Do the subjects share the tasks to perform or they work together? Are there difference between the two software? And between the remote and face-to-face interaction?)
- agreement - disagreement (Between the two subjects)
- participation (Does one of the two subjects assume a dominant position or is the participation equal? Which subject works more?)
- task conversation - other conversation (Are they having a task related conversation or not?)
- coordination statement (Which are the coordination statements?)
- remote coordination practices (Which are the coordination practices in the remote situation? Do interaction patterns exist?)
- face-to-face coordination practices (Which are the coordination practices in the face-to-face situation? Do interaction patterns exist?)
- calculate how many discoveries they make every five minutes
- (coding the) views (How many view do they exploit to reach a conclusion? Which view do they use more?)

While most codes are suitable for event recording (or cross-classifying events), only the task concerning the calculation of discoveries adopts a interval recording approach.

2.4. The Grounded Theory

The *Grounded Theory* is a qualitative research methodology developed by Anselm Corbin (who studied at the University of Chicago) and Juliet Strauss. The *Grounded Theory* is "a theory derived from data, systematically gathered and analyzed through the research process. In this method, data collection, analysis and eventual theory stand in close relationship to one another. A researcher does not begin a project with a preconceived theory in mind (...). Rather, the researcher begins with an area of study and allows the theory to emerge from the data [Strauss and Corbin, 1998, p. 12]".

In the introduction of their book "*Basic of qualitative research*" [Strauss and Corbin, 1998], the authors, before introducing the bases of their theory, outline the characteristics of a grounded theorist. She must have:

- o the ability to step back and critically analyze situations,
- o the ability to recognize the tendency toward bias,
- o the ability to think abstractly,
- o the ability to be flexible and open to helpful criticisms,
- o sensitivity to the words and actions of respondents,
- o a sense of absorption and devotion to the work process.

A grounded theorist has to possess flexibility and openness and must be able to manage ambiguity. However, all these features are not relevant if the researcher does not develop a new way of thinking about data in the world were she lives. The importance of this methodology is that it provides a sense of "*vision, where it is that the analyst wants to go with the research*" [Strauss and Corbin, 1998, p. 8] with the final goal of for gathering knowledge about the social world.

In the Grounded Theory methodology data are collected using the same techniques of other research methodologies. Data may be qualitative or quantitative or combination of both types since the authors advocate an interplay between qualitative and quantitative methods.

Strauss and Corbin sketch out the three major components of qualitative research of their methodology;

1. data (interviews, observations, documents, records, films etc);
2. procedures, used to interpret and organize data. These usually consist of conceptualizing, reducing, elaborating, and relating data (all these procedures are often defined as coding) by asking question and making comparison;
3. written and verbal report.

Particularly, the second point is more developed and further divided in three important steps that are the core component of the Grounded Theory:

1. *open coding*: the analytic process through which concepts are identified and their properties and dimensions are discovered;
2. *axial coding*: the process of relating categories to their subcategories, termed "axial" because coding occurs around the axis of a category, linking categories at the level of properties and dimensions;
3. *selective coding*: the process of integrating and refining the theory.

2.4.1. The open coding

The discovery and the analysis of concepts is the focus of open coding. The authors define a concept as "labeled phenomenon".

Indeed, during open coding data are broken down into discrete parts, closely examined and compared for similarities and differences. In the next analytic steps (axial and selective coding) data will be reassembled through statements about the nature of relationship. These statements of relationship are commonly referred to as "hypotheses".

Conceptualization is an abstraction. In the Grounded Theory this abstraction is reached by breaking down data into discrete ideas, events, acts and then by giving them representing names.

Once concepts begin to accumulate, the analyst should begin the process of grouping them or categorizing them under more abstract explanatory terms: categories. Once categories are identified it becomes easier to think about its properties (general or specific characteristics or attributes of a category) and dimensions (location of a property along a continuum or range) and to break it down into subcategories.

2.4.2. The axial coding

The purpose of axial coding is to reassemble the data that were fractured during open coding. Categories are related to their subcategories to form more precise and complete explanations about phenomena. Practically, axial coding is the act of relating categories along the lines of their properties and dimensions. Categories stand for *phenomena,* which are repeated patterns of happenings, events, or actions/interactions that represent what people do or say, alone or together, in response to the problems and situations in which they find themselves.

Axial and open coding are not sequential acts: one does not stop coding for properties and dimensions while one is developing relationships between concepts.

During the axial coding it could help to have a scheme that can be used to sort and organize relations: a *paradigm* [Strauss and Corbin, 1998, p. 8]. The basic components of the paradigm proposed by Strauss and Corbin are

- o *conditions*: a sets of events or happenings that create the situations, issues, and problems pertaining to a phenomenon and explain why and how persons or groups respond in certain ways;
- o *actions/interactions*: strategic or routine responses made by individuals or groups to problems, happenings, or events that arise under those conditions.
- o *consequences*: outcomes of actions/interactions.

This process of linking concepts can be viewed as a construction of hypotheses about concepts.

Instead of looking for properties, axial coding can also *analyzing data for process* and therefore looking at action/interaction and noting the changes occurring in different context and conditions. Bringing process into the analysis is essential; theory without process is missing a vital part of its story: how the action/interaction evolves. Process can be described as "the difference between a snapshot and a moving picture. Each one pictorial form presents a different perspective and gives insight, bit if one wants to see what happens or how things evolve, and then one must turn to the moving picture [Strauss and Corbin, 1998, p. 163]."

2.4.3. The selective coding

The process of constructing concepts is carried out during this stage by the analyst that reduces data from many cases to concepts and sets of relational statements that can be used to explain, in a general sense, what is going on.

When the major categories are finally integrated to form a larger theoretical scheme the research findings take the form of a theory. Selective coding is the process of integrating and refining categories.

The first step in integration is choosing a *central category* that represents the main theme of the research. The are several techniques that can be used to identify the central categories and to integrate concepts:

- o writing the storyline,
- o making use of diagrams,
- o reviewing and sorting of memos either by hands of by computer programs.

After having defined the theoretical scheme, the researcher has to refine the theory. A central category, like any category, must be defined in terms of its properties and dimensions. The central category has to be dense: all the salient properties and dimensions have to be identified.

Then, the theoretical scheme has to be validated: how well it fits with the row data and covers every salient features. One way to validate the scheme is to go back and compare the scheme against the raw data, doing a type of high-level analysis.

One common problem with theoretical schemes is that they fail to account for variation. However, this is not challenging: life does not fit into neat little boxes, there are always variations of every process.

2.4.4. Grounded Theory methodologies

In order to help the beginning researcher to sort out all the complex relationship, Strauss and Corbin create a *conditional/consequential matrix* to keep track of the various components of the analysis and to keep in mind the several analytic points. In particular this matrix is aimed at keeping track of the interplay of conditions consequences and the subsequent actions/interactions and to trace their paths of connectivity. For more details see Strauss and Corbin [1998, pp 181-199].

Concerning the sampling methods, the Grounded Theory applies *theoretical sampling,* a technique for "gathering data driven by the concepts that are derived from the evolving theory and based on the concept of *making comparisons*, whose purpose is to go to places, people or events that will maximize opportunities to discover variations among concepts and to densify categories in terms of their properties and dimensions" [Strauss and Corbin, 1998, p. 201]. These concepts should demonstrate to have relevance to the evolving theory because they should be repeatedly present in the data when comparing the different occurrences and they act as conditions that give variation to a major category.

Basically, during the sampling, the investigator looks for indicators (data) representative of relevant concepts and then compares these indicators for their properties and dimensions. In every step of the analysis the data are constantly corroborated and therefore the consequent interpretations are constantly validated or negated. Theoretical sampling differs on the basis of the coding procedures described above.

Sampling in open coding (*Open Sampling*) is open to all persons, places, and situations that will provide the greatest opportunity to discover, name, and categorize phenomena according to their properties and dimensions. The sampling can be carried where the researcher has the opportunity to gather data or where she knows she can find interesting data. The differences can emerge quite fortuitously or otherwise the gathered data can be reorganized according to the theoretically relevant concepts.

Sampling in axial coding (*Relational and Variational Sampling*) is aimed at looking for incidents that demonstrate dimensional range or variation of concepts and the

relationships among concepts. As in the open sampling there are different procedure to gather data.

In sampling in selective coding (*Selective Sampling*) the researcher's goal is to choose the sites, persons, and documents that will maximize opportunities for comparative analysis. This means returning to old sites, documents and persons or going to new ones to gather the necessary data, saturate the categories and complete the study.

The sample takes long until each category is saturated: *a*) no new relevant data seem to emerge, *b*) the dimensions and the property of every categories are well developed, *c*) the relations among categories are well established and validated. Therefore, the sampling in term of quantity of subjects or evidences cannot be planned in advance and continues until all the categories are saturated.

2.4.5. Grounded Theory and user modeling systems: an example

As a methodology and set of methods, the Grounded Theory can be applied not only in social sciences, but also in practitioner fields such as education, business, communication, anthropology, architecture, psychology and, of course, computer science.

In the field of user modeling (UM) and user-adapted systems, recently an evaluation study applying the Grounded Theory methods has been published in the UMUAI journal [Barker et al, 2002]. The work was concerned with the exploitation of a cooperative student model (collaboration between student and tutor in the construction of the model) of learner characteristics for a multimedia application. The multimedia learning application presented information in a different way on the basis of the individual characteristics of learners (language level, cognitive style, task and question level, and help level).

The evaluation of a complex educational computer application is a difficult process because it becomes hard to isolate the effects of any single variable. A goal of this research was to understand the many and complex interactions between learners, tutors and learning environment. Grounded Theory was used in order to understand this process because this method is able to integrate the range of qualitative and quantitative results.

The range of evaluation methods used in the study were:
- o   video of learners using the application,
- o   interview with learners,
- o   pre-test and post-test results comparison,
- o   data logging of user navigation and online tests and tasks,
- o   questionnaires related to user attitude,

- o tasks and questions results,
- o focus group studies carried out with two small groups of users,
- o staff evaluation of the course,
- o staff diaries,
- o interviews with staff involved,
- o a formal staff report of the experience,
- o expert evaluation of the multimedia application.

Before running the evaluation, a preliminary study was carried out in order to familiarize the researcher with the domain area and to produce a structure for the many categories, sub-categories, and variables involved in the study of the phenomenon.

The measures of learner performance (e.g., significant differences between pre-test and post-test scores) were effective. However, the Grounded Theory is aimed not only at understanding the effectiveness of the user modeling approach quantitatively, but also at evaluating how the application was used both by learners and tutors and their attitude to the user modeling approach.

An important stage in the Grounded Theory method is the investigation of main categories, subcategories and variables involved in the phenomenon under study.

The three main categories identified in the study and their corresponding subcategories were:

- o <u>the student model</u> → performance results, components of the student model, the cooperative student model
- o <u>the learning materials</u> → subject content, design features, usability, learning presentation strategy
- o <u>the management of learning</u> → the tutor, the learning environment

While some examples of variables identified in subcategories are

- o performance results → pre-test, post-test, on/off computer tests, on/off computer tasks;
- o usability → ease of use;
- o the tutor → involvement;
- o …

The process of discovering and organizing categories and subcategories takes place during the open and the axial coding. In these stages, also the research questions were elucidated: i.e. how the quality of learning was influenced by the use of the individually configurable multimedia application.

The selective coding process discovers as core category "*the quality of learning*", around which the research phenomenon can be understood: all the categories involved were associated to the quality of learning and causal relationships were discovered.

2.5. *Paul Dourish and the embodied interaction*

Another example of qualitative methods of research applied to computer science comes from the embodied interaction approach of Paul Dourish [Dourish, 2001a]. Dourish advocates not only the exploitation of ethnomethodological methods to the development of interactive software systems and user interfaces, but also he promotes a sociological stance to HCI. In opposition to the dominant cognitive approach to HCI based to a goal-driven approach to the human-machine interaction, Dourish resumes both sociology and phenomenology based approach to HCI (see [Suchman, 1987], [Winograd and Flores, 1986), more oriented to the embedded interaction between human and machine. Dourish sets out his theory in the book "Where the action is: the foundation of the embodied interaction". In my personal opinion, his theory is interesting not only for the proposed approach, but also for the open application of the embodied interaction to other fields of research, not yet analyzed by the author. A proposal of embodied interaction applied to the developed of user-adapted system will be sketched in 2.6.

The book starts with the observation that two trends in the development of human computer interaction suggest that we need new ways of interacting with computers. These two trends concern the massive increase in computational power and the expanding context in which we put that power to use (computation is now part of cellular telephones, microwave ovens, cars, TV). Dourish calls this new approach to interacting with computers *embodied interaction,* "an interaction with computer systems that occupy our world, a world of physical and social reality, and that exploit this fact in how they interact with us" [Dourish, 2001a, p. 3]. While the traditional computational model of HCI has been rationally built on a procedural foundation and set out its account of the world in term of plans, procedures, tasks and goals, Dourish's model of HCI places interaction at center of the picture. By this he considers interaction not only as *what* is being done, but also *how* it is being done.

2.5.1. *The tangible computing*

During the exploration of his historical model of interaction, Dourish notices a gradual incorporation of a wider range of human skills and capabilities in the interaction with computers. This allows computation to be made ever more widely

accessible to people without requiring extensive training by reducing the complexity of those interactions. The four stages in the historical development of user interfaces identified by Dourish are:

i) *the electrical interaction,* where the dominant paradigm of interaction was electronic. At that time every machine was a prototype and every programs designed for a specific computer.

ii) *the symbolic interaction,* where programming computers required less understanding of detailed construction of every specific machine. The primary form of programs moved from a numeric form (machine language) to other symbolic forms that were more readily understandable to human beings. A further progression along the symbolic path came with the development of early programming languages such as LISP and FORTRAN. Punched cards can be regarded as a primitive form of symbolic interaction, especially because they incorporate both data and instructions;

iii) *the textual interaction,* where interaction became an endless back-and-forth of written instructions between user and system. The textual interaction is characterized by a "grammar" of interaction with commands, parameters, arguments and options;

iv) *the graphical interaction,* where not only words are replaced by icons, but also new dimensions of interaction are opened up (i.e., the interaction happens in a two-dimensional space, visual metaphors are developed, spatial capabilities are exploited, etc)

The graphical interaction is still now (from the 80ies) the dominant paradigm of interaction with computers.

If we look back to the past thirty years, we can easily notice that the interaction with the personal computer itself has changed remarkably little: we interact with it in just the same way. The physical (keyboards, screens and mice) and the virtual (dialog boxes, scroll bars and menus) devices are the same, but also the ways in which the computer fits into our environment and our lives are pretty the same. This kind of interaction tends to demand our direct attention. So, is not only the form of the computer but also the computer-based activity that is changed remarkably little over the last twenty years.

Nevertheless, new forms of computation moving beyond the traditional boundaries of the desk towards an incorporation of our daily experience of physical and social world are arising more and more frequently. Dourish calls these new kinds of interaction *tangible computing,* an umbrella term that encompasses a number of different activities:

o the distribution of computation across a variety of devices, which are spread throughout the physical environment and are sensitive to their location and their proximity to other devices (i.e., printer and fax machines might be aware of the presence of handheld computers);

o the augmentation of the everyday world with computational power, so that pieces of paper, cups, pens, etc can be made active entities responding to their environment and to people's activities;

o how the above approaches can be harnessed to create environments for computational activity in which we interact directly through physical artifacts rather than traditional graphical interfaces and interface devices such as mice.

Therefore, tangible computing is exploring how to get computers "out of the way" and provides people with a much more direct - tangible - interaction experience. In particular, Dourish focuses his attention on the relationship between computers on the desktop and the world in which they (and we) operate.

So, as different researchers already proposed under the term of "ubiquitous computing" [Weiser, 1991], why dealing with a single, large, expensive, computer when you could harness many tiny, low-cost devices spread throughout the environment? Why not putting computation wherever it might be needed? And there are already devices (computers processors inside television set, car, microwave oven) with these purposes and working properly under an HCI point of view because they are organized around human needs and functions. And more and more projects are carried out in different research labs in all over the world (e.g., Dourish mentions the *active badge*, a little electronic device that can be located everywhere by a sensor network; the *digital desk*, a computationally enhanced desktop supporting interaction with both paper and electronic documents; the *reactive room,* a meeting room supporting not only face-to-face meetings, but also meetings distributed in space and time; for more details see [Dourish, 2001a, pp. 30-40].

As it can be expected, the interaction in tangible computing is slightly different. First, there is no single point of control of interaction and there is not even a single device that is the object of the interaction. Second, the tangible computing transforms the sequential nature of the interaction at the interface in a parallel interaction. Third, in tangible design, the physical properties of the interface are used to suggest its use. Nevertheless, interacting with tangible computing opens up a new set of challenges and a new set of (still unknown) design problems.

2.5.2. The social computing

In a parallel way to the tangible computing, the last decade has also seen increasing attempts to incorporate understandings of the social world into interactive systems. Dourish refers to this as *social computing.* Again this term encompasses a number of different activities:

- o the incorporation of social understandings into the design of the interaction itself. That is, it attempts to understand how the "dialogue" between users and computers can be seen as similar and dissimilar to the way in which we interact with each other;
- o the application of anthropological and sociological approaches to uncover the mechanisms trough which people organize their activities around computer systems, and the role that social and organizational settings play in this process;
- o how the "single-user" - one person sitting in front of one computer - interaction can be enhanced by incorporating information about others and the activities of others.

It might seem strange to look at interactive system design from a sociological perspective. Sociology is concerned with the structure and the function of society, while interactive systems are tools that people use. However, although that position seems immediately appealing, the significance of a sociological approach becomes clear when we look at the context in which computation is put to work. The context, for instance, is more social than technical. Furthermore, for Dourish, HCI and sociology share three common characteristics. First, they are concerned with the details of the organization of social conduct rather than broad social trends. Second, they are primarily oriented toward real activities and experiences rather than abstractions or models. Third, they all might adopt an anthropological perspective on collecting, interpreting and using field materials.

2.5.2.1. <u>Suchman's plans and situated actions</u>

In 1987, Lucy Suchman published "Plans and Situated Actions" [Suchman, 1987], which applied the ethnomethodology's techniques and perspectives to the organization of interaction between humans and technology. In doing so, she opened up significant new areas of investigation both for HCI researchers and ethnomethodologists.

Her initial concern was the problem of mutual intelligibility: the relation between observable behavior and the non-direct processes that make behavior meaningful. For psychological studies, the crucial processes are essentially cognitive, located inside the head of the actor. While for social studies, the crucial processes are circumstantial,

located in the relationship among actors, and between actors and their embedding situations.

What Suchman defined in this book was a critique to the notion of a "plan", one of the dominant paradigm for modeling human behavior in Artificial Intelligence, also focusing on the problem of human-machine communication. The "planning" paradigm models the human activity in term of formulation and execution of plans. Plans are script for sequences of actions and they are formulated through a set of procedures beginning with a goal, which is then decomposed into a sequence of sub goals. What Suchman did was to show how this model failed to take into consideration a range of ways in which social science had radically revised our notion of how people act in the world.

The planning model sees features of the world (and of our interaction with it) as stable, objective phenomena. In contrast Suchman presented a model of interaction with the world in which the apparently objective phenomena of the cognitive model were, instead, active interpretations of the world formed in response to specific settings and circumstances. In this model the sequential organization of the behavior is an ongoing and improvised activity. The actions are organized in response to the features of the settings in which they arise. The action is *situated.*

Then Suchman observed that the planning model that so dominates cognitive science was also the basis for the design of interactive devices. Therefore, the arising of a mismatch between the abstract and stable model of a system and the messy and immediate circumstances in which the users interact. She argued that artifacts built on planning model confuse *plans* with *situated actions,* and she recommends instead a view of plans as formulations of antecedent conditions and consequences of action that account for action in a plausible way.

2.5.2.2. **Dourish's and Button's technomethodology**

Suchman's model was not developed to solve problems of a specific design, but rather analyze how the plan model was used to support a whole range of technologies.

To create a deeper connection between sociological understandings and the design of interactive technologies Paul Dourish and Graham Button coined the term *technomethodology* [Dourish and Button, 1998]. Particularly, technomethodology is used to describe a deeper relationship between technological design and ethnomethodology that is able to satisfy to criteria:

  o first, to draw on ethnomethodology's fundamental insight about the organization of action as being a moment-to-moment, naturally occurring improvisational response to practical problems;

o   second, to relate these understandings to basic, fundamental principle around which software systems are developed.

To develop such a long-term objective, they explored particular areas of both technical and ethnomethodological interests trying to find overlaps and mutual orientation to common issues. Specifically, they explored the relationship between ethnomethodology's conception of "*accountability*" and the role that "*abstraction*" plays in the analysis and in the development of software systems. These two ideas are conceptually complementary, but some interesting problems for interaction lie in the differences between them.

Accountability. The notion of accountability is a fundamental feature of the ethnomethodogical perspective. "Acting rationally" and "perceiving actions to be rational" are reciprocal aspects of the same set of understandings. About Garfinkel's notion, as a feature of action, accountability means "observable and reportable" [Garfinkel, 1967, 1-2]. There is more than this, though. Accountability lies in the *reciprocality* of action and understanding. An action can be found to be rational by those who understanding it, "but rather that the methods of understanding and making sense of action and the methods for engaging in it *are the same methods.* In other words, being a competent member of some setting means being able to engage in action in ways that are recognizable to other members" [Dourish, 2001a, p. 79]. The organization of action serves to demonstrate what action is: actions are organized so as to reveal the kinds of actions they are.

The second aspect of accountability for Garfinkel, is *how* accountability arises as a feature of social action. The accountable aspect of activity is never a *commentary* on the activity, rather it is an intrinsic and inseparable feature of how the activity is woven into the fabric of action and interaction. At the same time, Garfinkel emphasizes that the accountability of action is not an absolute matter, it is an "endless, ongoing, contingent accomplishment" [Garfinkel in Dourish, 1998, p. 80]. The analytic concept of accountability emphasizes that the organization of an action provides others with the means to understand what it is and how to respond in a mutually constructed sequence of action. However, this idea does not mesh very well with the way in which we currently design interactive software system. The problem lies in the way in which software relies on a notion of abstraction.

Abstraction. Software systems are built from abstractions. User interfaces offer us abstractions in the form of generic user interface components or "widgets" such as menus, buttons, labels and scroll bars. The notion of "object" in software systems and programming languages is itself an abstraction. The instruction set is an abstraction

that hides a variety of possible implementations. Software systems, in other words, constitute a tower of mutually constituted abstractions right down to binary logic.

There are extremely good reasons why abstraction is such a good principle. First, abstraction makes it possible for us to treat a complex set of computational behaviors as simple high-level objects. Second, it also allows us to use a single abstract object (such as a scroll bar) to capture a range of potential needs and uses. Third, it helps to isolate one component from another so that they can be managed and maintained properly. Without these properties it would be impossible to build a modern software system.

The essence of abstraction in software is that it hides implementation. The implementation is in some ways the opposite of the abstraction: where the abstraction is the gloss that details how something can be used and what it will do, the implementation is the part under the covers that describes how it will work.

However, the idea of information hiding has become critical throughout the design of software systems. At the user interface the situation is more problematic. Within a system the different components will interact in fixed and predictable ways. Users are less predictable, though, and their actions less fixed depending on their goals, reasons, etc.

For ethnomethodology theory, the accountability is the key feature that enables people to interact: *the way that activities are organized makes their nature available to others:* accounts are representations that systems offer of their own activity. But this feature is exactly what is hidden by software abstractions: in the "information hiding" approach the information that is hidden is information about how the system is doing what is does, how the perceived action is organized. For instance, in the networked file servers the file servers are arranged so that they appear as part of the local file system. However, their performance is different and if the user is not aware of the actual organization of the file system some failure might be not understandable.

Dourish and Button [Dourish and Button, 1998] try to address this problem by introducing a form of accountability for the interface. Accountability in this sense means that the interface is designed so as to present, as part of its action, an "account" of what is happening. So, the account should not simply be an abstract description of the system's behavior but rather an explication of how the system current configuration is a response to the sequence of actions that has led up to this moment. As in the Garfinkel's analysis, the relationship between an account and the behavior it accounts for, is the key feature of the accountability. So, the account has to be strongly connected to the behavior that it describes. The account has to emerge along with the action. The offered account has to be an account of the current specific behavior of the system.

Dourish and Button [Dourish and Button, 1998] propose as solution a software design technique called "computational reflection". This technique emerged originally in the domain of AI programming. The basic idea of reflection is that there are two domains in the execution of any program: the domain about which the program is dealing (for instance, word and format for a word processor) and the domain of program itself (comprising its internal structures, program encoding, execution state, etc). Normally these two worlds of representation are kept separated, but reflection can provide a link between them. The reflective link allows a programming system to perform computations using not only the representations that refer to the outside world but also those internal representations that refer to its own operation. This gives a program the ability to describe its own internal state and even to operate upon itself by revising those internal structures. The link between the two domains is called the "causal connection" between the representation of a program and its own behavior. The causal connection provides the features required of the relationship between and account (representation) and the actual behavior of a system (program).

They propose the reflection as basis for interface accountability for this reason: because we know that people don't just take things at face value but attempt to interrogate them for their meaning, we should provide some facilities so that they can do the same thing with interactive systems. This proposal is radical because it is radical the relationship between technical design and social understanding and technomethodology is the most extreme proposal to bring these two elements together.

In summary, the central theme of social computing is that social action is *embedded*: it is firmly rooted in the settings in which it arises, where that setting is not just material circumstances, but social, cultural and historical ones as well. Moreover, social action is clearly organized and this social order emerges from practice (as well as in HCI practices emerge not from the designers of the system but from the actions of its users). So, the embedded approach to social action turned the attention on how the orderliness of social conduct was achieved and, given the role that technology places in social settings, "*the key question is to understand how the relationship between technology and social action comes to be worked out in different situations, and from these to understand how the features of technological design and the features of everyday social settings are related* [Dourish, 2001a, p. 97]".

2.5.3. From tangible and social computing to the embodied interaction

The starting hypothesis of the Dourish's book is that tangible and social computing are aspects of one and the same research program. He sets four arguments out.

First, tangible and social computing are based on the same underlying principles. In particular they both exploit our familiarity with the everyday world, a world of social interaction or physical artifacts. The role of everyday world draws on the *ways we experience the everyday world,* which is basically by interacting with it. Thus, the individual cannot be separated by the world in which that individual lives and acts.

This perspective comes about in contrast to the Cartesian "naïve cognitivism" which dominated the thinking of computer system designers and still persist to a considerable degree. This approach makes a strong separation between the mind, as the seat of the consciousness and rational decision and the objective external world, as a largely stable collection of objects and events to be observed and manipulated. From this perspective, a disembodied brain could think about the world just as we do. In contrast, the new perspective on which tangible and social computing rest argues that a disembodied brain could not experience the world in the same way we do, because our experience of the world is intimately tied to the ways in which we act in it. Physically our experiences cannot be separated from the reality of our bodily presence in the world; and socially, too, the same relationship holds because our nature as social beings is based on the ways in which we act and interact.

Second, the core element that tangible and social computing have in common: the idea of *embodiment.* Dourish gives two definition of embodiment " [Dourish, 2001a, pp. 100-101]:

1) *Embodiment 1. Embodiment means possessing and acting through a physical manifestation in a world.*

2) *Embodiment 2. Embodied phenomena are those that by their very nature occur in real time and real space.*

Following the Dourish's definitions, embodiment does not mean simply physical reality, but also denotes a form of participative status. Embodiment is about the fact that things are embedded in the world, and the way in which their reality depends on being embedded. For the proponents of tangible and social computing, the key of their effectiveness is the fact that we, and our actions, are embodied elements of the everyday world and we are familiar with the "real-world-ness".

Why is embodiment relevant to this sort of interaction with computers?

First, the designers of interactive systems have increasingly come to understand that interaction is intimately connected with the settings in which occurs. The adoption of anthropological techniques has underlined the role of physical and social environments. Second, this focus on settings reflects a more general turn to consider work activities and artifacts in concrete terms rather than abstract ones. Instead of

developing abstract account of mythical users, HCI increasingly employs fields studies and observational techniques to stage "encounters" with real users, in real settings, doing real work. These encounters are often very revealing, as they often show that the ways the work gets done are not the ways that are listed in procedural manuals, or even in the accounts that the people themselves would tell you if you asked. Attention to details and to actual cases leads to a concern with how interaction is manifest in the interface. Tangible computing reflects this concern by exploring new opportunity to manifest computation and interaction in radically new forms, while social computing seeks ways for interaction to manifest more than simply the programmer's abstract model of the task, but also specifics of how the work comes to be done. Third, there is recognition that, through their embodiment in the world we occupy, the artifacts of daily interaction can play different roles.

Third, the idea of embodiment is not a new phenomenon. The notion of embodiment plays a special role in one particular school of philosophical thought, phenomenology, originated in the latter part of nineteenth century and last for the following hundred years through a number of distinct intellectual positions (Husserls, Heidegger, Schutz, Merleau-Ponty, Wittgenstein).

Phenomenology is primarily concerned with the elements of human experience: how we perceive, experience and act in the world around us. What differentiates it from other approaches is its central emphasis on the actual phenomena of experience, where other approaches might be concerned with abstract world models with a truth independent of our own experience. In contrast, the phenomenologists argue that the separation between mind and matter (the Descart's *res cognitans* and *res extensa*) has no basis in reality. Thinking does not occur separately from being and acting: the way I encounter the world gives it meaning for me the way I act in the world reflects different meanings. Consequently, phenomenology has attempted to reconstruct the relationship between experience and action without this separation. Perception begins with what is experienced, rather than beginning with what is expected. The model is "*sum ergo cogito*" rather than "*cogito ergo sum*" [Heidegger, 1927].

However, phenomenology is not only about perception, but it is also concerned with action, with understanding, and with how these are all related to each other, as part and parcel of out daily experience as participants in the world. For instance, Ludwig Wittgenstein [Wittgenstein, 1953] develops related approaches to topics such as language and meaning. In his famous sentence "the meaning of a world is its use in the language" he wants signify that meaning is embedded in the practice of the language. For Martin Heidegger [Heidegger, 1927], instead, the nature of being – how we are in the world – shapes the way that we understand the world, because our

understanding of the world is essentially an understanding of how we are in it. And the most important aspect of the way in which we encounter the world is that we encounter it practically. *It is the way in which we act that makes the world meaningful for us.*

Fourth, on the phenomenological understandings can be created a foundational approach to the embodied interaction (a phenomenological approach to HCI was already proposed by Winograd and Flores [1986]).

The Dourish's embodied interaction, indeed, is not simply that is a form of interaction that is embodied, but rather that it is an approach to the design and analysis of interaction that takes embodiment to be central to, even constitutive of, the whole phenomenon. Tangible and social computing both reflect this central concern with embodiment. Tangible computing attempts to capitalize on our physical skills and our familiarity with real world objects. It attempts to move computation and interaction out of the world of abstract cognitive processes and into the same phenomenal world as our other sort of interaction. The use of sociological approach is motivated by the "situated" perspective [Suchman, 1987], which is grounded in the relationship between social action and the settings in which it unfolds, the relationship of embodiment. And the origin of this concept was developed in the phenomenological tradition. For the phenomenologists:

- o embodiment is not simply a "physical manifestation", it means being grounded in everyday, mundane experience. The source of action and the meaning are in the world: embodiment is a participative status, a way of being;
- o the action in the world are fundamental to our understanding of the world and our relationship with it;
- o the embodied practical action is the source of meaning: we find the world meaningful primarily with respect to the ways in which we act within it.

In summary, social computing similarly recognize that the meaning is something that users create through that ways in which they interact with technology and with each other, and it opens up the opportunity to explore and negotiate meaning in the course of interacting with and through software systems. So, the major lesson drawn from the phenomenological work is that embodiment is about the relationship between action and meaning.

Therefore, Dourish concludes that *embodiment is the property of our engagement with the world that allow us to make it meaningful* and *the **Embodied Interaction** is the creation, manipulation and sharing of meaning through engaged interaction with artifacts* [Dourish, 2001a, p. 126].

The idea of embodiment can be used in two ways:

1) as basis for an approach to design that is oriented toward the way in which people interact with systems as fundamentally embodied phenomenon. The underlined idea is that the activity and the interaction with the real phenomena of experience are central, rather than focused on internal or purely cognitive interpretations;

2) as a way of uncovering issues in the design and use of existing technologies: a stance we can take on the design of interactive systems.

The primary characteristic of technologies supporting embodied interaction is that they variously make manifest how they are coupled to the world. The embodied interaction perspective begins to illuminate not just we act *on* technology, but how we act *through* it.

2.5.4. Dourish's design principles

The original motivation for exploring embodied interaction was to help design new systems. The difficulty of articulating the relationship between theory and design has persistently dogged interdisciplinary work in HCI. This is not least because theory and design are fundamentally different sort of activities, carried out by different people with different training and presented to different audience.

The design implications of field studies should arise through an explicit dialogue between researchers from different disciplines. Both theory and design gain value from being put together.

The core argument of the book is that social and tangible computing share a common foundation in embodied interaction. At the heart of tangible computing is the relationship between activities and the space in which they are carried out. Tangible computing explores this in three related ways: through the configurability of space, through the relationship of body to task, and through physical constraints. Social computing instead, argues that interaction with software systems needs to be seen in a broader context: the context in which it draws is the socially constructed setting within which the interaction takes place. Social computing introduces a model where the sequential organization of interaction does not simply result from the "execution" of a formal plan in the user's head, but instead arises from a process of continual responses to circumstances within which it was being produced.

This has two implications for the design:

1. supporting the improvised sequential organization action by giving users more direct control over how activity is managed (for example by organizing the

interaction as informal assemblage of steps rather than a rote procedure driven by the system);

2. helping the process of improvised, situated action by making the immediate circumstances of the work more visible.

The broader idea of embodied interaction points out that action and meaning arise in specific settings, physical, social, organizational and so forth.

Dourish pointed out six main principle aimed at taking elements from theoretical understandings and at showing how they are particularly important for design [Dourish, 2001a, pp. 155-188]:.

2.5.4.1. <u>Computation is a medium</u>

Certainly computers provide a medium for communication: they represent and convey information. However, they do not make computation the central element of communicative act. This is and idea already suggested by the proponents of computers in education. Nonetheless the idea of Dourish is different. First of all, he notices the existence of communication between the designer of a system and a user trough the medium of the system itself: the structure of the system communicates to the user some set of expectations that the designer held for its use and this communication is achieved by *modulation*. Media are modulated when they are transformed in some way to carry information. In the case of embodied interaction, the modulation must encompass not only the technology, but also the practice in which the technology is embedded. So, the meaning is transmitted not only through a system but also through the practices that surround it (for instance, people develop expectations about the information that can be found on the World Wide Web).

The most obvious way to observe how a system modulates its effects on a user's actions is to see how our own activities are transformed when we interact with the system. The computer systems augment and amplificate our own activities and they are embedded into a set of practices.

2.5.4.2. <u>Meaning arises on multiple levels</u>

Objects carry meaning on multiple levels: as entities in their own right, as signifiers of social meaning, as elements in systems of practice, and so on. Systems or artifacts supporting embodied interaction need to be designed with an orientation toward the multiplicity of meanings that may be conveyed through them. The different levels of meaning involve artifacts and representations in different ways (for instance the dimensions iconic and symbolic). Design needs to consider how those different levels

of representation will be manipulated and controlled by the users (e.g., are the users acting "on" or "through" the artifact?)

2.5.4.3. <u>Users, not designers, create and communicate meaning</u>

and

2.5.4.4. <u>Users, not designers, manage coupling</u>

Traditional interactive system design ascribes two sets of responsibilities to the designer: the responsibility for the form and the function of the artifact and the responsibility for its use. For the first, the designers have the primary responsibility, even if new approaches such as User-Centered Design and Participatory Design have underlined the importance of the active role of end-users in the design of software systems. The second responsibility ascribed to the designers must be designed with some expectation of its final use. However, designers are continually surprised at the uses to which their artifact are put, or the ways in which they are incorporated into the activities of users. So, how technology will feature as an aspect of working practice cannot be predetermined by the designer, but instead will emerge from the specific, situated activity in which the technology is incorporated.

Embodied technologies are used to create and communicate meaning and because they can only have meaning through the way in which users incorporate them into working practices, then clearly the manipulation of meaning and coupling (intended here as an intentional connection that arises in the course of interaction) are primarily the responsibility of users, not of designers. These observations can have an impact on *designer's stance.*

The designer's stance is, for Dourish, the designer's conception of her role in the interaction between the user and the artifact. In the traditional approach the designer manage the interaction between user and artifact thorough control of the design parameters for the artifact. This stance is reflected in the tools available to interactive system designers: task-analytic methods to model activities in which the user is engaged, *user modeling* methods to understand the user point of view in the course of interaction, cognitive-evaluative techniques to assess the cognitive impact of different designs, etc.

This stance has to be transformed when we recognize that users play a much more active role during the interaction. Therefore, the designer should focus on ways for the user to understand the tool and how to use the tool in each situation instead of designing the ways to use the artifact. The first resource concerns the ability to operate on entities at different levels both acting *with* them and acting *through* them. While, in contrast, conversational approach separates the use of an artifact from its manipulation and configuration.

2.5.4.5. **Embodied technologies participate in the world they represent**

As for Heidegger [Heidegger, 1927] the meaning arises from engaged action in the world, the Dourish's embodied perspective rejects the traditional separation between representation and object: they are entities that participate to a single coextensive reality. Similarly, the technology of embodied interaction participates in the world they represent.

Technically mediated communication involves the encoding of a communicative act into some representation (text, audio) and *this representation* is interpreted by the remote participant in addiction to the content of the communication.

So, the representation works on multiple levels, and so interactive systems need to allow people to operate on them at multiple levels: in different contexts, the same entity may be an object of action or a means by which some action is achieved.

2.5.4.6. **Embodied interaction turn action into meaning**

The relationship between action and meaning is central to the idea of embodiment. The core idea of and embodied interface is the ability to turn action into meaning. *Meaning does not reside in the system itself, but in the ways in which it is used.*

Features of the design afford particular ways of understanding it. For instance, within a community of practice [Wenger, 1998] (groups of people sharing histories, identity and meaning through their common orientation toward and participation in practical services; the communities of practice are the social grouping within with the meaning is formed, negotiated developed, and communicated) the technology does not simply afford certain sort of actions, but it also reflects particular sets of assumptions, convections and practices.

2.5.5. *Conclusion and directions*

Embodiment is a feature of interaction, not of technology. It is rooted in the ways in which people and technologies participate in the world. Embodiment is about engaged interaction rather then disembodied cognition; it is about the particular rather then the abstract, practice rather than theory, directness rather than disconnection.

Embodied interaction is not a technology or a set of rules. It is a perspective on the relationship between people and systems. The question on how it should be developed, explored, and instantiated remains an open research problem.

2.6. *Discussions*

At the end of this excursus I want to analyze how the described methodologies can be applied to the evaluation and the development of user modeling and user-adapted systems.

The exploitation of qualitative methods to the evaluation of user modeling and user-adapted systems is not a new idea, even if quantitative methods are largely applied. As discussed in 2.4, Grounded Theory, for instance, has been applied to the evaluation of a user modeling system and helped to discover new concepts to take into account in the user model. Other qualitative methods of research have been successfully applied. For instance, see [Oppermann, 1994].

Evaluating user in a qualitative way requires a fewer number of users involved then in case of quantitative research. However, this has always been one of the critics moved to qualitative researchers. Their defense has always been that qualitative methods allow to reach a deeper knowledge of the subjects involved that compensate the less representative sample involved (see discussion in the final Conclusions of the thesis).

In fact, what this kind of methodologies can offer is a more accurate knowledge of the real behavior of a user sitting in front of an interactive system compared to the artificial situation of lab environment. This information can be useful *i)* during the development of an adaptive system by singling out, for instance, the dimension for modeling the users and *ii)* for a system revision after the evaluation. The problem concerns the difficulty of modeling an interaction by taking into account a situated action instead of a plan, for example, predetermined by "search space" of goals and actions.

The two perspectives seem to be opposite, but I want to find some point of contact. In particular my question is how an embodied approach can be applied to user modeling and user-adapted systems.

Under a phenomenological point of view, we reach the meaning by acting in the world, so in case of user-machine interaction a subject reaches the knowledge about the system only by using the system and interacting with and through it. And also for "the system", the subject becomes meaningful only during the interaction.

Another key point of phenomenological perspective is that experience and interaction come before meaning, while the Cartesian view considers action arising from meaning as the expression of internal mental states. So, the way things are organized shape our understandings of those things.

A logical conclusion of both these observations could be constructing the knowledge base of an adaptive system only during the interaction, by learning from the real user behavior since we cannot predict the user behavior until she experienced the system. Indeed, this is the point of view of machine learning techniques and collaborative filtering system that adapt the interaction without prior knowledge of the user.

Nevertheless, there are some ways to model the user in advance also under a embodied perspective. The user model could be originated, for instance, by the observation of real users interacting with similar systems or with the system to model, if it is already implemented. Therefore, interpretative techniques such as contextual inquiry, cooperative and participative evaluation and ethnography (see 2.1.2.2) can be applied to monitor and to have feedback from the user in every design step.

To extract relevant dimensions we could, for example, analyze *work practices* and look for common patterns emerging from different users' actions. On the basis of existing correlations between users and practices, information on how users understand the system can be gained and exploited to model the users.

To offer personalized recommendations, instead, we could build the knowledge base by monitoring the user choices. Then, we could propose the system's recommendations to the users and asking them to evaluate such proposals and discussing with them about their choices. Finally, revising the knowledge on the basis of user feedback.

Following Dourish's advices, the designer should focus on ways the user understands the tool and how she uses it in each situation instead of merely designing the ways to use the artifact (this is similar to the "tool paradigm" approach for interactive systems – the idea that a system should present itself to users as a tool without constraining how the tool is to be used [Dourish, 2001b]). Therefore, Dourish proposed techniques [Dourish, 2001b] such as visualizing the behavior of software systems, visualizing security, populating the social workspace, ad-hoc and emergent information structures. All these methodologies are aimed at giving cues, not about the low-level software implementations, but about the experience of computation. The question is allowing users to see the consequences of their actions, and understand how those actions can be transformed to yield different results [Dourish, 2001a].

Following this perspective we could link user modeling and adaptive systems to technomethodology (see 2.5.2.2), and in particular to the idea of providing facilities to interrogate interactive systems to discover their meaning. The consequence in user modeling and adaptive systems is making the user aware about how adaptation works (this is not a new idea, see [Oppermann, 1994; Höök, 2000) and adapt the ways in which these facilities are presented on the basis of user model.

However, different users probably understand the tool in different ways. So users having a background in computer science have a different approach in understanding how an adaptive system works, compared, for instance, to users having a background in Communication Science. So, also in this case I advocate the need of different suggestion tailored on the basis on user profile.

In conclusion, the most important lessons learned at the end of this excursus are

- o the importance of user observation in her real context (social, cultural, organizational);
- o gathering field data and studying working settings;the importance of usage studies that point out the unexpected uses of technology that the designers had never intended;the attention to user practices and practices shared in the communities;
- o user involvement and user participation in the system design;
- o the link between meaning and experience.

3. Evaluation of user-adapted systems in practice

After having described methods and techniques to evaluate interactive software systems, I want to propose how these methods can be exploited to evaluate user-adaptive systems according to the three tasks in which Kobsa, Koenemann and Pohl divide *personalized hypermedia applications[44]* (Kobsa et al., 2001). Thus, Section 1 sketches the three tasks, Section 2 proposes the classification of evaluation methodology according to the tasks, and Sections 3, 4, 5 describes evaluation of three different adaptive systems.

3.1. The three tasks characterizing personalized hypermedia applications.

In their review of personalized hypermedia presentation techniques Kobsa, Koenemann and Pohl divide the personalization process into these three major tasks.
- o   acquisition method and primary inferences,
- o   representation and secondary inferences,
- o   adaptation production.

3.1.1. Acquisition Method and Primary Inferences.

This task is aimed to identify and to gather the information necessary to construct an initial user model. This process can be further divided in three steps:

1. identifying the available information about
   a. <u>*user's characteristics*</u> (demographic data, user knowledge, user skills and capabilities, user interest and preferences, user goals and plans);
   b. <u>computer usage behavior</u> (selective actions, temporal viewing behavior, ratings, purchases and purchases-related actions, usage frequency, situation-action correlations, action sequences);
   c. <u>usage environment</u> (software and hardware environment, locale information such as user's location or usage local).
2. making the collected information available to the adaptation component of the application;
3. constructing initial user model, usage model, and environment model.

The information listed in Point 1 (user/usage/environment) can be obtained either by monitoring the user's behavior or by external sources. For instance, while some user

---

[44] The authors define as *personalized hypermedia application* a a system which adapts the content, structure, and/or presentation of the networked hypermedia objects (web pages) to each individual user's characteristics, usage behaviour and/or usage environment.

data can be supplied by the user, most usage data can be inferred from observation. The acquisition methods to obtain the data in Point 1 can be divided in:

- **user model acquisition methods**:
    - user-supplied information (explicit questions, controlled queries, tests, exercises, etc),
    - acquisition rules, which could be, for examples, acquisition rules that are typically executed when new information about the user is available, such as observed user actions or interpretation of user behavior;
    - plan recognition, which deals with reasoning about the goals that the user may pursue and the action sequence (plan) she performs to achieve them;
    - stereotypes reasoning, which consist in classifying people into categories and to make predictions about them based on a stereotype that is associated with each category;
- **usage model acquisition methods**, which are methods aimed at modeling the user behavior as direct basis for system personalization. Machine learning algorithms can be applied for these purposes;
- **environment data acquisition methods**, which deal with acquiring software environment information (i.e., information about the web client obtained from the HTTP headers), hardware constraints (which are difficult to asses in a implicit way) and locale information about the physical environment (i.e., locality information actively provided by mobile devices)

3.1.2. Representation and Secondary Inferences.

This task is aimed to represent the acquired user/usage/environment information appropriately in a formal system, let them available for further processing and to draw further assumptions. Several types of representation approaches and inference techniques can be applied:

- ○ **deductive reasoning**
    - logic based representation and inferences: methods based on different logic based reasoning such as concept formalism, proportional calculus, modal logic, etc;
    - representation and reasoning with uncertainty: in order to cope with the uncertainty present in user modeling, methods that rate the validity of user model contents by evidence rules can be applied (Bayesian network, fuzzy logic, etc)

- o **inductive reasoning**
    - ▪ <u>learning</u>: inductive reasoning about the user involves monitoring users' interaction with the application and drawing general conclusion based on a series of observations. Learning algorithms can be employed to acquire user profile, such as features-based techniques, neural networks, and explicit user ratings.
- o **analogical reasoning**
    - ▪ <u>clique-based filtering (collaborative filtering)</u>: methods that adapt the system to the individual user on the basis of the behavior of her "interest neighbors", other users that show similar interaction behavior;
    - ▪ <u>clustering user profiles</u>: methods that form explicit user profiles using machine learning methods and statistics. Basically, clustering algorithms are applied to the available profiles in order to find similar users and to form group profiles.

3.1.3. Adaptation Production.

This task is oriented to the generation of contents, presentation and structure adapted on the basis of a given user, usage, environment model. For details about this task, see also [Brusilowsky, 1996].

- o **adaptation of content** changes the information that is presented in hypermedia pages. Personalized contents can be aimed at offering:
    - ▪ optional explanations,
    - ▪ optional detailed information,
    - ▪ personalized recommendations,
    - ▪ theory-driven presentation,
    - ▪ optional opportunistic hints.

Possible techniques to adapt contents to different used are:
- ▪ <u>page variants</u> → are different version of the pages in which adaptation occurs;
- ▪ <u>fragment variants</u> → each adaptive fragment of a static page change at runtime;
- ▪ <u>fragment coloring</u> → the content remains unchanged, while for each user certain element of the page may be colored in a different way
- ▪ <u>adaptive stretchtext</u> → is elastic text that the user (or the system, in personalized hypermedia system) can extend or collapse;

- adaptive natural language generation → natural language generation techniques are applied to create alternative text description for different users.

o **adaptation of presentation and modality** changes the way in which information is conveyed to the user, while the content stays the same. Adaptation concerning multimedia presentation is often based on explicit user's preferences.

o **adaptation of structure** refers to changes in the way in which the link structure of hypermedia documents or its presentation to users is changed. Adaptation of structure can be realized by applying:

  - collateral structure adaptation → links adaptation present in fragment variants;
  - adaptive link sorting → ranking the link lists on the basis of their relevance for each user;
  - adaptive link annotation → exploitation of different colors and symbol codes to annotate links in a personalized way;
  - adaptive link hiding and unhiding → removing the visible indicator of a link in order to reduce the hyperspace (then the cue may be unhidden, for instance, when the user has all the necessary prerequisites);
  - adaptive link disabling and enabling → removing the functionality of a link without changing its visual appearance;
  - adaptive link removal/Addition→ deleting the link anchors completely.

Possible personalization functions of adaptation of the structure are concerned with the offer of

  - adaptive recommendations: recommendations concerning products or information, navigation recommendations, etc;
  - adaptive orientation and guidance: adapting the navigation style by offering different links to each user (e.g., personalized next buttons in educational systems);
  - personal view and spaces: supporting users in creating personalized views and personalized "information spaces", based mostly on usage data (e.g., adaptive short list of recent URLs; personalizes spaces on portal sites).

3.2. Evaluation methods according to Kobsa's tasks

In Table 3.1 I propose how the evaluation techniques described in the first two Chapters can be used during the three personalization tasks singled out by Kobsa et al.

First of all, I advocate a layered approach, since, as described in 1.4.1, user-adapted systems need a evaluation that differentiates, at least, problems concerning content adaptation and interface adaptation.

Then, I want to underline that evaluation methods for user-adapted systems result well suited not only for the mere evaluation, as in regular HCI applications, but also as knowledge sources for the development of the adaptive application.

So, for instance, task analysis can be used to analyze not only the way people perform their jobs, but also how different kinds of users perform their jobs and then modeling the interaction on the basis of these different models.

In case of heuristic evaluation, what we need to know is if an interface works not for a generic user, but how can work for different users. Then, we can project the interface on the basis of the experts' suggestions.

Evaluations performed when the system is complete, instead, can lead to refinement and updating of the user model. And so on.

This is of course not a new idea. A similar approach has been already proposed by Paramithys, Totter and Stephanidis [Paramithys et al., 2001] as outlined in 1.4.1. The difference is that *i)* I propose the exploitation of evaluation techniques taking into account different layers of adaptation (the Kobsa's tasks) compared to those ones taken into account by Paramithys et al.; *ii)* the techniques here proposed encompass a wider range of evaluation methods.

In addition, as Paramithys et al. already noticed [Paramithys et al., 2001], since the concept of the typical user of a system cannot be applied in adaptive systems, each individual evaluation task has to take into account a particular user having characteristics encoded in some type of user profile and in a particular context of use.

Dix [Dix et al. 1998] listed the following factors distinguishing evaluation techniques:

- o   the stage in the cycle at which the evaluation is carried out,
- o   the style of evaluation,
- o   the level of subjectivity or objectivity of the technique,
- o   the type of measures provided,
- o   the information provided,
- o   the immediacy of the response,
- o   the level of interference implied,
- o   the resources required.

In the following Table I will try to pinpoint some more factors distinguishing the evaluation techniques exploited in the evaluation of user modeling and user-adapted systems.

	Evaluation methods	Acquisition Method And Primary Inferences	Representation And Secondary Inferences	Adaptation Production
Collection Of User's Opinion	*Interviews, Questionnaires, Existing Users Surveys*	o To know which user's dimensions acquiring o To collect user's opinions useful to define primary inferences o To analyze correlations between user's dimensions and opinions, tastes, behaviors, etc o …	o To set the user model dimensions on the basis of collected information o …	o To collect information on how the system works (useful also on the final phase of an experiment) o To collect user's suggestions o To collect real user's preferences o …
	Focus group	Focus group with real users can offer information useful to every personalization tasks. The advantage is that the focus group can be oriented by the evaluator on the basis of the evaluation goals. This method can be applied in participative design evaluation and used to gain user's feedback for every design phase		
	Having user evaluating items		o To refine and update the user model	o To check the correctness of recommendations
User observation	*Think aloud protocols*	Think aloud protocols involving real users can offer information useful to every personalization tasks. It has to be performed when the system (or a prototype) is running. This method can be applied in co-operative design and co-operative evaluation		
	Task analysis	o Useful to goal and plan recognition methods	o To set cognitive dimensions of the user model o …	o To adapt the interface to user's tasks and goals o …
	Cognitive and socio-technical models	o	o To model the user on the basis of the adopted model	o To adapt the interface on the basis of the goals set by the model o …

	Observing users in context	(see ethnographic studies, contextual inquiry, sequential analysis)		
	Logging use	It can be considered as a summative method and can offer clues for every design phases		
Predictive evaluation	*Heuristic evaluation*			o To establish interface adaptations o To check the correctness of interface adaptations
	Domain expert appraisals	To get suggestions useful for the generation of inferences and recommendations and content adaptation		
	Parallel Design			o To develop different interface solutions for different users
	Cognitive walkthroughs			o To observe how users exploit the interface and get clues on how interface adaptations work o …
Formative Evaluation	*Wizard of Oz simulation*	To evaluate the adaptations without implementations		
	Mock-ups			o To test the interface adaptations in the early phase of the design
	Scenario- based design			o To test the interface adaptations on the basis of proposed scenarios
	Prototypes	For a first testing phase. On the basis of the implemented features, one (or all) the tasks can be evaluated		

Summative Methods	*Usability testing – Acceptance test*	o To check the usability of the (adapted) interface and the system's performance
	Controlled experiment	o For a complete (quantitative) evaluation of the system. o To refine and update the user model
	Ethnographic studies	o For a complete (qualitative) evaluation of the system. o For the individuation of the categories of the user model. o To refine and update the user model
	Contextual inquiry	o For a complete (qualitative) evaluation of the system o For the individuation of the categories of the user model
Selection process evaluation methodologies	*Precision and recall, Training set and test set, Evaluation of the ordering, MAE and RMSE, Reversal rate, Sensitivity measures, Utility metrics, Simulation*	o To test recommendations. It can be performed also in the first stage of the design o To refine and update the model
	Grounded theory	o For a complete (qualitative) evaluation of the system. o For the individuation of the categories of the user model. o To refine and update the user model
	Sequential analysis	o For a complete (quantitative) evaluation of the system o To refine and update the user model

Table 3.1. Evaluation methods according to Kobsa's tasks

3.3. *Evaluating an electronic program guide*

In this Section and the following ones I describe four different evaluations I carried out during these last two years. This section describes the evaluation of an Electronic Program Guide.

3.3.1. Introduction

With satellite and cable TV, the convergence of TV and Internet and the advent of digital networks, the offer of TV channels will increase in the near future. Consequently, it will be very difficult for the users to find their favorite programs. They will be exposed to an information overload similar to those known in the World Wide Web. On one hand, in such a scenario personalized filtering techniques will become fundamental to reduce the huge amount of broadcasted programs. On the other hand, the growth of the TV-offer will allow users to differentiate their own choices. In this way the users will be able to impose them in a world that has seldom paid attention to personal preferences. Thus, the knowledge of individual tastes and habits will take an importance just unknown in the past. Therefore, it will be necessary the presence of an intermediary between the TV broadcasters and the viewers, such as a personalized electronic programs guide (EPG), in order to lighten the user from the burden of search. The purpose of an EPG, actually, is to recommend in a timely fashion the programs which best match the individual viewing preferences.

3.3.2. The Personal Program Guide[45]

The Personal Program Guide (Ardissono et al., 2001; Difino et al, 2002) is a user-adaptive Electronic Program Guide that tailors the recommendation of TV programs to the viewer's interests, taking several factors into account, such as her viewing habits during the different times of day. This system captures an individual model for each registered user and employs it to generate an EPG whose contents and layout are tailored to the user watching TV[46]. Moreover, the system automatically records the supposed preferred TV programs or suggests them to the user. The system is aimed at supporting the personalization of the interaction since the first time a user views the EPG and at achieving precise descriptions of the user's preferences in the long term. To this purpose, the system integrates user modeling techniques based on explicit preferences (elicited by questioning the user), stereotypical information about classes of TV viewers, and unobtrusive user modeling techniques aimed at determining the

---

[45] This work has been developed at the Dipartimento di Informatica – Università di Torino in cooperation with TI LABS within the project "Modellizzazione automatica dell'utente nell'interazione su Web".
[46] At the current stage, we have focused on the personalization of the EPG to individual TV viewers. The management of household viewing preferences is part of our future work.

user's preferences on the basis of the observation of her viewing behavior. The system is based on a multi-agent architecture and is designed to run on the user's Set Top Box, where it maintains the user models of individual TV viewers and generates the EPG. The decentralization has clear advantages in the preservation of the viewers' privacy, as the user model is stored locally to the Set-top box. Moreover, the continuous analysis of the user's actions on the TV is possible, therefore supporting the revision of the user model based on a complete picture of her viewing behavior.

The **User Modeling Component (UMC)** is the core element for the personalization task. This agent maintains the whole information about socio-demographic data and preferences of the registered users. The UMC is composed of four modules:

- the **UMC Manager** that manages the user model exploited within the system for generating the personalized EPG. This module infers the user's preferences by relying on a set of user modeling modules (UM Experts, see the following points) that apply different user modeling techniques for the estimation of such preferences. The UMC Manager combines the experts' predictions, which could be conflicting, into a Main User Model that represents the integration of the different points of view on the user's preferences;

- the **Explicit Preferences Expert** that infers the user's preferences on the basis of her general interests and her declared preferences for broad program categories;

- the **Stereotypical UM Expert** that exploits stereotypical information about viewing behavior to predict the user's preferences, given her socio-demographic data and her declared hobbies and interests;

- the **Dynamic UM Expert**, which infers the user's preferences by unobtrusively monitoring her viewing behavior, i.e., the actions performed while she consults the EPG or she watches TV programs.

The UMC Manager integrates the experts' points of view on the user's preferences by merging their predictions in a weighted way, depending on the presumed reliability of the predictions, and caching the resulting estimates into the Main User Model. To describe the reliability of a prediction, the expert's confidence in the prediction itself has been employed. This confidence is a subjective evaluation, performed by the expert without taking user feedback into account, and depends on the estimation of the quality of the data used to generate the prediction.

In the following, I will focus on the component of the system useful to clarify the formative evaluation *i)* of the PPG selection process and *ii)* of the three interface prototypes described in 3.3.4. For more details about the other components of the systems see [Ardissono et al., 2001; Difino et al, 2002].

3.3.2.1.   **The Stereotypical UM Expert**

The exploitation of sociological stereotypes seems to be usual in the mass-media world. Thus, we decided to generate our user modeling knowledge base starting from an analysis of existing surveys about TV viewers. Particularly, we examined the lifestyles surveys that cluster the population into groups according to consumers' preferences, socio-cultural trends, and homogeneous behaviors. Especially, we concentrated on a lifestyles' study, Sinottica, conducted by Eurisko data analyzers [Calvi, 1986]. Given the completeness of the considered viewpoints and the reliability of collected data, we decided to build the stereotypes knowledge base starting from the Eurisko lifestyles. However, the information regarding the lifestyles is not defined in a formalized way. Thus, we exploited a formalism to structure the information characterizing each user class in order to represent in a formalized way the lifestyles descriptions [Torasso and Console, 1989]. Moreover, we structured the stereotypes in two main parts, assuming a plausible correlation among homogeneous user groups and their preferences:

- o   *a profile*, containing the classification data of individuals belonging to the represented stereotype;
- o   *a prediction part*, containing the preferences typical of such individuals.

A similar approach has been adopted in SETA, a prototype toolkit for the construction of adaptive Web stores [Ardissono and Goy, 2000]. While the classification data are used to evaluate how close the individual viewer using the EPG matches a stereotypical description, the preferences are used to enable the user modeling system to make initial predictions by exploiting stereotypical information. The Eurisko lifestyles description has been used for the profile of the stereotypes, which has been further split into two main parts: *personal data* (age, gender, education level, type of job, geographic zone); *interests*.

Regarding the prediction part of stereotypes, we initially analyzed a survey on the exposure to the TV, made by Eurisko in collaboration with Auditel, the "super partes" company which daily picks up information about TV audience [Auditel, 2000]. We analyzed these information items considering the average audience reception rating (number of average viewers in every minute of a program) and the share (percentage of viewers of a program compared to the overall audience in the same time slot) [Casetti and Di Chio, 1998]. To obtain more detailed information, we decided to merge the Eurisko/Auditel audience data and the information about interests. We assumed an existing correlation between the user's interests and the programs concerning his interests. Moreover, we refined such collected data by comparing it with the audience data of Eurisko Big Map [Casetti and Di Chio, 1998], a sociographic analysis of Italian

society. Finally, we included in the prediction part two temporal dimensions: the watching frequency and the viewing time.

For more details concerning the Stereotypical Expert see [Gena, 2001; Gena and Ardissono, 2001; Ardissono and Goy, 2000]

3.3.3. *Overview of the functionalities offered by the system*

This section describes the functional specification of the Personal Program Guide, which is perceived by the user as a personal assistant offering advanced TV services and helping her to easily manage the record/memo actions. The Personal Program Guide supports multilingual access and the current prototype is accessible in Italian and in English.

In the following subsections, the main facilities offered by the system are discussed, by using as an example the GUI of the prototype developed for demonstrating the Personal Program Guide on desktop environments. While this description is aimed at explaining the main facilities offered by the EPG, the proposed TV interface, to be developed for the TV-Set environment, is discussed in section 3.3.4.4.

3.3.3.1.   **Browsing the Personal Program Guide**

In order to view the EPG, the user has to log on into the system by entering her name and her password. The first time she logs in, she has to register and to provide a set of optional pieces of information, such as socio-demographic data, interests, TV programs preferences, etc.; these data are used to initialize the user model. After the login phase, the system shows the main EPG interface that is mainly devoted to the list of TV programs (visible on the central portion of Figure 3.1). By default, this list is referred to the current time and the TV programs are ranked by taking into account the profile of the logged user.
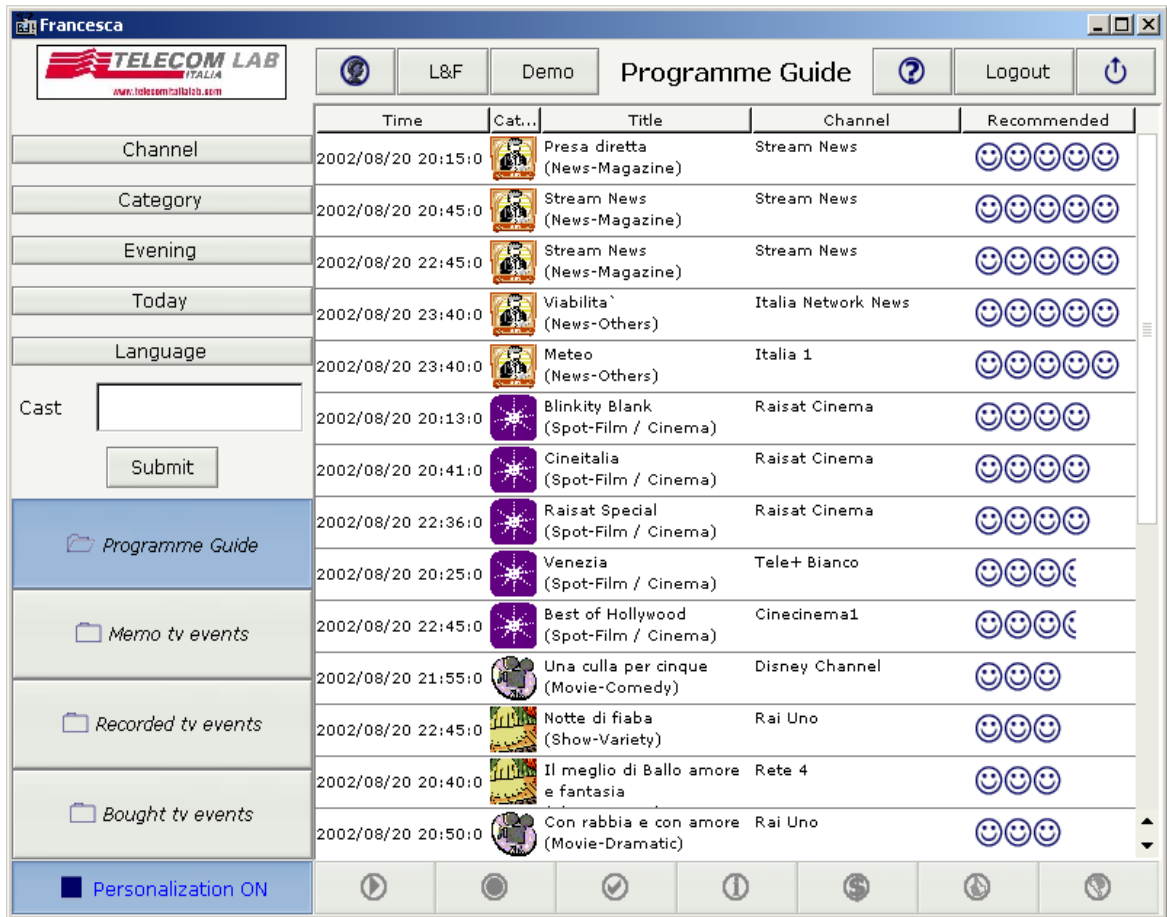
Figure 3.1. Personal Program Guide: PC simulator main window.

Apart this list, which occupies the most part of the display, the interface includes three command areas.

The left area shows:

- six search constraints that can be used to query the program guide (Channel, Category, Viewing Time, Day, Language and Cast);
- a list of buttons that can be used to retrieve the archived programs (Memo TV Events, Recorded TV Events, Bought TV Events);
- a button used to switch between the personalized and the regular program guide (Personalization ON/OFF).

At the top of the interface, from left to right, there are the following buttons:

- a button to access the user-profile (displaying a "persona" icon);
- a button to access the session "Look and Feel" (L&F) to set up different skins for background and colors;
- a "Demo" button to get information about the facilities offered by the EPG;
- three standard buttons (help, logout, exit).

The bottom bar includes, from left to right, a collection of buttons for requesting the main EPG facilities. These buttons are represented by standard icons and are associated to the following actions: Play, Record, Memo, More Info, Buy, Evaluate a program (Like and Dislike buttons). It should be noticed that, when the action associated to a button is not allowed for some reason, the button color is faded. For example, the "Play" option, which enables the user to view a broadcast program, can only be selected after the program has started.

3.3.3.2. <u>Recommendation of TV programs</u>

The user can get the information about TV programs in two modalities. In the default mode, the personalization facilities are used (Personalization ON) and the TV programs are ranked by taking her profile into account: the less suitable programs are filtered out and the most promising ones are shown at the top of the recommendation list (see Figures 3.1 and 3.2). The other mode disregards the personalization. In that case, the list of TV programs is sorted on the basis of the programs starting time. The user may switch from one option to the other by clicking on the left-down button labeled "Personalization ON/OFF".

The degree of recommendation of a program is displayed by showing a set of "smiling faces" close to the main program description. This does not represent an objective evaluation of the quality of the program, but the degree of matching between the program characteristics and the user's preferences, estimated by the system. Since it is not always so clear that by default the best items are ranked at the top of the list, and the user can also sort and filter the programs in different ways [see Section 3.3.3.3], the exploitation of such icons improves the user understanding of the system behavior.
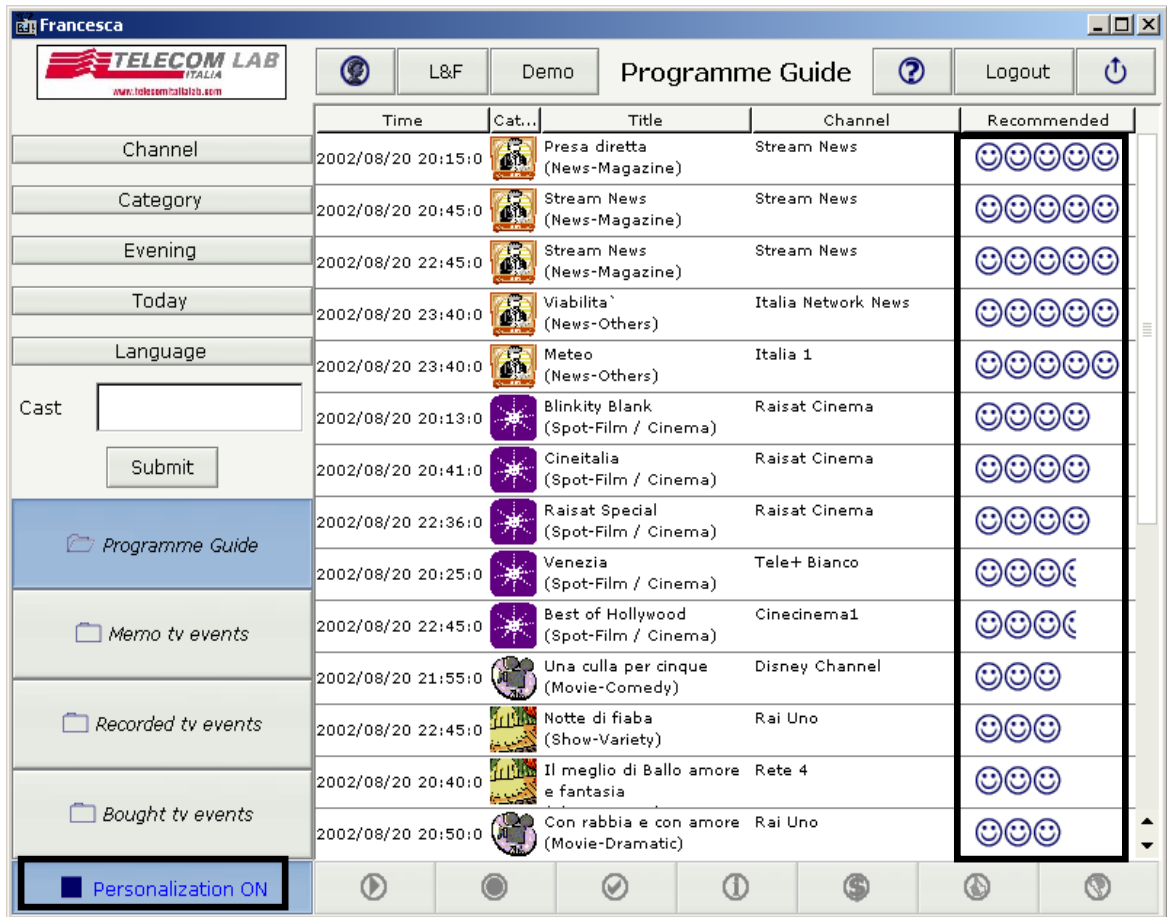
Figure 3.1. Personalization ON and highlighted recommended level column.

To view a program, the user has to select the "Play button" (bottom bar of the GUI), after having selected the program. Then, the guide is turned off and the screen is tuned on the TV mode. If the user selects a pay-per-view program, she also has to start the payment procedures by selecting the "Buy button" before having the program available.

3.3.3.3. Sort and filter

The available programs can be sorted by starting time, category, title, channel and recommendation by selecting the corresponding buttons at the top of the recommendation list: see Figures 3.1 and 3.2. By selecting the search constraints (shown on the left side bar), the user can filter the whole program guide, focusing on a particular channel, category, language, time band (morning, noon, afternoon, evening, night), day and cast. The sort and filter facilities are accessible both in the "Personalization ON" and in the "Personalization OFF" mode.

3.3.3.4. Like, dislike and more information

The logged user can easily ask for more information about a program. By selecting the "More information" button, a pop up window will show details about the highlighted program, if they are available to the system. Such details include the title, subtitle, year, country of production, language, cast, content description, parental rating concerning the program (Figure 3.3). Not all this information is always available, but it depends on the category of the individual program and on the information collected by the system in its local database.

The user can also provide the system with feedback about the recommended programs by rating them. In particular the user can click on the "Like" or "Dislike" buttons after having highlighted a program.
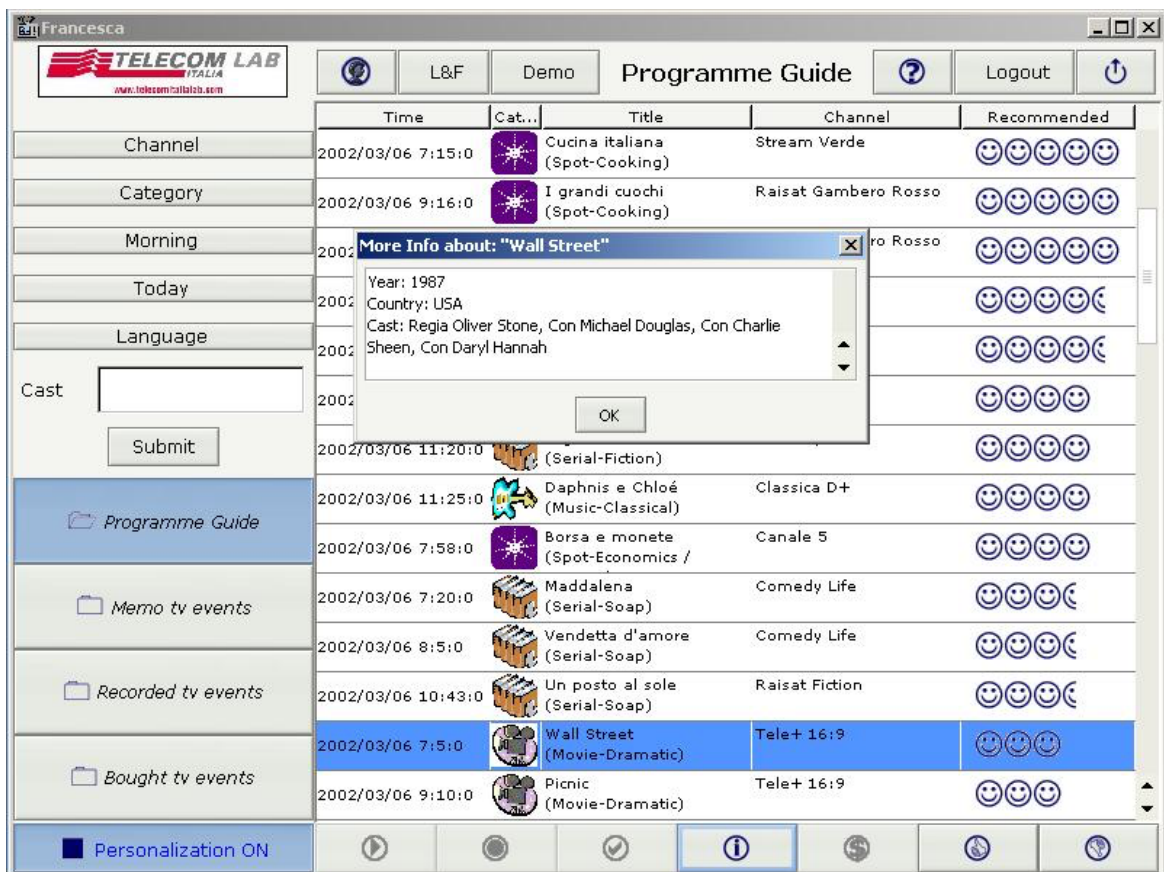
Figure 2.3. More Info about a TV program.

3.3.3.5. Memo and record

If the user selects the "Memo" button, the system will advice her about the start of the highlighted program (see 3.4). By selecting the "Memo TV Events" menu voice, all the requested "Memo Events" are listed. Otherwise, if the user wants to record a program she has to select the "Record" button after having highlighted the program. By

selecting the "Recorded TV Events" voice menu, all the recorded programs are listed. No more than one program can be recorded at the same time, but there are no constrains about the number of concurrent "Memo Events".
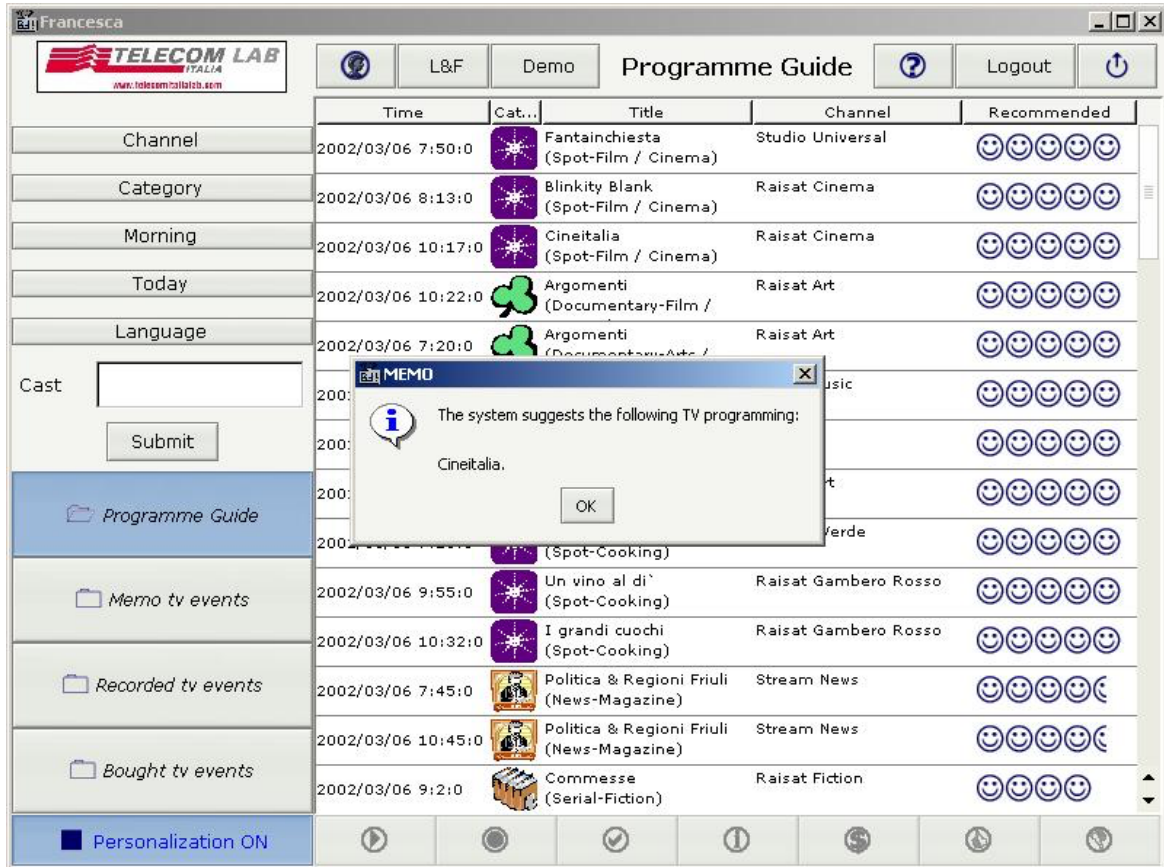
Figure 3.3. Memo suggestion.

3.3.3.6. **Proactivity**

The Personal Program Guides supports the user in the selection of programs and, possibly, focuses her attention on incoming programs she has not considered yet. The user is required to specify which level of initiative she wants the system to display in the recommendations. Three choices for Memo and Record support have been defined: disabled, low support, high support. The system analyzes the set of available programs and, for each registered user, it retrieves the most interesting programs she has not yet considered. When a program is found, depending on the initiative level specified by the user and on the ranking of the program, the system can either autonomously record it or include the program in the user's "memo list". If the user is logged in, the memo and record suggestions are notified by means of an alert window. Otherwise, the memo suggestion is ignored and the record suggestion starts the automatic

recording of the program. In this last case, as soon as the user logs in, an alert message is shown to advise her about the newly recorded program.

3.3.4. Experiments

3.3.4.1. <u>The evaluation of the system's recommendation capability</u>

Since in our system the final recommendation of TV programs is based on the contribution of three different UM Experts, an evaluation of the system recommendation capability has to take into account the different contributions of these three modules. As the complete evaluation of the system is possible only with the system running on a TV box, the evaluation of the Dynamic UM Expert's predictions is not available at moment. Thus, in this formative evaluation, the attention has been focused on testing the Stereotypical UM Expert predictions because the Explicit UM Expert translates in a direct way the explicit user's preferences (and interests) into preferences of the categories introduced in the General Ontology.

The first part of the Stereotypical UM Expert evaluation concerns its capability of properly classifying subjects into the right stereotypes, while the second one concerns its recommendation capability.

Subjects. 62 subjects, 22-62 aged, with different education and different social level. They were all target users of the PPG.

Procedure. The subjects have been interviewed in order to collect their socio-demographic data, their interests and their preferences for TV programs. Then, the gathered information has been entered into the system to evaluate the degree of users classification and the accuracy of the recommendations. The survey was conducted in October-November 2001 and the participants were Italian citizen living in - or in the suburbs of - the city of Turin, in the North of Italy.

Questionnaire. Three main topic areas were identified in the questionnaire: general question, information about general user's interest (books, music, sport, religion, etc) and preferences for categories (Movies, News, etc) and subcategories (Action Movies, Cooking Programs, etc) of TV programs. The final questionnaire was made up of 4 questions where both the questions and the answers were fixed. The questionnaires were auto-filled by the users to avoid any possible interviewer's interferences and gained a week after the distribution. The questionnaire was anonymous and introduced by a written presentation explaining the general research aims. For the items concerning general data participants were required to thick the appropriate answer from a set of gives answers. In the other questions, users had to express their level of agreement with the options concerning the given questions by choosing an item of a 3-point Likert scale.

Measures. Concerning the evaluation of the effectiveness of the stereotypical classification, we have compared the system classification with the classification of an Eurisko lifestyles specialist.

To evaluate the distance between the system predictions and the users' preferences we have calculated the MAE, RMSE. In general, a lower value of MAE means better results in recommendations (see 1.2.5). While to test the accuracy of the selection process we have measured the precision of the collected data (ratio between the user-relevant contents and the contents presented to the user, see 1.2.1).

3.3.4.2. The evaluation of stereotypical classification

The comparison between system classification and the lifestyle specialist the shows that 70% of the users have been correctly classified by the system, while the remaining 30% have been incorrectly classified for two main reasons:

- o the system classification fails when the user's interests are different from those evaluated according her socio-demographic data. For example, if a user *a* has socio-demographic data typical of stereotype *A*, but interests typical of stereotype B, she will be classified as belonging to stereotype A. Then, the prediction will be incorrect because *a* will prefer programs recommended for users belonging to stereotype B. In order to balance the contributions of the two sets of classification data, the interests are also considered as explicit preferences and managed by the Explicit Preference Expert. However, since in the FACTS project [Bellifemine et al, 1999], by analyzing viewing histories noticed that the explicit preferences declared by users are often inconsistent with their own viewing behavior, the system classification (and the derived predictions) are taken into account until new real user data are collected by the Dynamic UM Expert;

- o the data provided by the Eurisko survey does not cover the whole Italian population. For instance the "retired" stereotype is merely referred to low-income users and therefore all the remaining retired users (e.g., retired managers) are not considered. On the contrary, only high-income teenagers (high school/college students) are classified and all the remaining ones are ignored. Therefore this lack of information has to be filled to improve the coverage of the stereotypical KB and the consequent correctness of the system classification.

3.3.4.3. The evaluation of system's recommendations

The TV program predictions generated by the Stereotypical UM Expert have been then compared to the explicit preferences expressed by the users.

In order to compare the Stereotypical UM Expert predictions and the users' preferences, the preferences values generated by the Explicit Preferences Expert have been exploited instead of the mere collected qualitative values (low, medium high). The preferences values generated by the Explicit Preferences Expert are reliable measure of users' preferences because the system does not incorporate further inferences during the generation of these values, but it simply propagates the explicit user preferences in the General Ontology. Moreover, the Stereotypical UM Expert does not take into account the user explicit preferences to generate its own suggestions. Therefore, after having entered the user's socio-demographic data and the interests (exploited by the Stereotypical UM Expert) and her preferences values in the User Explicit Profile, the differences between the values generated by the Explicit Preferences Expert and those ones generated by the Stereotypical UM Expert have been calculated. In this way, MAE is obtained using similar measures.

68 different TV program category predictions have been compared with the corresponding users explicit preferences, with possible values ranging between 0 and 1. The obtained mean absolute error value is 0,26 with a standard deviation of 0,17 (see Table 3.2). This result cannot be considered satisfactory and also the precision is quite low. Probably the MAE value has been strongly influenced by the percentage of subjects (about 30%) incorrectly classified in the stereotypes as described above.

Stereotypical UM expert *vs.* explicit users preferences	MAE = 0,26 Precision=0.40	SD = 0,17	Number of programs compared = 68 per each subject
Main User Model (Stereotypical UM E. + Explicit UM Expert) *vs.* explicit users preferences	MAE = 0,10 Precision=0.51	SD = 0, 071	Number of programs compared = 112 per each subject

Table 3.2. The evaluation of system's recommendations

Since the final recommendations of the system are generated by the integration of the three UM Expert, the distance between the Main User Model recommendations, which incorporates the three Experts' predictions, has been calculated. Again, as explicit users' preferences the values generated by the Explicit Preferences Expert have been considered. Since at the moment are not available sufficient data about the real behavior of the subjects in a sufficiently long time window, the Dynamic UM Expert returns an empty list of preferences. Therefore, the predictions integration has been only provided by the Explicit Preferences Expert and the Stereotypical UM Expert. As in the evaluation described above, the system's predictions are expressed in a value ranging between 0 and 1 and, this time, 112 different TV programs categories predictions have been compared with the corresponding users explicit preferences (the

integration of two experts increases the number possible comparison since more General Ontology's categories can be taken into account).

In this second stage, the obtained mean absolute error value is 0,10 with a standard deviation of 0,071 (see Table 3.2) and the precision value is o,51. The improvement of the MAE confirms our hypothesis about the validity of the integration of different source of information. The main reason of this improvement is due to the management of the incorrect stereotypical classification. As described above, 30% of evaluated subjects have been incorrectly classified into stereotypical descriptions. Most of them have been classified almost in all the stereotypes and therefore the Stereotypical UM Expert has a very low confidence in its predictions, which are too general and too approximately corresponding to the real user's preferences. On the contrary, since the Explicit UM Expert is more confident, its predictions have been more weighted during the final integration. Thus, the final Main User Model's predictions have been closer to the user's preferences than those ones suggested by the Stereotypical UM Expert and this explains the different values of MAE (see Table 3.2). However, the precision value has to be still improved, since the percentage of user-relevant contents does not match our expectations.

Both the contribution of the Dynamic UM Expert and a broader coverage of the stereotypical KB can still decrease the difference between the predictions generated by the system and the user's preferences. Moreover, by adding the Dynamic UM Expert's predictions, a list of real TV programs can be produced and evaluated by the users in order to discover if they really would watch the suggested programs (in the currents experiments only the distance between categories and sub-categories of TV programs evaluated by the users and their corresponding system recommendations have been calculated).

In conclusion, this first formative evaluation has given interesting contribution for a refinement and updating of the system. A broader evaluation will carried out taking into account also the contribution of the Dynamic Um Expert and after having implemented the suggested corrections.

3.3.4.4. <u>**The TV interface design**</u>

In the initial stage of the system, to test the main functionalities of the EPG, the system was implemented in a PC set environment. In order to transfer the system in a TV set environment, the current computer based interface had been modified according to the constraints of the new device and the different kind of interaction. However, the functionalities offered by the EPG are the same, likewise the users events captured by the system.

Following a user centered approach to the interface design we decide to carry out a formative evaluation also for the interface. Three different interfaces prototypes (see 1.1.4) have been designed and these static prototypes have been tested with 24 subjects.

Issue on TV-based interfaces. The interaction within the multimedia world is now so inclusive and sophisticated that the TV interfaces, even the most elaborated ones, cannot compete with the computer based ones. The TV based interactive applications are very different from the computer based ones. The interaction techniques are different: there is no mouse and no object direct manipulation. The users target is wider and less homogeneous: the potential users could be either people that seldom use the computer or do not use the computer at all. Therefore, the computer metaphors should be avoided, since not all the target users are familiar with windows, desktop and menus. Moreover, the context and the focus of the interaction are completely different. The TV is often located in a living room and its social role in the family and its influence in the contemporary society are well known phenomena in sociology, mass-communication, psychology, and so forth.

Until now the research trends in the TV-based interaction have been focused on the integration by simplification and the major efforts have been directed towards the access personalization and the interaction reduction [Visciola, 2001]. Concerning the users' expectations, test carried out highlighted a strong easy-to-use request [Buczak et al, 2002] and expectation of print-like functionality [Baudish and Brueckner, 2002].

Moreover, the user engagement in the interaction with the TV has always been poor, because it has always required not too much initiative: the users are in the "receive mode" and they prefer receiving information than provide it [Norman, 2001]; therefore the interface should not to offer too much functionalities. The ideal interface should be a self-evident interface requiring a moderate user effort to be learned, hopefully without need of printed help.

Following the Mohageg's and Wagner's definition of *information appliance* [2000], an EPG can be described as a computer enhanced consumer device dedicated to a restricted cluster of tasks. The main differences between these kinds of devices and the PC based ones are the less elaborated interface solutions, the wider audience, the lower users' expectations, the request for an easier learning level and the limited purposes of the appliance. Moreover, when an information appliance is designed for an entertainment environment further considerations have to be taken into account: i*)* users are more relaxed; *ii)* experience is more pleasant; *iii)* tasks are less structured; *iv)* users' attention is lower; *v)* interface should not be too much invasive since the interaction with the content is more important then any other factors.

As far as more practical suggestions for the EPG interface design are concerned, Web TV design rules [Nielsen, 1997a-b] can be borrowed in addition to more general usability guidelines.

However, the Web TV has a set of serious design constraints that, can guarantee a good level of usability [Nielsen, 1997b] by limiting the possible (bad) choices in interface design. For instance: *i)* the screen resolution is low and therefore the choices of background colors are limited, *ii)* the movements allowed by the input devices are slow and limited, *iii)* the interaction is more distant, *iv)* the cursor key button are largely exploited as basic tool for the interaction, and so on.

Focusing on the EPG usability issues, a set of useful guidelines arose from the evaluations carried out by the Serco Usability Services [Serco, 2000] and from the analysis of some existing EPGs [Barbieri et al., 2001; Mc Donald et al., 2001].

The interface prototypes. Three interface prototypes have been designed and implemented on a computer screen with low resolution (640 x 480 pixel). The proposed layout is quite similar to that one described in 3.3.3 and modified following the constraints and the design guidelines for TV based interfaces.

Prototype A [Figure 3.5] has been designed as the most complex one. The screen has been divided in five frames:

- o *a left frame* containing the search constraints and the list of menu voices;
- o *a top central frame* grouping the general utilities (user profile, help, exit, logout, etc);
- o *a main central frame* containing the programs list (time, channel, theme, title, suggestion rate);
- o *a bottom central frame* containing a set of buttons for the main EPG functionalities;
- o *a bottom frame* for the paging buttons and for a title bar displaying the purpose of the selected functionality.

The goal of the interface is to collect in a single display both the information about the programs and the main functionalities of the EPG in order to avoid multiple switching.

In prototype B [3.6] the central part of the screen has been completely devoted to the programs list. Compared to prototype A, there is no left frame and therefore the access to the search functionalities and to the main menu voices is available in the top frame. A "*Search*" and a "*Menu*" buttons have been added to connect other screenshots dedicated to these tasks. For example, after having selected the "Research" icon a screenshot dedicated to the advanced research will show the possible research options [Figure 3.8].

In prototype C [Figure 3.7] the strongly recommended programs are emphasized using a larger font size and different colors to make them more visible than the less promising ones. This last solution has been designed to test the effect of adaptation presentation techniques for the personalized version of the EPG [Brusilovsky, 1996].

As described in section 3.3.3.2, in all the considered interfaces the degree of recommendation of a program is depicted by a set of "smiling faces" closed to the main program description and showing the degree of matching between the characteristics of the program and the preferences of the user.
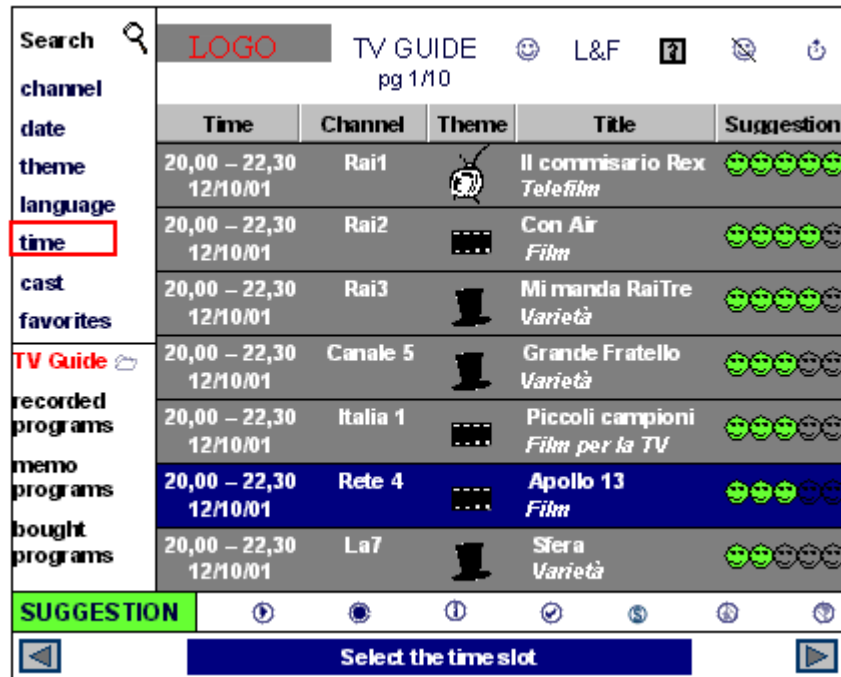
Figure 3.4. Interface prototype A

Figure 3.5. Interface prototype B

Figure 3.6. Interface prototype C

Figure 3.7. Prototypes B-C: the search screenshot

3.3.4.5. Evaluation of the proposed user interfaces

Evaluating prototypes can be particularly effective at solving problems such as developing menu hierarchies that users can understand and grouping and labeling

information. Moreover, these tests allow designers to make changes before it is too late and they make it possible to incorporate user feedback into the early design process. Since benefits can be gained from a user-centered approach, an early and continuous evaluation of the Personal Program Guide system has been planned.

Subjects. 24 subjects, 22-62 aged (a subset of those ones interviewed in the former test)

Procedure. Since the prototypes were static only qualitative evaluations were possible. During the test, after a general explanation of the system, the users were guided to the exploration of the interface through simulated scenarios (a number of paths through the interfaces had been realized to simulate some possible scenarios), then they had to answer questions about the meaning of the icons, buttons and the labels, the grouping choices and their preferences for the prototypes [Figures 3.18-3.20].

Results. The collected results showed that most subjects preferred prototype C (45.45%), followed by prototype A (40.9%), while 13.6% of the subjects preferred prototype B. Since the third and second prototypes have the same basic layout, most subjects (59.05%) preferred the less advanced version of the EPG, which is mostly dedicated to the program list.

Concerning the appraisal of the adaptivity choices, 55% of subjects preferred prototype A [Figure 3.4] and suggested the application of prototype C adaptation techniques [Figure 3.6] to A interface. Therefore, considering this rate and the strong preference for prototype C (45.45%), most users appreciated the proposed adaptation techniques (using higher contrast colors and larger font size to emphasize the most suggested TV programs).

The meaning of the icons has also been evaluated and the less comprehensible ones have been changed following the users advises.

Concerning the users satisfaction and their suggestions, most users were enthusiastic about the EPG functionalities and the main suggestions were: to improve the ease of use; the importance of "More Information" availability about programs; the possibility to have favorite channels list; a clear distinction between "Pay per View" programs and not; the visual icons for categories and subcategories of programs; the possibility to make comparisons among different channels during the same time slot; the audio version of the EPG for blind people; the possibility to have correlated information about the program by means of links (e.g. information about the soundtrack of a movie); the list of the scheduled time slots of a repeated program; the possibility to set transparent and background colors; the possibility to recover the last search results after having selected a program by the "Play" functionality; the availability of different kinds of information for different types of TV programs.

This first prototypes evaluation underlined a request of simplicity (most users preferred the simplest interface) and ease of use. Moreover, the visual cues and the color solutions have been highly preferred to the textual ones. These suggestions are useful to proceed in the development of TV interfaces. Then, a complete usability evaluation should be planned with the system running in a TV-Box and participants interacting with a remote control (or a wireless keyboard) and watching a TV screen. Moreover, to improve the usability of the system and to speed the interaction with PPG, the main functionalities (play, record, more info, paging, home, back, recent etc.) will be also accessible from a remote control, together with the four-cursor keys, as several studies suggested [Nielsen, 1997a-b; Norman, 2001].

Following the suggestion of the above results, our current interface choices are addressed to the further development of prototypes B and C (less complex interface with and without adaptation solutions). Moreover, to offer different background and different style choices to the users, interface skins[47] will be available in the session "Look and Feel" to set up personalized backgrounds and colors. In order to exploit the stereotypical information also for this kind of suggestion, after the first login the system will try to guess a particular skin on the basis of the user classification in the classes of TV viewers. As other studies [Fucs, 2001] already suggested, to provide a more intuitive system interaction, personality styles can be taken into account to satisfy internal needs in addition to functional needs. Since the survey exploited for stereotypes KB [see 3.3.2.1] provides psychological and behavioral descriptions about the lifestyles groups, this information can be used to infer preferences for graphics and styles. Therefore, general statements about stereotypical personalities have been translated into concrete issues that could be relevant to the EPG interface design. By exploiting the two dimensions (doing vs. thinking, strength vs. smoothness, personal goals vs. social interest) used in the Eurisko Big Map [Eurisko, 2000], where the lifestyles groups are placed in a bi-dimensional space, the PPG stereotypes have been grouped into broader classes. For each class a basic skin has been defined, also taking into account other information such as age, social status and gender. For instance, a more colored and modern layout is proposed to younger and personal goal-directed users, while a clearer and classic interface is proposed to older and more reflexive users. However, these are only rough suggestions and the user will be able to change the skin following her personal tastes by accessing the "Look and Feel" session.

---

[47] A skin is a graphic file used to change the appearance of the user interface but not the offered functionalities.

3.3.5. Related work

Several information filtering tools are used to recommend items in Web-based services such as Web stores, electronic libraries and TV listings services; e.g., [Resnick and Varian, 1997], [Greening, 2000], [NetPerceptions, 2002] and (Fink et al., 1998) for an overview. Several examples can be found in the TV programs recommendations field. For instance, MovieLens [MovieLens, 2002] is a research site run by the GroupLens Research group at the University of Minnesota that uses collaborative filtering technology to make recommendations of movies/videos that the user might enjoy. The predictions the user gets are personalized to her tastes, which are learned by asking her to rate at least 10 movies that she have seen before. The core feature of the system is the GroupLens [GroupLens, 2002] recommendation engine that is able to suggest items on the basis of explicit evaluations provided by the user and implicit information gained by the observation of the user's behavior.

However, these systems, based on techniques such as collaborative and content-based filtering for personalizing the suggestion of items, are typically designed as monolithic architectures, which can hardly be modified, or integrated with new modules to enhance their functionalities. Although this is not a problem for the Web search applications they have been designed for, it limits their applicability to other domains. For instance, these systems run on central servers and store the information about users and products in their own databases. Some activities, e.g., collaborative filtering, can currently be performed only in a centralized way, but the decentralization of tasks is essential to the TV world, which imposes privacy issues on the treatment of the users' data and severe constraints on the time needed to download information. For instance, the PTV Listings Service system [Cotter and Smyth, 2000] is a commercial system that generates personalized EPGs accessed by browsers and WAP phones. The system is based on a centralized architecture and uses collaborative and content-based filtering techniques to select programs for the EPG. A profiler agent maintains information about the user's preferences for programs and other, more general preferences, concerning channels, watching time, etc.; as the profiler agent continuously tracks the user's behavior, real-time personalization of the EPG is supported. To overcome the central server problem and the lack of connection with the TV, Smith et al. proposed the integration of PTVplus (the current evolution of PTV available through PC, PDA and WAP phones) and GuideRemote [Smyth et al., 2002a], an interactive universal remote control that integrates Internet and the LCD technology. From PTVplus the user can download the personalized guides into the GuideRemote, which combines the facility of a handheld EPG and a remote control. On the return path, the GuideRemote can capture the user's selection and PTVPlus can use this information as grading source. On the recommendation quality point of

view, another interesting feature offered by PTVPlus [Smyth et al. 2002b] is the exploitation of data mining techniques to extract hidden relationships between programmes in PTVplus user-profile cases.

Further specific commercial systems have been developed for digital TV. The Singularis S.3P system [Singularis, 2000] generates personalized EPGs that can be accessed both on the Web and from the TV device. Moreover, it offers a personalized video recording facility to autonomously download on digital VCR the programs assumed to be of interest to the user. TiVo [Tivo, 2000] and Replay TV [SONICBlue, 2002] support a personalized management of digital VCRs, however, they differ in the type of preferences used for selecting programs. TiVo generates customized listings and automatically records programs, on the basis of the user's choices and explicit feedback, but it also tries to reason about the user feedback to produce better suggestions. In contrast, Replay records TV programs on the sole basis of the user's explicit choices.

On the side of different user modeling techniques integration in the TV domain, Buczak et al. [2002] combined recommendations from various constituent recommendation algorithms to improve the user's trust in their TV show recommender. They fused the implicit recommenders (based on individual and household viewing history) and the explicit one (based on explicit users' preferences) using a neural network. The final results showed that such a fusion network performs well for users it has not yet encountered. Van Setten et al. [2002] combine prediction techniques to optimize the personalization of TV programs. Their system chooses a techniques taking into account three factors that can cause the dynamic of personalization: the usage lifecycle, the information lifecycle and the system lifecycle. Their test results indicated that the prediction strategies improve prediction quality by using that best combination of techniques in a particular situation.

The exploitation of stereotypes-based techniques is not new in the user modeling field [rich] and it seems to be particularly suitable to address the *cold start problem* by achieving a quick and effective way of initialize a user modeling system. Concerning the use of TV viewer stereotypes, Kurapati and Gutta [2002] proposed the exploitation of stereotypical patterns of TV shows derived from a sample set of users observed for a period ranging from 5 months to 2 years. They derived stereotypes by applying clustering techniques to the view history data set. A new user has to choose her closest stereotypes to start her profile, which will evolve during the time towards a specific profile. To avoid users having to set up their profiles by going trough all the programs genres, Barbieri et al. [2001] let the users select a predefined stereotype such as Movie Lover, Film Freak, and Documentary Buff. Goren-Bar and Glinansky [2002] proposes the Family Interactive TV system that filters TV programs on the basis

of stereotype groups. Their stereotypes are classified according to age and occupation groups. During the first login phase, the user has to grade the programs categories and to tell what is her probability to be in front of the TV in any given 2-hour slot to generate a preferences/hour vector. In order to not annoying the user with the need to identify each time she wishes to watch TV and to manage the problem of multiple users at once, the system tries to guess the current user(s) and suggests several programs on the basis of the preferences/hour vectors generated by the system. Another system [Pearson, 2001] tries to address the problem of the mandatory login phase proposing a speech interface solution by a close-talking microphone to recognize the user.

3.4.   *Evaluating an adaptive web site*

This section describes an adaptive web site [Gena et al., 2001a] and the results of its two different evaluations. Indeed, after having performed the first empirical evaluation [Gena et al. 2001a; Gena, 2002], we decided to implement a team of presentation agents to support the adaptivity [Gena et al., 2002]. This choice was led by the fact that most of the users we tested had detected the difficulty to recognize some adaptive behavior, if nobody explained it before. But when they were informed by the experimenters about the adaptive features of the site, they found them very useful. Therefore, we thought to substitute the human presentation by a team of agents and then test the new components of the adaptive site.

3.4.1. Introduction

On the one hand, the advent of high-speed Net access in the offices and industries, due to the broadband diffusion, makes the web-based applications an effective alternative not only to traditional software applications but also to standard office tools. Free email, scheduling software, address databases, calendars and other web-based utilities are often used instead of standard desktop applications. Moreover, these tools have the advantage of being accessible from any computer with Internet connection. It is now a long time since the major portals like Netscape, AOL, Yahoo, Excite, Lycos started to offer to the users this kind of free services because they early understood the potential of such an important competitive advantage.

On the other hand, the big amount of information presented in a large number of web sites let often the users "lost in the hyperspace" [Conklin, 1987] and unable to quickly find the relevant information. In these cases the exploitation of personalization techniques can help the users to filter out the information on the basis of their needs and preferences. Most of the commercial web sites that offer some kind of personalization are called *adaptable*: the user is in control of initiation, proposal, selection and production of adaptation [Kobsa et al., 2001]. For instance, MyYahoo, MyExcite, MyLycos gives the chance to set the basic site layout by letting choose to the user the preferred items in the preferred order. Another way to personalize is to examine the behavior of the user visiting a web site and autonomously adapt the interaction on the basis of her inferred interests. These system are called *adaptive*. Adaptivity and adaptability often co-exist in the same application in order to allow the user correcting or adding the adaptive feature and checking her interest profile. On the web, not many example of these systems are now available. For instance, the UM2001 web site [Schwarzkopf, 2001] attempts to recognize the various visitors' interests and offers shortcuts to potentially relevant documents, reminders to

important deadlines etc. Due to the complexity of the system it is probably slow for most applications and because of its structure the user must interact quite a while before appreciating the adaptivity.

It is well known that the adaptable sites require an additional user effort that can cause the paradox of the active user [Carroll and Rosson, 1987]. Users often refuse to visit the sites that impose to respond to an interview first because they would save time getting their immediate task done. Moreover, users are usually uncomfortable in answering to personal questions and Manber et al [2000] observed that the majority of users do not customize the web pages when they have the possibility to do it. On the contrary, users require quick answer to their questions that only a tailored solution can give. Unfortunately, the most of adaptive sites require a period of interaction before being able to show the personalization to the users.

This paper describes the development and the preliminary evaluation of an adaptive commercial web site offering a wide range of services and tools commonly exploited by whoever uses a computer for professional purposes. Because of this large amount of different web-based applications supplied by the system, we decided to personalize the interaction with the user in order to tailor services to the individual user's needs. Thus, we designed our system as a *personalized hypermedia application*, a system which adapts the content, structure, and/or presentation of the networked hypermedia objects (web pages) to each individual user's characteristics, usage behavior and/or usage environment [Kobsa et al., 2001]. The system is a rule based adaptations system mainly based on usage frequency data of the user. In order to evaluate how the adaptivity increases the success in retrieving information and reduces the amount of actions needed to solve the tasks we have made a test comparing the site with and without adaptivity.

3.4.2. The services

After a preliminary questionnaire given to fifty people to know the real needs of potential users, we developed E-tool[48], a commercial web site offering a wide range of tools for working users. The aim of the system is to catalogue in a homogeneous environment, a web site, a set of working tools that are often available in different part of World Wide Web. As in the computer desktop, all the most used applications are always available to the user, also in our system all the tools are always present as menu items in a "speedbar" [Debevc, 1993]. Thus, the user could save time having an easy access to all the necessary working tools using the site as a dynamic web-based desktop.

---

[48] http://www.e-tool.it.

The site is targeted to a specific type of users, particularly for professional users who seek for services and tools now available on the web but usually exploited in non computer-based applications. For instance, instead of searching an unknown foreign word on a traditional dictionary, the user can find meaning and translation directly at a mouse click distance.

The main classes of potential users are: employees, secretaries, managers, lawyers, and business consultant. Anyway, everyone can find some useful tool to improve her work and study or to organize her free time.

In the English version, we classified the offered tools into fifteen general categories on the basis of their explicit function:

- o Utilities: to gather information about the European low, the council union, etc;
- o Calculators: to calculate km, loans;
- o Telephone: to find phone number, names, business address;
- o Traveling: to find addresses, maps, departures, arrivals;
- o Business: marketplace, search job, euro pages;
- o Translations: to translate words, expressions and sites in different languages;
- o Converter: to convert currency, measures, etc
- o Diary: an electronic scheduler for appointments, meetings, etc..;
- o Reminder: to remind automatically by e-mail important dates, appointments, etc;
- o Search engine: to search the web;
- o News: on line newspapers collection;
- o Stock market: quotations and financial news;
- o Weather forecast;
- o Links: a list of useful web sites.

The site is engineered to be used like a basic tool for general purposes automatically tailored for each individual user. We decided to exploit personalization techniques in order to provide value to the customers and therefore create long-term relationships with repeat users.

3.4.3. The development of the system

Following Kobsa's classification of *personalized hypermedia application* (Kobsa et al., 2001), we divided our personalization process into these three major tasks.

3.4.3.1. **Acquisition Method and Primary Inferences.**

This task is aimed to identify and to gather the information necessary to construct an initial user model. In order to accomplish this purposes, we decided not to force the

user to enter information as interest, preferences, and knowledge about the domain, etc... Therefore, we have chosen to show immediately all the contents. If the user decides to create and maintain her personal version of E-tool she has only to provide some demographic data (name, surname), her e-mail and username and password for the login. If she want set by herself the preferences she can do by accessing to the session "Your profile" which permits to add bookmarks. It is important to underline that in our system the user cannot manage personalized site views [Brusylovsky, 2001] as commonly in the adaptable sites, but she can only add bookmark which will be highlighted by applying adaptation presentation techniques as described in the session "*Adaptation Production*". In fact, the system is mainly a browsing oriented adaptive annotation system that attacks visual cues to the link in order to help the users to select the most relevant one [Brusylovsky, 2001].

In any case, the system gathers the usage information necessary to the user model by observing the user behavior. Thus, the user model is mainly based on usage data instead user data. The system also uses heuristics to determine positive and negative evidence of user's interest [Mladenic, 1999]. We assume that the information items can be divided in two different classes: interesting and non- interesting. The links selection is considered as positive example, while the non-selected pages are considered as negative example in the sense of user disinterest [Kobsa et al., 2001]. The system also keeps track of the origin of the link selection (speedbar, left menu, main page, secondary pages) to have an automatic evaluation of the system usability and to make more refined inferences regarding the user model. This collected data are stored in a database and accessible to the system in order to be processed to construct the secondary inferences.

3.4.3.2. __Representation and Secondary Inferences.__

This task is aimed to represent the acquired user information appropriately in a formal system, let them available for further processing and draw further assumptions about the user. In our system this process is generated every time the user log on into the system and the user model is consequently runtime built. We decided not to make immediate changes during the same user session, but refresh the user model only once per session. This choice is led by the fact that we didn't want to confuse the user by adapting the site during the same session.

The usage statistics are transformed into explicit assumptions and the system learns about the user by processing them, and after a given period of time makes adaptations based on the inferred assumptions. This process is carried on by combining a set of inference rules and a set of selective queries on the collected data. The system extracts from the interaction history the recently most frequently used

pages and the possible explicit user's preferences. While the links selection is medium indicator of interest reinforced by the usage frequency, the user's explicit preferences are a strong indicator of interest. On the basis of the inferences, the system assigns the corresponding priority value to the selected pages and creates an adaptive short list of bookmarks sorted by these indications. This priority value corresponds to a confidence value assigned by the system. Moreover, if the user uses the bookmarks proposed by the system, the interest indicator will grow and consequently the confidence value. The bookmark list is comparable to the list of recently used files at the end of the "File" menus, in that it provides an automatic shortcut for probable actions (Debevc et al., 1997].

The bookmarks list will be exploited also to generate personalized messages based on the similarity of the features of suggested items and the features of items the user liked in the past by applying a feature based filtering technique. We distinguished two kinds of personalized messages: commercial suggestions (commercial offers and banners) and recommendations that try to anticipate the user's need. These last one are aimed to highlight the links that could interest the user on the basis of her profile. The exploitation of recommendation is also a way to partially skip the possible wrong assumption generated by not considering the negative examples. In fact, it is common to overlook a page and therefore classifying objects not visited as negative examples sometimes could led to dangerous assumption [Schwab et al., 2000]. The recommendation is randomly selected between a list of tailored suggestion and changes every session.

The bookmarks list is available to the user in the section "Your profile", which allows the user to manually correct and update it, if it is necessary. In this way the user always has the control of the system adaptivity. However, the user can also ignore the adaptive modifications by choosing to make the interaction anonymous.

3.4.3.3. <u>**Adaptation Production.**</u>

This task is oriented to the generation of contents, presentation and structure adapted on the basis of a given user model. Before discussing this session we want to describe the layout of the web site in order to introduce the structure of the interface. We designed a simple interface [Nielsen, 2000] structured in a set of frames (3. 9).
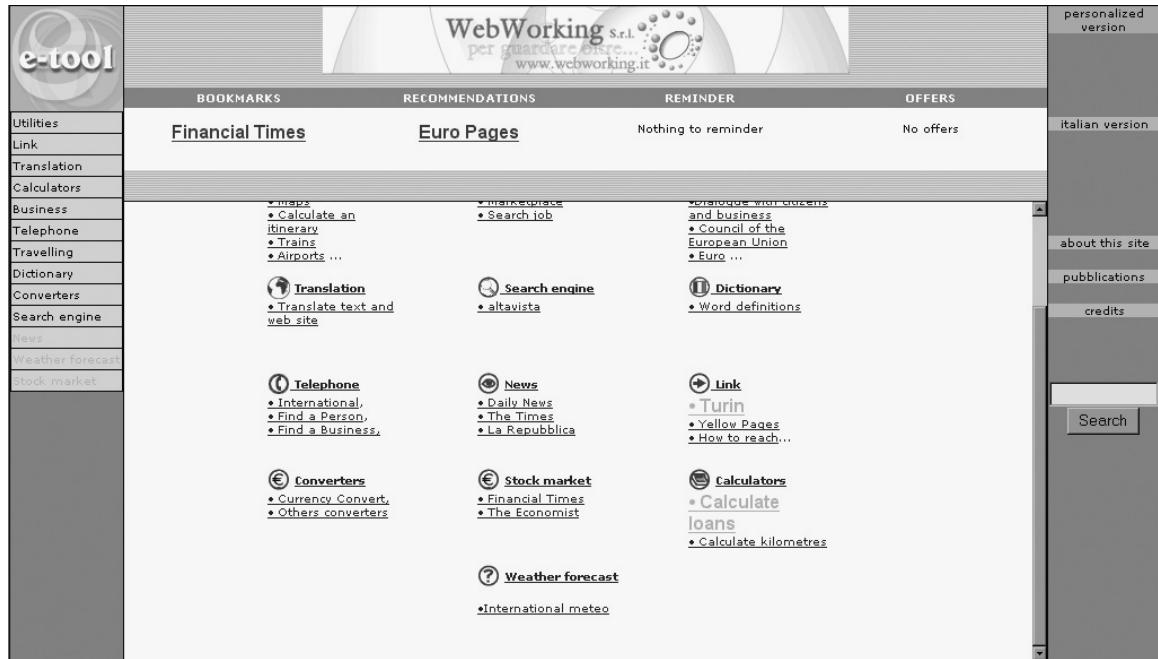
Figure 3.9. The layout of the home page.

- o a top frame containing a personalized message box,
- o a left frame containing the list of main categories services,
- o a central frame containing the list of links to the effective services,
- o a right frame containing information about the site and other utilities (the search form, the login form, the link to "Your profile", etc).

The adaptation of the content is realized by presenting personalized messages as described above. The messages are shown in the top frame that is handy for the user and therefore designed as a speedbar. In this frame are also available

- o a set of shortcuts to the most used utilities,
- o the recommended link,
- o a commercial offers tailored on the supposed preferences of the user,
- o the day's event to remind,
- o a keyword-marketing banner.

Moreover, in the central page we have added a personalized welcome page message.

The adaptation of the structure is realized

- o by applying the adaptive link annotation [Brusylovsky, 1996] to the bookmarks in the central frame : the corresponding links are highlighted with different font color and size in the pages that appear in the central frames;

o by applying the adaptive link sorting [Brusylovsky, 1996] to the left frame menu where are gathered the links to the available utilities grouped by functions. On the basis of usage frequency and user interest, the links are ranked each session in a more refined way. In contrast, the less used or unused links will be lightly faded but not disabled.

While the other adaptations are mainly aimed to improve the navigation between pages, the adaptive link annotation is aimed to improve the effect of information presented within a page. Every adaptation of the site is managed by a set of parametric adaptation rules based on the user profile.

3.4.4. The knowledge base

The structure of the domain is explicitly described in a knowledge base founded on an inheritance net and implemented in a database. This is a conceptual representation of the services categories that describes their features and their relations in a inheritance net. Figure 3.10 shows a portion of the taxonomy. Every category has as sons the respective subcategories of "Suggestions" which have the unique function of grouping in a more refined way the services of a category in order to generate more precise recommendations. In fact, the suggestions groups don't appear to the user, which see the site organized in groups of categories and their respective services. For instance, the category traveling has these services: Find Address, Maps, Itinerary, Train, Airports in Europe, Turin Airport, Malpensa Airport, Frankfurt Airport, London Airport, Paris Airport. In the taxonomy, these services are grouped in three more precise subcategories (Cities, Airports, Alternative traveling) to make more refined inferences on the user preferences, but the user is unaware of this classification.

Every node of the net is conceptually a frame: every feature of the node is represented by a set of slot. For instance, the final node, which always represents the description of the effective service, is characterized by these slot: *service name, its own category, its own suggestion groups, the file path, the description of the service.*

This taxonomy is easily re-usable in other domains which can be described as a inheritance net and which could easily take advantage from a similar system. For instance, a portal is usually organized in different sections containing, services, tools, etc. On line newspaper are characterized by a set of standard sections containing different kinds of news. Moreover, the inferences rules applied to generate the user model are parametric. Thus, the system could be easily "adapted" to other domain without strong efforts.
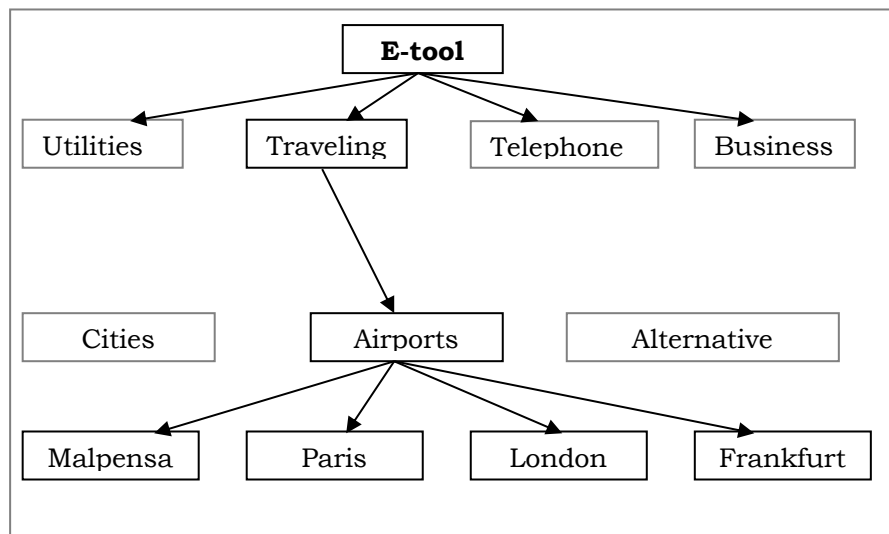
Figure 3.10. A view of the services taxonomy.

3.4.5. The test

While most evaluations of adaptive systems are comparisons of the system with and without adaptivity, where the last one has often not been designed optimally for the task [Höök, 1997; Espinoza and Höök, 1996], the two versions of E-tool are quite similar. As described above, the most relevant changes in the adaptive version are in the top frame (personalized messages) and in the left frame (sorted menu items). Our choices have been also led by the fact that we want to leave the interface predictable after the adaptations. The surprising changes and the unpredictable behavior of the system disorient the users and cause the negative effects that could compromise the effectiveness of adaptive systems [Shneiderman, 1997].

Anyway, to demonstrate that adaptive system can improve usability it can be shown that without the adaptive capability the system would have performed less effectively (Benyon, 1993]. Therefore, our test is aimed to measure the task completion time, the amount of within page navigation and finally the satisfaction of the users. Since our system shows the adaptation after a certain number of interactions in a given period of time, we decided to simulate this scenario instructing the users in a pre-test phase and let them just "click around" to make the interaction more natural. Moreover, we structured our test by a repeating task: at the beginning of the test the user has to accomplish a task in the non-adaptive version (NAV) and at the end of the test she has to repeat the same task in the adaptive version (AV) to test the effectiveness of the adaptive changes.

Subjects. 14 subjects, 24-35 aged, all with high knowledge of the Internet and web browsers and all using the Internet during job time. These subjects are the kind of users that the site was projected for.

Procedure. The subjects were split in two groups (seven subjects each) and randomly assigned to one of the two groups. Everyone had to solve eight tasks by exploiting utilities given by the site (for e.g. "Look for the telephone number of John Red" or "Translate this word." , and so on).

Experimental tasks. Group 1 had to solve four tasks in the NAV and four tasks in the AV. On the contrary, Group 2 had to solve the last four tasks of Group 1 in the NAV and the first four tasks in AV in order to compare the results of the two versions.

The solutions of the four tasks in the AV could be reached by exploiting the supposed *facilities* offered by one of the applied adaptation techniques (see Figure 3.9):

- o   a sorted menu-item (left frame) or a annotated link (central),
- o   a annotated link (central frame),
- o   a recommendation (top frame),
- o   a bookmark (top frame).

Every group had a repeated task: the first task in NAV was repeated at the end in the AV (bookmark) to simulate the real running of the system. After the test, we asked the subject eight question about their viewpoint of the system.

Measures. Task completion time and subjects' satisfaction (questionnaire responses).

Experimental design. Crossed 2 x 4 factorial design with factor A (the presence/absence of the adaptation) represented as between-subjects variable and factor B (the 4 different tasks to carry out) as a within subjects variable ("**crossed" mixed within-subjects factorials design,** see 1.1.5.5).

Therefore, the result can be analyzed as two separated mixed within-subjects factorials design by considering two crossed groups (Table 3.3):

- o   the former one, where the four tasks (**Tasks b1-b4**) in the AV are completed by subjects of Group 1 and then compared with the same tasks completed in the NAV by subjects of Group 2 (see the thicker arrow of Table 3.3);
- o   the latter one, where the four tasks (*Tasks b1-b4*) in the AV are completed by subjects of Group 2 and then compared with the same tasks completed in the NAV by subjects of Group 1 (se the dashed arrow of Table 3.3),

	Task b1	Task b2	Task b3	Task b4	Task b1	Task b2	Task b3	Task b4
Adaptive (a1)	S1 – G1	S1 – G1	S1 – G1	S1 – G1	S1 – G2	S1 – G2	S1 – G2	S1 – G2
	S2– G1	S2– G1	S2– G1	S2– G1	S2 – G2	S2 – G2	S2 – G2	S2 – G2
	S3– G1	S3– G1	S3– G1	S3– G1	S3 – G2	S3 – G2	S3 – G2	S3 – G2
	...	...	...	...	...	...	...	...
Non Adaptive (a2)	S1 – G2	S1 – G2	S1 – G2	S1 – G2	S1 – G1	S1 – G1	S1 – G1	S1 – G1
	S2 – G2	S2 – G2	S2 – G2	S2 – G2	S2– G1	S2– G1	S2– G1	S2– G1
	S3 – G2	S3 – G2	S3 – G2	S3 – G2	S3– G1	S3– G1	S3– G1	S3– G1
	...	...	...	...	...	...	...	...

Table 3.3. "Crossed" mixed within-subjects factorials design.

Before this test, we had made a preliminary evaluation with another 14 subjects that showed some errors in the first design of the interface and in the first choices of adaptation. Particularly, the adaptive link annotation was only applied to the general categories of most used items (e.g., the system annotated "Telephone" instead of "Find number") and the items of the speedbar were not enough visible (too little font size, not high contrast between background and foreground). Moreover, the different color of the annotation often resulted less visible compared to the standard links color.

For these reasons, the results of this first test didn't show relevant differences between the AV and the NAV. Therefore, we decided to highlight every single most used link with a higher contrast color and to emphasize the and re-design the speedbar. After these changes we made the test described before of which the results are shown in the next session.

3.4.6. Results
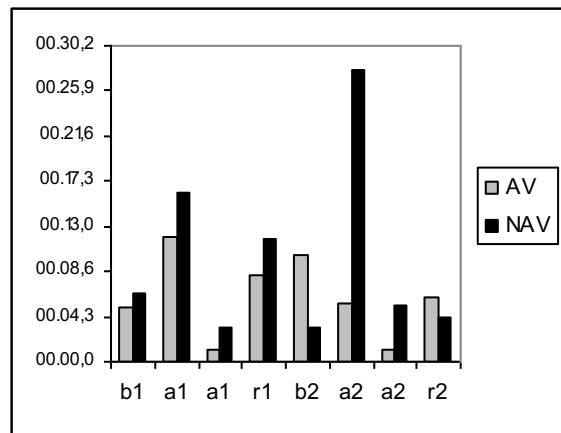
Figure 3.11. The test results (average seconds per task).

The results of the test are shown in Figure 3.11 (n. 1 means that Group 1 used the AV, n. 2 means that Group 2 used the AV) and analyzed in the following Section.

3.4.6.1. Adaptation in Group A

ANOVA

$F_A(3,36) = 2.05$

The main effect of factor A is not significant, and we have retained the null hypothesis. So, the presence/absence of adaptations does not change the subjects' performance.

$F_B(3,36) = 1.60$

The main effect of factor B is not significant, and we have retained the null hypothesis. So, the effect of different tasks does not change the subjects' performance.

$F_{AXB}(3,66) = 0.06$

The obtained value of F does not exceed the critical value, so we can conclude that statistically significant interaction effects are not present in the data. The presence of adaptations in the different tasks does not affect the completion times. Since the interaction is not significant, **main effects** (see 1.1.5.4) have to be investigated.

MAIN COMPARISONS

$F_{Acomp}(1,12) = 2.05$

The difference between the average ratings given by the presence or not of adaptations is not significant.

$F_{Bcomp}(1,36) = 12.14$, $p<0.01$

In this case, I compared the different average values obtained by tasks b2 and b3 and tasks b1 b4 in both level of A. I grouped the tasks in this way since in the adaptive version tasks b2 and b3 could be completed by means of annotations techniques, while tasks b1 and b4 (bookmark and recommendation) by means of facility displayed in the top frame. In this case the differences are significant. Therefore, the tasks b2 and b3 obtained significant different results compared to the other ones because of the nature of the tasks. In conclusion, the presence of annotation techniques (b2 and b3) is not responsible of the different results, but the nature of the tasks affects the subjects' performances.

3.4.6.2. **Adaptation in group B**

ANOVA

$F_A(3,36) = 1.14$

The main effect of factor A is not significant and we have retained the null hypothesis. So, the presence/absence of adaptations does not change the subjects' performance.

$F_B(3,36) = 2.70$

The main effect of factor B is not significant and we have retained the null hypothesis. So, the effects of different tasks do not change the subjects' performance.

$F_{AXB}(3,66) = 3.11$, $p<0.05$

The obtained value of F exceeds the critical value, so we can conclude that statistically significant interaction effects are present in the data. So, in this case, the presence of adaptations together with the different tasks affects the completion times. Therefore, in this second comparison both the presence of adaptation and the different tasks affect the completion time and **simple effect and simple comparisons** (1.1.5.4) have to be investigated.

SIMPLE EFFECTS

$F_{A \ at \ b1}(1,48) = 3.01$

$F_{A \ at \ b2}(1,48) = 226.68$, $p<0.01$

$F_{A \ at \ b3}(1,48) = 8.92$, $p<0.01$

$F_{A \ at \ b4}(1,48) = 21.69$, $p<0.01$

The between-subjects simple effects show that there are significant differences between the performances of the two groups in tasks b2, b3, and b4. So the presence of adaptation techniques changes the subjects' performance for these tasks. However, in case b4 the completion time in AV is higher, so the positive contribution of

adaptation techniques is only present for task b2 and b3 (where annotation technique is exploited).

$F_{B \text{ at a1}}$ (3,36) = 7.70, p<0.01

$F_{B \text{ at a2}}$ (3,36) = 68.32, p<0.01

The within-subjects simple effects show significant variability between the different tasks at both levels of A.

SIMPLE COMPARISONS

$F_{A \text{ comp. at b1}}$ (1,48) = 3.01

$F_{A \text{ comp. at b2}}$ (1,48) = 226.68, p<0.01

$F_{A \text{ comp. at b3}}$ (1,48) = 8.92, p<0.01

$F_{A \text{ comp. at b4}}$ (1,48) = 21.69, p<0.01

As in case of simple effect, the between-subjects simple comparisons show that the presence of adaptation produces significant differences between the performances of the two groups in tasks b2, b3, and b4.

$F_{b1-b2 \text{ comp. at a1}}$ (1,36) = 0.03

$F_{b1-b3 \text{ comp. at a1}}$ (1,36) = 0.73

$F_{b1-b4 \text{ comp. at a1}}$ (1,36) = 0.25

$F_{b2-b3 \text{ comp. at a1}}$ (1,36) = 0.47

$F_{b2-b4 \text{ comp. at a1}}$ (1,36) = 0.45

$F_{b3-b4 \text{ comp. at a1}}$ (1,36) = 1.85

$F_{b1-b2 \text{ comp. at a2}}$ (1,36) = 10.07, p<0.01

$F_{b1-b3 \text{ comp. at a2}}$ (1,36) = 0.06

$F_{b1-b4 \text{ comp. at a2}}$ (1,36) = 0.03

$F_{b2-b3 \text{ comp. at a2}}$ (1,36) = 9.55, p<0.01

$F_{b2-b4 \text{ comp. at a2}}$ (1,36) = 11.36, p<0.01

$F_{b3-b4 \text{ comp. at a2}}$ (1,36) = 0.07

The within-subjects simple comparisons show that the significant difference within tasks are only present at level A2 and concerns the differences among task b2 and the other ones (b1, b3, b4).

3.4.6.3. <u>Conclusion</u>

The calculation of ANOVA in the two groups sketched in Table 3.3 shows contradictory results. The results calculated in 3.4.6.1 show no significance interactions, while the

results calculated in 3.4.6.2 highlight significant interaction effetcs between the two factors (in particular due to annotation techniques). Moreover, in this latter case, the adaptivity reduces the difference among different tasks within the same group.

To solve the contradiction a further evaluation is required, involving an higher number of subjects.

However, after having observed the users and having discussed with them about their experience with the adaptive site, we believe that link annotation (a) is more effective than the other techniques here presented. The users scan the page content and if some chunk of information is highlighted the attention is focused on it. This technique is also effective because when users often exploit a utility they tend to remember the general collocation of the link and by emphasizing it improve the retrieval of its position. By observing the users we noticed that the link annotation decreases the within page navigation since the user learns the meaning of the annotation.

The exploitation of bookmarks (b) was related to the repeated task. At the end of the test the users repeated the first task with a difference: the task began from the last viewed page (the other start from Home). This change is due to the fact that the expected utility of bookmarks increases when the user is in another page and instead of going back Home she can click on the bookmark in the top frame. Five users of Group 1 used the bookmarks and the other two went back Home. Instead, two users of Group 2 exploited the bookmarks, the other four went back Home and only one clicked on the left menu. In fact, Group 1 had better results then Group 2.

Only one user per group used the recommendations to solve the third task and therefore the differences are not relevant.

Subjects	Which version would you use on line?	Which adaptation is it more effective ?	Is it personalization useful?
1	personalized	annotation	yes
2	personalized	annotation	yes
3	personalized	annotation	yes
4	personalized	bookmarks	yes
5	personalized	annotation	yes
6	personalized	annotation	yes
7	personalized	annotation	yes
8	personalized	bookmarks	yes
9	non-personalized	bookmarks	no
10	personalized	bookmarks	yes
11	non-personalized	annotation	no
12	personalized	annotation	yes
13	personalized	annotation	yes
14	personalized	bookmarks	yes

Table 3.4. The final interview of the test.

The final interview showed that the most of the users were satisfied with the site and preferred the AV (only two didn't like it). They found adaptation useful to accomplish repeated tasks and to avoid information overload. They mainly preferred the link annotation and only four voted for the bookmarks that are generally not immediately perceived, as recommendations. All of them but one found the left menu extraneous to the rest of the site (never used in the required AV tasks) and suggested adding icons and stretch-text containing the subcategory links, and removing the fade and the sorting techniques.
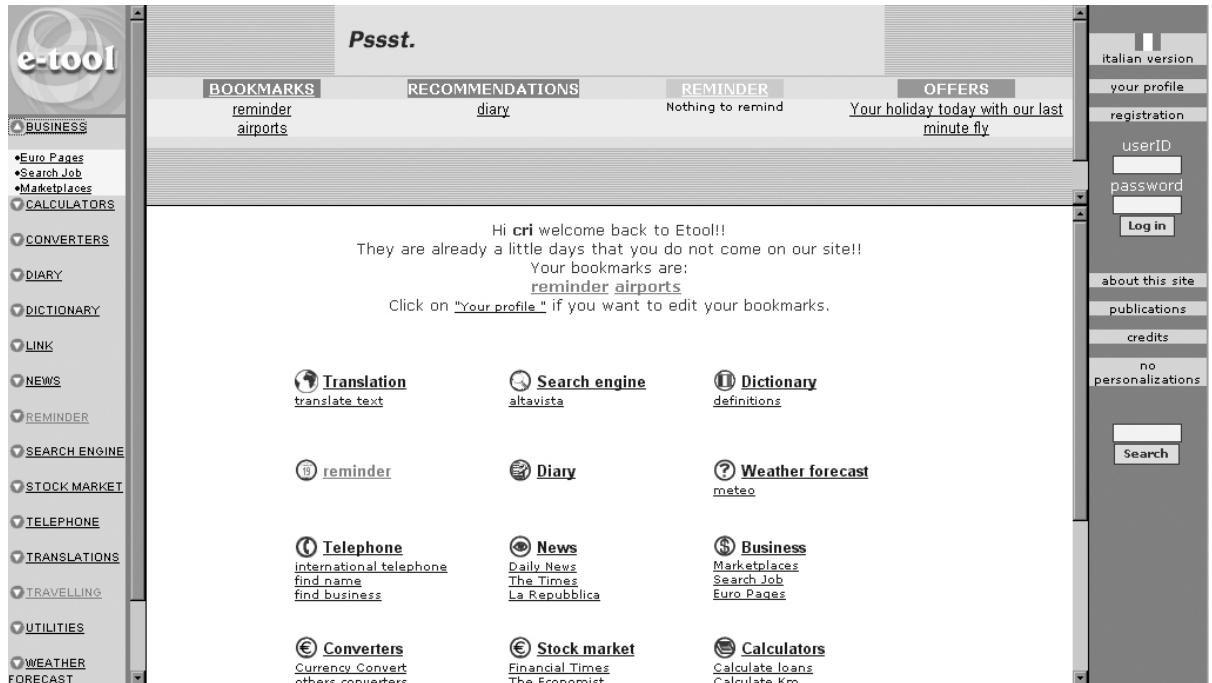
Figure 3.12. The new version of the site.

3.4.7. Discussions

These indications are useful but not exhaustive. We have made the changes suggested by the users trying to cover the lack of our design (see Figure 3.12). Particularly,

- o concerning the left frame, we have removed the fading and the sorting techniques. We have added stretch text to every categories link containing the list of subcategories items. Moreover, we have annotated the link to the most used categories and subcategories by applying the same kind of annotation used in the central frame in order to keep the consistence;

- o concerning the speedbar, by using different colors we distinguished the bookmarks, the recommendations, the reminder and the offers. We also reduce the number of possible presented bookmarks near to the Miller's number 7 ±2 [Miller, 1956] in order to reduce cognitive overload of the users;

- o concerning the central frame, after the welcome message, we added the bookmarks list and the link to the user profile in order to give to the user the possibility to immediately change the list. We repeated the list because the test showed that the user attention is initially focused on the central zone of the page. Moreover we extended the link annotation technique to all the pages describing the subcategories.

We are also studying new possibilities to improve the adaptation perception of the current version of the site. In addition, we want to test the role and the effectiveness of the commercial suggestion and their impact on the users. Therefore, we decided to continue to evaluate the site also by analyzing the on line behavior of its real users to prove its utility in a longer period of time.

In addition, we decided to implement and to test the introduction of a team of interface agents that could help the user in the recognition of the adaptive behavior of the system. This new features of the site and its evaluation is described in the following paragraph.

3.4.8. *The second evaluation*

3.4.8.1. __Introduction__

The exploitation of interface agents is a controversial issue both in HCI [Norman, 1998; Shneiderman, 1998] and in AI [Maes, 1994; Lieberman, 1997]. Following the assumption that humans treat computer socially [Reeves and Nass, 1996] the main goal of developing interface agents has been the replication of human-human communication. Several types of interface agents have been developed. These agents can be either human-like [Cassel, 2000] or more cartoon-like [Lester et al, 2000] and they can have a full body or only a face.

On the one hand, the failure of both "Microsoft Persona Project" and the "Microsoft Office Assistant" has probably shown the unsuccessful exploitation of some of these interaction techniques. Moreover, the management of agents instead of the direct manipulation of objects is a question still unsolved in HCI [Kay, 1990].

On the other hand, studies have shown that using personas can improve the users' satisfaction with the system providing a more personal and social interaction [Maes, 1994; Vassileva and Okonkwo, 2001].

While the traditional interface agents' design is oriented toward conversational interfaces, where the user and the agent "take turns" acting, a new trend of research has been focused on the development of presentation agents. André and Rist [2000] introduced the notion of presentation teams. They proposed performances given by a team of characters as a new form of presentation. The basic idea is to communicate information by means of simulated dialogues that are observed by an audience. Descamps, Ishizuka et al. [Ishizuka et al, 2001] created a language called Multimodal Presentation Mark-up Language (MPML) to command interface agents to do presentation so that users can easily add such attractive presenters on their web pages.

Previous evaluations of the site [see 3.4.6] had shown that most users did not immediately perceive some of the core adaptation components (shortcut to the most used utilities, content-based recommendations, personalized commercial offers), probably due to the more complex learning of the system behavior. However, after a human presentation given to the subjects at the end of the test, they found these components useful to their browsing goal. Therefore, we decided to replace the human presentation with an agent presentation in order to increase the short term learning of the site. Then, we tested the site with this new component and the results are here described.

3.4.8.2.   **Character Behavior**

Since we wanted both users to learn more about the system and they enjoy their visit to the site, we decided to exploit a team of agents. Moreover, as André's and Rist's [2000] study suggested, we also believe that people can learn more about a subject matter if they are willing to spend more time with a system.

We exploited the Microsoft Agent™ package [Microsoft, 1999] that includes a programmable interface to predefined characters. We chose the wizard Merlin and the parrot Peedy. These characters are able to move on the computer screen focusing the user's attention on a particular point and to talk aloud using a Text-To-Speech engines (TTS engines). In addition, all the words appear in a *word balloon* [Figure 3.13]. We characterized them with different personalities in order to create more brilliant dialogues and to have more defined roles. We exploited the different personality traits deriving from the opposition "adulthood/youth". Merlin is older and therefore is wiser, more serious and rational. He behaves like a real wizard using a magic wand and talking in a magician way. Peedy is a parrot that needs to be taught. Since he is young, he is curious but also inexpert.

Figure 3.13. A screenshot of the general presentation.

To give him a stronger characterization, we portrayed him as a pirate's parrot lost after a sinking, now surfing in the Internet Sea. Suddenly, he meets Merlin that will be guiding him in the site exploration. This fantastic script is aimed to improve the emotional experience making these virtual agents a psychological entity that should support and guide users in a real experience. This could be considered as an example of *affective computing*, where the agents' performance is realized in an emotional virtual world [20].

3.4.8.3.   **The Agents' Performance**

The agents' performance has been divided in *i)* a general presentation of the site and its adaptive features and *ii)* a personalized presentation. The former presentation is

aimed to focus on the adaptive features that are probably not so usual and familiar to the users. Moreover, letting users know about the adaptive changes could be a way to avoid possible disorientation. This general presentation is coded in a fixed script and is optional. The user can voluntary decide to watch it and he is always in the control of the agent's behavior [Norman, 1998] because he can easily switch them off. The presentation is organized as follow: *i)* a brief description of the site structure, *ii)* an explanation of adaptivity and *iii)* a detailed explanation of the adaptive changes of the site.

On the opposite, the personalized presentation is shown only after the login and it is a free optional choice for the user. Merlin manages this performance when some adaptive feature appears. For this second part we decided to exploit only one character, because no explanations are needed. After a personalized welcome message, he advises the user about the adaptive changes, also showing their position in the page layout and particularly focusing on the top frame, where most adaptations are concentrated. This script is not fixed but is managed by the same set of inference rules exploited to generate the adaptive behavior of the system. For instance, after a given number of interactions in a given period of time a link is added to the user bookmarks list. As consequence, the link will appear in the bookmarks list at the top of the page and will be highlighted with different font color and size both in the central and in the left frame. In addiction, a content-based suggestion will appear at the top of the page along with content-based commercial messages (offers and banners). The first time a bookmark is added, the user is unaware of this process. Thus, Merlin will describe all these adaptive changes. However, since the complete presentation is quite long, we hypothesized that after a couple of times the user will know the adaptive behavior of the site. Moreover, a long presentation could be a waste of time for the user. Therefore, after a while Merlin will perform only the welcome message and the eventual daily appointments to remind without showing the adaptive features described above. If some other adaptive change occurs (e.g., the addition of a new bookmark), the user will be able to notice the new adaptive features. Anyway, in every moment the user can switch off the agent option by accessing to the session "Your profile".

3.4.8.4. **Evaluation of Presentation Agents**

Figure 3.14. The non-adaptive web site.

Subjects. We evaluated 24 subjects, 20-26 aged, with a medium-high knowledge of Internet and web browsers and good computer skills.

Procedure. The study was conducted in the Computer Science lab of the Communication Science building, Turin University. The experiment ran on high-end Pentium PC's with color monitors. The subjects were randomly assigned to the experimental and the control group. The first group ran the experiment using the presentation agents (Group A), the second one without (Group NA).

Experimental tasks. During the first part of the test Group A saw the general presentation performed by Merlin and Peedy. Then, after a couple of minutes of letting them just "click around", they were introduced by the Merlin's personalized presentation in an adapted version of the site. The adaptive changes of the site were correlated to the four tasks the users had to accomplish during the second phase of the test. After that, they had to accomplish four tasks in another non-adaptive site called Xtorino, a portal web site about the city of Turin. Before carrying out the task, the main features of the site were introduced by a Merlin and Peedy performance. Finally, they had to fill in a questionnaire regarding their experience. The users of Group NA had to solve the same set of tasks with the same site except the agents' presentation was turned off. They had more time to become familiar with the site and

to simulate a real condition and they were asked to carefully read the on line explanations about the site.

As in the first experiment described in 3.4.5 the four tasks in the adaptive site were related to the exploitation of supposed *facilities* offered by one of the applied adaptation techniques (see Figure 3.9):

- o a annotated link (central frame),
- o a recommendation (top frame),
- o a bookmark (top frame),
- o the presence of a commercial offer (top frame).

Measures. Task completion time, number of clicks to complete a task, the link source to reach the goal and subjects' satisfaction (questionnaire responses).

Experimental design: 2 x 4 factorial design with factor A (the presence/absence of the presentation agents) represented as between-subjects variable and factor B (the four different tasks to carry out) as a within subjects variable (**mixed within-subjects factorials design,** see 1.1.5.6).

3.4.8.5.   <u>Adaptive site results</u>

	Annotation	Bookmark	Recommendation	Offer
□ EXPERIMENTAL	11.76	4.30	19.16	23.65
■ CONTROL	5.99	5.41	12.76	47.22
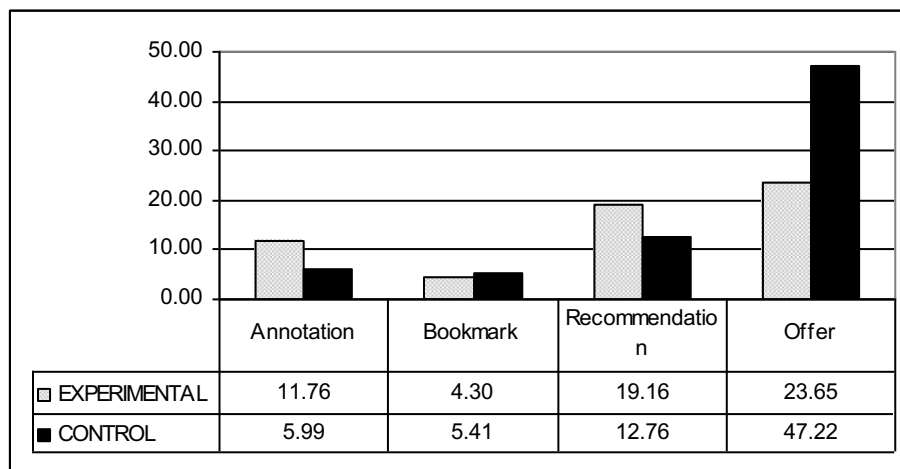
Figure 3.15. E-tool test results (average seconds per tasks).

ANOVA

$F_A(3,66) = 0.25$

The main effect of factor A is not significant, so we have retained the null hypothesis. The presence/absence of the agents does not change the subjects' performance.

$F_B(3,66) = 4.67$, $p<0.01$

The main effect of factor B is significant, so we have rejected the null hypothesis. The effects of different tasks change the subjects' performance.

$F_{AXB}(3,66) = 1.25$

The obtained value of F does not exceed the critical value, so we can conclude that statistically significant interaction effects are not present in the data. So, the presence of agents in the different tasks does not affect the completion of the tasks. Since the interaction is not significant, **main effects** (see Chapter 1.1.5.4) have to be investigated. However, as shown in 3.15, since the task concerning the commercial offer has obtained slightly different average times, we were curious to know if this difference were significant. So, we decide to investigate also **simple effect and simple comparisons** (see 1.1.5.4)**.**

MAIN COMPARISONS

$F_{Acomp}(1,22) = 0.36$

The difference between the average ratings given by the presence/absence of the agents is not significant.

$F_{Bcomp}(1,66) = 3.43$

In this case, I compared the different average values obtained by the annotation technique (first task) and the other three techniques applied (bookmarks, recommendations and offer) that are all displayed in the top frame. The differences are not significant.

SIMPLE EFFECTS

$F_{A \text{ at } b1} (1,88) = 1.44$

$F_{A \text{ at } b2} (1,88) = 0.05$

$F_{A \text{ at } b3} (1,88) = 1.77$

$F_{A \text{ at } b4} (1,88) = 23.97, p<0.01$

The between-subjects simple effects show that the unique significant difference between the performances of the two groups concerns the forth task (commercial offer). Therefore, only in this case the agents produce significant differences.

$F_{B \text{ at } a1} (3,66) = 25.32, p<0.01$

$F_{B \text{ at } a2} (3,66) = 15.63, p<0.01$

The within-subjects simple effects show significant variability between the different tasks at both levels of A (presence/absence of the agents).

SIMPLE COMPARISONS

$F_{A \text{ comp. at } b1} (1,88) = 2.07$

$F_{\text{A comp. at b2}}$ (1,88) = 0.07

$F_{\text{A comp. at b3}}$ (1,88) = 2.5

$F_{\text{A comp. at b4}}$ (1,88) = 34.52, p<0.01

The between-subjects simple comparisons show that the presence of the agents produce simple significant differences only in case of task b4 (the offer task).

$F_{\text{b1-b2 comp. at a1}}$ (1,66) = 0.50

$F_{\text{b1-b3 comp. at a1}}$ (1,66) = 0.59

$F_{\text{b1-b4 comp. at a1}}$ (1,66) = 1.29

$F_{\text{b2-b3 comp. at a1}}$ (1,66) = 2.02

$F_{\text{b2-b4 comp. at a1}}$ (1,66) = 3.42

$F_{\text{b3-b4 comp. at a1}}$ (1,66) = 0.18

$F_{\text{b1-b2 comp. at a2}}$ (1,66) = 0.003

$F_{\text{b1-b3 comp. at a2}}$ (1,66) = 0.42

$F_{\text{b1-b4 comp. at a2}}$ (1,66) = 15.55, , p<0.1

$F_{\text{b2-b3 comp. at a2}}$ (1,66) = 0.49

$F_{\text{b2-b4 comp. at a2}}$ (1,66) = 15.98, p<0.1

$F_{\text{b3-b4 comp. at a2}}$ (1,66) = 10.85, p<0.1

The within-subjects simple comparisons show that there are significant differences among different tasks (B) only at level a2. All the significant results concern the differences between b4 (the offer task) and the other tasks, while in a1 the presence of agents reduce these differences.

3.4.8.6. <u>**Non-adaptive site results**</u>

	Find movies	Find hockey	Find CVs	Find artist's authorization
☐ EXPERIMENTAL	4,58	26,46	36,97	42,21
■ CONTROL	7,14	24,73	18,64	40,15
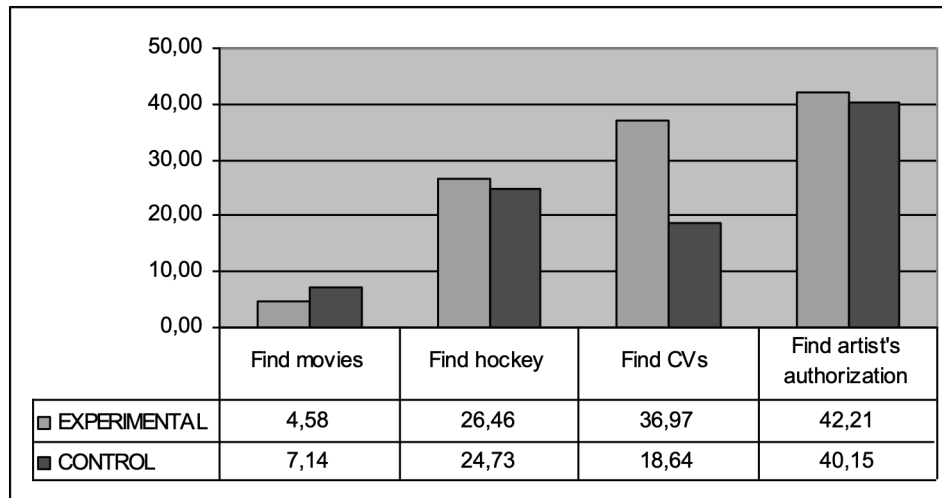
Figure 3.16. Xtorino test results (average seconds per tasks).

ANOVA

$F_A(3,66) = 0.13$

The main effect of factor A is not significant, so we have retained the null hypothesis. The presence of the agents does not change the subjects' performance.

$F_B(3,66) = 9.61$, $p<0.01$

The main effect of factor B is significant, so we rejected the null hypothesis. The overall effects of different tasks change the subjects performance.

$F_{AXB}(3,66) = 1.07$

The obtained value of F does not exceed the critical value, so we can conclude that statistically significant interaction effects are not present in the data. The presence of agents in the different tasks does not affect the completion of the tasks.

Since the interaction is not significant, **main effects** (see 1.1.5.4) have to be investigated.

MAIN COMPARISONS

$F_{Acomp}(1,22) = 0.56$

The difference between the average ratings given by the presence or not of the agents is not significant.

$F_{Bcomp}(1,66) = 34.33$, $p<0.01$

In this case, I compared the different average values obtained by the first task and the other three tasks. The particularity is that only the first task is placed in a list of quick links on the left side of the home page (see Figure 3.14). The difference is significant, so we can conclude that links placed on the Quick Links Menu can contribute to obtain significantly faster results.

3.4.8.7.   __Conclusions__

Concerning the completion time, there were no significant different results between the performance of Group A and Group NA both in the adaptive and in the non-adaptive site [Figure 3.15 and 3.16]. The only exception was for the last task regarding the commercial offer in the adaptive site. The subjects were asked to find the offer, which was collocated in the top frame, and the Group A was significantly faster then Group NA.

Moreover, before filling in the questionnaire, we checked their information retaining about the top frame asking them the content of the daily event to remind which is collocated close to the commercial offer. 50% of the subjects in the experimental group remembered it against the 17% of those ones of the control group.

Regarding the exploitation of adaptive techniques to reach their goals, we were particularly interested in the influence of the agents on the users' browsing behavior. The results showed that:

- o   92% of users in Group A used the suggested annotation in the central frame and 67% of the Group NA;
- o   50% of users of Group A skip the regular collocation of the link and exploited the bookmarks collocated in the top frame. Only 16% of control group used this utility;
- o   no users of both groups used the content based recommendation;

As our first test partially demonstrated [see 3.4.6], the annotation technique quite effective. However, this technique only highlight the links in their regular collocation while the bookmarks and the recommendations are collocated in the top frame of the page layout dedicated to the personalized messages.

The results showed that a good number of users of the experimental group exploited the suggested bookmarks, but no one used the recommendation. A reason could be the higher familiarity of Internet users with the exploitation of bookmarks.

Concerning the results of post-test questions, 75% of the subjects judged the agents helpful, and 50% of them wanted to have the agents in the site. Most complaints regarded the too long agents' performance.

The comparison between the two sites showed that 42% of the users preferred the agents in the adaptive site, 16% in the non-adaptive, 33% would have the agents on both sites and 8% neither.

3.4.8.8.   **Discussions**

The final results of the test showed that the exploitation of presentation agents partially modified the users' browsing behavior. 50% of users in the experimental group noticed the bookmarks, but they completely skip the content-based recommendation. Their recollection about the event to remind and the commercial offer of the top frame was quite good. This means that probably the agents' presentation reinforced the users' attention on this part of the page layout.

In conclusion, our results proved that the exploitation of agents' presentation could be qualitatively successful in more complex systems where a longer training phase is required to learn how to use them (e.g., an adaptive web site vs. a regular web site). Furthermore, the agents' presence has not to be invasive and time-consuming for the users and their exploitation must always be optional.

Moreover, we believe that this kind of technique could also be successful both in entertainment and in educational web sites where younger users are involved as other studies have already demonstrated [Vassileva and Okonkwo, 2001].

3.5. *Evaluation of an on-vehicle adaptive tourist service*

This last Section is dedicated to the description of an on-vehicle adaptive tourist service and the evaluation exercise we performed [Console et al., 2003].

3.5.1. Introduction

The goal of providing large amounts of information at the driver's fingertips is becoming more and more important in the last decade. The dashboards of modern cars include, and in some cases integrate, an on-board computer, a GSM telephone, possibly with GPS, a navigation systems, besides more common devices such as radio, CD/DVD player, etc. The presence of these electronic systems, and the fact that people are spending more and more time on cars, suggested car manufacturers the design of new systems for providing to the various types of services on board the car (news, information about facilities and tourist locations, etc). The availability of these kinds of systems represents an opportunity but could also become a problem for the driver: the services may be very useful or even necessary, but the use of the systems may distract the driver causing serious dangers for active and passive safety. This led car manufactures and providers of these systems to study the design of the systems from the point of view of ergonomics and human computer interaction. In the last few years some researchers suggested that *adaptation* and *personalization techniques* can play a very important role in vehicle on-board applications and can significantly contribute to solve the problems previously mentioned. Believing in the potentials of such techniques applied to this area, in the 2000 we started the study of a framework for on-board adaptive systems and implemented a prototype, MASTROCARONTE, which exploits adaptation and personalization techniques to provide *tourist information* to a driver (see [Console et al. 2001], [Console et al. 2002] for a discussion on the framework and a description of the system).

Given the complexity of adaptive systems, due also to the discretionary choices they unavoidably carry out, evaluation is considered a very important subject in the user modeling and adaptive systems community (see [Chin 2001], [Chin and Crosby 2002]). As recommended by the paradigm of user-centered design [Norman et al. 1986], it should be performed already in the design phases to get immediate feedbacks from users. For such a reason we started an evaluation exercise of the first prototype of MASTROCARONTE. As a specific objective of our evaluation, we aimed at showing that indeed adaptation and personalization can contribute to the achievement of *two major goals*: *i)* checking whether the first items suggested by the system are indeed in accordance with the user's preferences and needs and *ii)* whether the mode and format of the presentation is in accordance with the user's features and contextual situation.

3.5.2. The system under evaluation

In the past two years we defined a framework and architecture for on-board adaptive systems, implementing a prototype application MASTROCARONTE, for providing tourist information to the driver [Console et al 2002]. In the following we analyze the principles of the framework and architecture, providing concrete examples taken from MASTROCARONTE.

On board adaptation is based on *explicit models* of the user and of the context. The former includes some features that are application independent (e.g., general preferences but also cognitive characteristics such as visual capabilities, ability to capture information on a screen) and some that are application dependent (e.g., interests or propensity to consume are relevant for the tourist services domain). The context model includes many pieces of information such as the location of the car (from a GPS) as well as driving conditions and context (e.g., type of road, traffic, speed, weather, time of the day, presence of passengers; travel information such as duration, distance from home, direction, …). These pieces of information can be obtained from sensors available on the car (e.g., speed or travel information) or inferred (e.g., the presence of passengers can be inferred from the seat belts sensors; the traffic conditions can be estimated from the position, the date and time of the day and the speed).

All three forms of adaptation (content, interface, behavior) are equally important and the system should infer as much information as possible about the user autonomously, without requiring a lot of interaction. This means that the model must be refined or revised continuously according to the user's behavior. We decided to base the initial model on stereotypical information and then track the user's behavior to collect data for revising the model. In the case of MASTROCARONTE the stereotypes have been derived from a psychographic research about the cultural behavior of the Italian population, while the possibility of learning from the user's behavior is based on a log of the interactions with the system (specifically, of the tourist facilities on which the user requested details or which she called or for which she required information about the route or which she visited). Statistics derived from this log are used by a set of rules that refines the user model accordingly.

The same system must support multiple users of a car and the same user can use the system on multiple cars. Therefore, we decided to store all information about a user (and in particular the user model) on a smart card that can be inserted in the dashboard on any car on which the system is installed.

An on-board adaptive system should be based on a distributed architecture with a client on the car and servers accessed via a GSM or GPRS connection. This means that exchange of information should be minimized. In the case of MASTROCARONTE, the

tourist databases are located on the server but are also replicated on the car on a CD/DVD. Since in this way the information on the car may be out of date, we devised a protocol such that only the pieces of information that are out of date are transferred from the server to the car. The tourist database contains information about hotels, restaurants and tourist locations; each item is annotated with extra information to be used for personalized selection.

The intelligence is distributed among several agents, some of which are on-board the car and some of which are on the server; in particular:

1) The agent that is in charge of "behavior" adaptation is on-board and is mainly in charge of deciding if and when the system must activate autonomously and the type of service/request that is most useful to the user at each activation. MASTROCARONTE activates in special conditions, according to the user's preferences; for example it may suggest an hotel suitable for the user if it is late at night, the user is traveling since many hours and is not directed toward home.

2) The agents for "content" adaptation are distributed. One is located on-board and is in charge of (i) contacting the server, sending relevant portions of the user and context model and (ii) once a response from the server is received, filtering out or ranking the information/services to be displayed to the user, given the full user and instantaneous context models. Another agent is located on the server and receives requests from the car, performing a first filtering and ranking of the services/information to be sent to the car.

3) The agent for "mode of interaction and interface" adaptation is located on the car and selects the most appropriate mode of interaction, given the user model and the context and driving conditions. In MASTROCARONTE the alternatives are the following: a vocal interface (to be used in situations when the driver is alone and cannot be distracted) a graphical interface, with five different styles, ranging from a very simple one in which the items in the reply are presented one at a time, to more complex ones. Also the number of extra functions (automatic connection to the phone, connection to the route planner, bookmarking) is adapted to the user preferences and context. 3.15 contains a scheme with the main components of our architecture, instantiated on the MASTROCARONTE application.
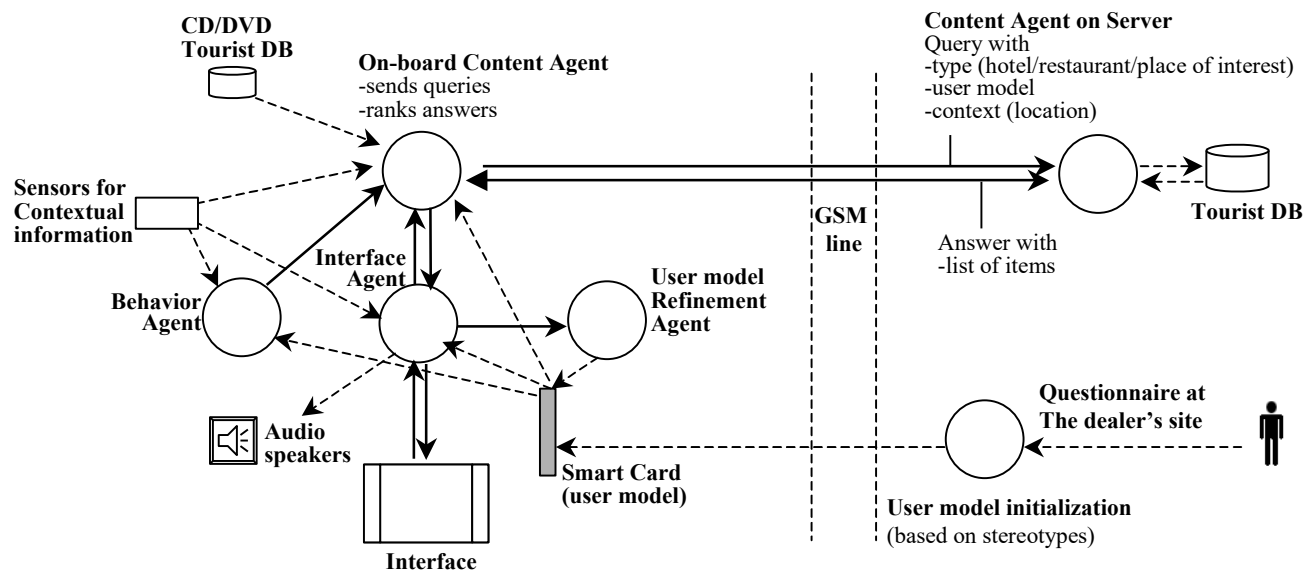
Figure 3.17. Architecture of MASTROCARONTE

MASTROCARONTE has been implemented under the VxWorks real time operating system running on Magneti Marelli car navigation systems. In particular, it runs on a laptop simulator of this operating system. The agents and server components are implemented using the CLIPS rule-based system and the exchange of information is based on XML. The service database is a multipurpose relational one and, for the evaluation, it has been populated with tourist information about the Turin area. It contains about 400 restaurants and 150 hotel, and also information about places of interest, which, however, have not been involved in this preliminary evaluation.

For the purpose of the evaluation exercise in the next sections, it is worth spending a few more words on the user model, the interface and on the rules used by the content adaptation agent and interface agent to rank items (hotels and restaurants) and to decide the presentation format.

The *USER MODEL,* implemented as a Bayesian Network, contains several pieces of information about the user and is initialized starting from stereotypes about the behavior of the Italian population[49], which provide probabilistic information about the level of interest of classes of users for aspects such as: traveling, art, history, nature, visiting museums. They also provide information about the propensity to spend and consume, whose values highly contribute in the recommendation of hotels and restaurants. Finally, the user model contains information regarding the user's receptivity (i.e., the ability to capture information, estimated from parameters such as the user's age, her stereotypical classification, familiarity with electronic devices and interfaces and visual problems). The initialization is performed off-line, e.g., at the dealer,

---

49 We defined the stereotypes starting from a psychographic study (Sinottica Eurisko) about the Italian population.

Examples of descriptive features are: age, school degree, job, geographic area, etc.

storing the result on the user's personal smart card. These initial estimates are then revised by the system by performing analyses on the actual behavior of a specific user: MASTROCARONTE tracks the actions of the user and updates her model based on statistic taken from this log of actions. This process may also produce more detailed information that cannot be estimated in general, such as the preferences as regards the type of food or the habits regarding hotels and restaurants.

The *INTERFACE* for presenting information to the user includes two media: audio speakers and a screen. As regards the screen we defined a number (five in the prototypes) of presentation styles. Each style defines a format (number of items to be displayed, fonts, colors) as well as the number of extra services that the user can access; these services include the possibility of calling a restaurant/hotel, the possibility of getting detailed information about it, the possibility of asking for the route to reach it, the possibility of bookmarking the item (these services are very useful for getting feedback about the user's behavior).

The *RANKING OF ITEMS* provided by the server is based on a set of rules that make two kinds of evaluation. The first one provides a score to each hotel or restaurant, according to the user's characteristics, namely, her propensity to spend (coming from the predictions in her model) and an estimation of her preferences, such as preference for a type of food or a type of place (computed with another set of rules). The second evaluation regards the contextual information (time of the day, distance, type of area – metropolitan, extra urban, etc.) and is used to weight the previously selected items.

The *SELECTION OF THE INTERFACE,* medium and format, is more complex as it involves several parameters, such as the user receptivity and preferences (but also estimates of the user tiredness), the driving conditions (e.g., speed, traffic), contextual information (e.g., time of the day, weather, etc.). This is achieved by means of a set of rules that operate in steps, first deciding the medium, then the format.

In order to make the process clearer, let us consider an example of interaction. In case of a query concerning restaurants, the agent on the car sends information about the location of the car and about some of the user's features and preferences (preferences about food, propensity to spend). The agent on the server retrieves the restaurants that may be of interest, sending a list of references to the car. The agent on the car asks for up-to-date information for those restaurants whose reference is unknown. Then it ranks the list according to the complete user model and to the instantaneous context conditions. Finally, it passes the ranked list to the interface agent.

In the following we show, as an example, the layouts loaded by the system with three different users/contexts:

o   a university student with high level of receptivity (young, familiar with electronic devices, no visual problems, etc.), medium/low propensity to spend, while driving at a low speed with no traffic (Figure 3. 18)

o   the same student, while driving at a high speed, but with non traffic and in a straight way (Figure 3.19),

o   a middle age lady with high propensity to spend (Figure 3.20).

In all the case, the screenshots on the left show the ranked list of selections, displayed according to the selected format and style; the screenshots on the right show the details of a restaurant, using again the selected format and thus activating different sets of services. The screenshots are exactly as shown on the dashboard's display of the car.

Figure 3.18 – Recommendations for the user 1 (university student) in context 1 (low risk level)

Figure 3.19 – Recommendations for the user 1 (university student) in context 2 (medium risk level)

Figure 3.20 – Recommendations for the user 2 (middle age lady) in context 1 (low risk level)

3.5.3. The evaluation and its methodology

In the UM community the importance of systems evaluation has been strongly advocated (see [Chin 2001], [Chin and Crosby 2002], [Petrelli et al. 1999]) and now it is a shared principle. As regards the automotive environment, in our opinion, it requires considering *several aspects*: first of all, of course, the matching between the real users preferences and the features of the items suggested by the system. Second, the correct weight to external conditions, like distance, time pressure, etc., in the selection of the items. Finally and most important, the analysis must evaluate if the system adapts its content, presentation and behavior in order to, and respecting the requirement of, being safe and not intrusive. Since the system is still under development, we performed *i)* a *formative evaluation*, which is aimed at checking the initial choices and getting clues for future revision, concerning the knowledge base implementation and the correctness of adaptation rule; *ii) a predictive evaluation*, based on HCI experts estimation, concerning the interface design choices.

To obtain reliable users' data we needed an accurate and quick way to collect self-reported information from target users. Thus we decided to exploit a questionnaire we personally distributed to

Subjects and sampling procedure. 107 users identified following a proportional layered sampling strategy, where the population is divided into layers, related to the variables that have to be estimated, and containing each one a number of individuals proportional to its distribution in the target population. We identified eight groups characterized by different age, sex, education, job, technology expertise, geographic area, etc. (that are the same descriptive data used by the system to classify each user). Every group identifies a potential user of the system. For instance group s1 (the 5% of our sample) is characterized by age: 26-35, sex: male, education: high school, job: autonomous workers, technology expertise: medium, etc.; while *group* s8 (the 23% of our sample) is characterized by age: 36-45, sex: male and female, education: high school/degree, job: manager/ professionals, etc..

The questionnaire. To obtain the desired information the questionnaire collected two sets of data: (a) information useful to the system to classify users and to generate recommendations and interfaces adaptations; (b) information about the real users' preferences useful to calculate the distance between system's recommendations and real users preferences. Six main topic areas were identified in the questionnaire: personal data, information about visual problems, familiarity towards computer and interfaces, food and restaurant preferences, restaurant prices preferences, hotel prices

preferences. The final questionnaire was made up of 14 questions where both the questions and the answers were fixed.

Procedure. The questionnaires were auto-filled by the users to avoid any possible interviewer's interferences and gained a week after the distribution. The questionnaire was anonymous and introduced by a written presentation explaining the general research aims. For the items concerning personal data, visual diseases, computer and interfaces, the participants were required to thick the appropriate answer from a set of gives answers. In the other questions, users had to express their level of agreement with the options concerning the given questions by choosing an item of a 5-point Likert scale. The survey was conducted in September-October 2002 and the participants were Italian citizen living in - or in the suburbs of - the city of Genoa and Turin, in the North of Italy.

We entered the data set (a) in a PC simulator version of the system to generate the system responses. This version contains a service database that is multipurpose relational one and, for the evaluation, it has been populated with tourist information about the Turin area.

Measures. We analyzed the correctness of recommendations by the exploitation of two *statistical accuracy metrics* (Sarwar et al, 1998), MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) that are aimed at evaluating the closeness between the numerical recommendations provided by the system and the numerical ratings entered by the users for the same items. More precisely, MAE (Mean Absolute Error) and RMSE evaluate the distance between the system predictions and the collected users opinions -set (b)- by means of rate vectors (both user's and system's items were expressed on a scale ranging from 0 to 5). Obviously, higher values mean worst recommendations. To test the accuracy of the selection process we have also measured the precision of the collected data (ratio between the user-relevant contents and the contents presented to the user, see 1.2.1).

We used a similar approach also for evaluating the accordance of the proposed layout with contextual conditions and user's cognitive capabilities. On the simulator we set two contexts (different for speed and traffic level). The layout loaded by the system was then compared with that ones chosen by two HCI experts (again the distance is on a scale from 0 to 5).

3.5.4. Results of the evaluation

First of all, we observed that in all cases the system was able to select facilities close to the user's location, giving priority to the closest ones at specific times (e.g., closest hotels at night). We can say that the requirement for MASTROCARONTE of giving a correct weight to external conditions, like distance, time pressure, etc., in the selection

of the items, is satisfied. As regards the content of the recommendation, namely the requirement of correspondence between the user's preferences and the features of the items selected, the obtained results are here summarized:

Restaurants				Hotels	
MAE	RMSE	MAE*	RMSE*	MAE	RMSE
1,44	1,75	1,05	1,49	1,87	2,08
Precision = 0.41				**Precision** = 0.2	

Table 3.5. Experimental results. (*) means without the restaurant prices predictions

As it can be seen, the results are not so satisfactory. Concerning restaurants, we noticed that the most of the distance was due to the price. The recommendations concerning prices were too optimistic: most participants chose lower price levels (notice in Table 3.5 the better results obtained without calculating price prediction). The price problem also explains the higher values concerning Hotels that are recommended mainly following price estimations. As we will see, one of the possible reasons could be due to the current economic situation, which is probably different compared to that one considered by the psycho-graphic study used to build stereotypes and we have to adapt the price recommendations to the current propensity to spend. Thus, 3.5 gives a snapshot of the results and provides the important information regarding prices. However, such a piece of information and others, not represented in the table, come from a complex analysis aimed at explaining the reason of that 1,44 or also 1,05 distance.

Several are the ways that can be followed to investigate the cause, or causes, of the distance to finally finding out where the system has to be modified. In particular we have identified two types of approaches, which could be defined, for simplicity, as backward and forward or, also, as bottom-up and top-down. What do they mean? The first one begins the investigation from the data, looking for cases, or group of cases, with anomalous statistics. The second type of approach starts the investigation with no knowledge about the collected data, makes hypothesis and tests the component of the system that can produce errors of adaptation.

We investigated exclusively the results concerning MAE and RMSE, since the overall contents of the database are not exhaustive and the calculated precision is merely indicative.

3.5.4.1. __The bottom-up approach__

We started with the first type of approach and, in order to perform this analysis, we disaggregated the results by considering different groupings of the users. First of all, we considered three ways to group the subjects involved in the test:

- *Sampling Groups* = classes of subjects with common socio-demographic features (in our case we singled out eight groups labeled as s1, s2, s3, ..., s8)

- *Profile Groups* = classes of subjects having the same predicted profile. In other words, classes of subjects having the same antecedents used in the adaptation rules. The profiles are transversal respect to the stereotypes. A profile includes *i)* similar predictions of stereotypes concerning the propensity to spend, technology expertise, receptivity; *ii)* similar rules which estimates the receptivity of the driver; *iii*) socio-demographic features (age). We singled out fifteen groups labeled as pf1, pf2… pf18)

- *Prediction Groups* = classes of profiles, namely subjects belonging to some profiles, for which the system produced the same recommendations (in our case we singled out five groups labeled as pr1, pr2, pr3, pr4, pr5). These are clearly related to the previous ones, in the sense that each prediction group includes a set of profile groups, in our case, for example we have that pr2 = pf3 ∪ pf6 ∪ pf10

By comparing the behavior of the systems for these different groups, it is possible to get hints to understand for which reasons the system does not provide a good advice on some users. As a first result of this deeper evaluation, we noticed that the recommendations changed significantly according to the different predictions groups of users: some groups received better recommendations than others (see Table 3.6). As noticed above, the five *prediction groups* cluster participants with different socio-demographic features (they crossed the initial sampling groups) but common recommendations. For instance people belonging to group pr2 are 26-35 years old and have e medium propensity to spend, while people belonging to group pr4 are 46-65 years old and have a high propensity to spend. Within group pr2 there are males and females, with different education level and different professions (employees, autonomous workers, managers) and within group pr4 there are males and females, with different education level and different professions (employees, managers, teachers). The results in the table suggest that either the accuracy of classification or the prediction for pr4 is more problematic than for pr2.

groups	Restaurants		Hotels	
	MAE	RMSE	MAE	RMSE
group pr1	1,46	1,78	2,05	2,27
group pr2	1,08	1,40	1,00	1,27
group pr3	1,29	1,53	1,85	2,05
group pr4	1,54	1,85	1,92	2,11
group pr5	1,72	2,05	2,02	2,24

Table 3.6. *Prediction groups*' results

In order to have a better evaluation of these results, we wanted to understand if these differences were uniform among all the *Profile groups* that belong to the same *Predictions group* or not. The idea was that, in case we observed some *Profile groups* with very high MAE with respect to the belonging *Prediction group*, we could hypothesize that the negative result of the *Prediction group* is due to one only *Profile group*. As noticed above, given this acquisition, it becomes possible to limit the probable causes of error: it can be a classification problem or a personalization problem regarding only the specific values of that *Profile*. The results of this evaluation are reported in Table 3.7. As you can see, we have computed the distances MAE and RMSE between each *Profile Group* and the belonging Prediction Group, and then we calculated the Standard Deviation (SD) inside each *Profile Group*. We decided to exclude (also in the following table), from such a computation, *Profile Groups* with less then five subjects. But the result was that many groups could not be classified, making all the remaining statistics not significant.

Predictions groups	Profile groups	Restaurants			Hotels		
		MAE	SD	RMSE	MAE	SD	RMSE
group pr1	Group pf1	1,662	0,408	1,904	2,133	0,149	2,315
group pr1	Group pf2	1,434	0,292	1,749	1,990	0,235	2,238
group pr1	Group pf4	1,397	0,212	1,767	2,067	0,893	2,408
group pr2	Group pf3	0,980	0,164	1,370	0,900	0,150	1,139
group pr2	group pf6	0,829	0,136	1,260	1,029	0,167	1,302
group pr3	group pf7	0,86	0,212	1,17	2,05	0,229	2,19
group pr3	group pf8	0,82	0,10	1,13	1,45	0,21	1,64
group pr4	group pf13	1,14	0,27	1,63	1,94	0,34	2,13
group pr5	group pf17	1,27	0,21	1,74	1,97	0,19	2,22

Table 3.7. Prediction and Profile groups' results

If we consider the Standard Deviation of pf1, we see that it is 0,408 while the Standard Deviation of all the subjects is 0,3. It means that the group is not very

homogeneous or that it is homogeneous with a medium-low deviation. The value of the SD is important because it allows to understand if the deviation of the group is due to a single isolated case or not. Once we have seen that it is not an isolated case to determine the deviation, the problem becomes understanding the cause of the deviation of the Profile group, i.e. if it is for a problem of classification or of personalization. This second evaluation can be done easily, because we know exactly the values of the dimension used by the system to classify (the values that identify the Profile). For example, for the group pf1 we judged that adaptation rules were right defined. So we were reasonably sure that it was a classification problem.

Finally it is even possible to identify the exact component in charge of the bad classification, just calculating the MAE for the different features of the advice. For example, for restaurants, the features prices, kind of food, type of places, etc. are correlated with specific user model dimensions, which can be computed using the stereotypes (e.g., propensity to spend and thus price) or the rules (type of food and of places). For pf1 we discovered a problem in the propensity to spend (stereotypes).

We now move to the last analyses regarding the bottom-up approach. Let us consider the *Sampling groups*. Also in this case we have different results for different groups (see Table 3.8).

groups	restaurants		Hotels	
	MAE	RMSE	MAE	RMSE
group s1	1,39	1,67	1,64	1,75
group s2	1,41	1,69	1,81	2,08
group s3	1,36	1,71	1,00	1,18
group s4	1,59	1,92	2,06	2,28
group s5	1,18	1,55	1,31	1,58
group s6	1,49	1,78	2,05	2,23
group s7	1,22	1,51	1,85	2,06
group s8	1,42	1,73	1,70	1,88
not-classified	1,50	1,81	1,93	2,11

Table 3.8. *Sampling groups*' results

The *Sampling groups* are the initial 8 groups collected by the sampling strategy. It happens that 5 subjects result as not classifiable in any group. Also in this case there are differences between groups: for instance *group s6* receives the best restaurant recommendations while *group s3* receives the worst recommendations.

Finally, we compared these results with those obtained for the prediction groups. First of all an ANOVA comparing the groups' results showed that the different results are due to a significant correlation between the kind of group taken into account (independent variable) and its related recommendations (dependent variable).

Predictions groups (103 subjects for 5 groups): restaurants MAE: $F(4, 98) = 9,27$, $p<0,01$; restaurants MAE no prices: $F(4, 98) = 5,33$, $p<0,01$; hotels MAE, $F(4, 98) = 26,83$, $p<0,01$.

Sampling groups (107 subjects for 9 groups): restaurants MAE: $F(8, 98) = 2,74$, $p<0,01$; restaurants MAE no prices: $F(8, 98) = 1,71$ $p<0,01$; hotels MAE, $F(8, 98) = 9,40$, $p<0,01$.

All these results (expect for restaurants MAE no prices for Sampling groups) show significant dependencies. In summary, by analyzing the groups precisely, we can thus get clues on the parts of the Knowledge Base that have to be revised. A part from the specific changes to the Knowledge Base that were suggested by the tests (which are not relevant for a reader), it is worth noting that this methodology of aggregation of the evaluation was very interesting, providing interesting insights on the system's knowledge bases and behavior.

3.5.4.2. <u>**The top-down approach**</u>

Given the problems with not comparable *Profile groups*, and anyway to extend our analysis, we also followed this approach and obtained meaningful results.

Test of the correctness of the KB. Our starting hypothesis was that each stereotype shares homogeneous preferences, behaviors and lifestyles. The inexistence of this supposed correlation could be due to these factors: i) the stereotypes are too generic and therefore they cannot be used for a specific domain; ii) the entered data are too old and do not reflect the current situation.

We could accomplish this test using the questionnaire and comparing the answers with the Eurisko classes. The answers in the questionnaire showed that users almost always selected ranges of prices lower than those predicted by the Life Styles. This finding was also confirmed by the computation of the MAE and RMSE (see Table 3.8). Therefore the need of updating our Knowledge Base, concerning the supposed propensity to spend emerged clearly. The other dimensions (technology expertise, receptivity) derived from the stereotypical knowledge seemed to be well suited for the system's purposes, and therefore non-generic.

Test of possible errors in the KB implementation. The research we exploited describes the lifestyles in a qualitative way and thus we had to translate them in probabilistic values. Errors are frequent in processes like this, due to a misunderstanding in tuning the estimates. For instance we found that working young

people are described as having a high propensity to spend. However, the test demonstrated that their propensity to spend is related to their current situations, so they generally prefer non-expensive restaurants even if they like to go out for dinner. As a consequence, we concluded that there is the need of splitting the propensity to spend in frequency and value, but Eurisko does not have this distinction and we have no means to know its interpretation of the dimension. A confirmation of this problem comes from the MAE calculated with respect to students. A possibility could be to define a new Knowledge Base, exploiting a domain specific survey to the target population.

Test of the Rules in charge of the personalization. In the system, a set of rules associates user features (age, propensity to spend) to hotels/restaurants features (price, kind of restaurant, kind of food, etc..). The test showed some problems, for instance, that restaurant suggestions for 25-36 years old are better than restaurants suggestions for 20-25 years old. Then, within the first group the suggestions are better suited for people characterized by a medium propensity to spend. From this analysis, now we know the associations that have to be revised.

Finally, let us move to the evaluation of the personalization choices as regards the format and layout of the presentation. As regards layouts, the evaluation was made as follows. We interviewed two HCI experts, asking them to suggest, for each one of the profiles the best interface choice. We then compared these "gold standards" with the personalized layout selected by the system. We used the following procedure: the two experts proposed the better interface for groups of subjects sharing the same features used by the system to generate layout recommendations (age, technology expertise, receptivity) in a given contexts. Then we calculated the distance (MAE and RMSE) between the real system's proposals and experts' suggestions. Here the results:

Layout Context 1		Layout Context 2	
MAE	RMSE	MAE	RMSE
0,18	0,18	0,09	0,09

Table 3.9. Layout prediction' results

The closeness between the system's proposed layout and the HCI experts' suggestions confirmed the appropriateness of layout adaptations choices.

3.5.5. Conclusions

In their review of personalized hypermedia presentation techniques Kobsa et al. [2001] divide the personalization process into these three major tasks: *i)* acquisition method

and primary inferences, *ii)* representation and secondary inferences, *iii)* adaptation production. In this evaluation exercise we take into account this process by performing a layered evaluation where the evaluation is decomposed into different layers corresponding to the high layers described above. We performed a formative evaluation for the points *i* and *ii* and a both a formative and predictive evaluation for point *iii* (which includes both content and interface adaptations). The final results suggest and give clues for a revision of the current knowledge base and adaptation rules concerning the final contents recommendations, while the interface adaptations obtained good evaluation results. The next step in this user-centered iterative evaluation process will be a testing with subjects interacting with a version of MASTROCARONTE currently running on a Fiat Punto to have both users rating the predictions generated by the system and interacting with the on-board car system.

3.5.6. Post test considerations

At the beginning of the paper, we stated that the goal of this work was to present a first evaluation of MASTROCARONTE, showing the advantages and the potentials of adaptation and personalization techniques in the provision of on-board applications. To achieve that, we have reviewed the main principles and works in the field of User Modeling and Adaptive Hypermedia, from the Web applications to mobile and on-board ones. Then we have enlighten the main points of interest of the field, showing constraints and new opportunities, noticing also the contrast between the large interest for the field and the lack of scientific literature about that. Afterwards, we have presented our proposal of architecture for on-board adaptive services and finally we have illustrated methodology, results and future steps of a preliminary evaluation of the different aspects of a tourist prototype of MASTROCARONTE.

The results of the evaluation exercise demonstrate that these claims can be accomplished in practice. The best way to draw a conclusion is to consider some of the guidelines for the design of on-board HMI, noticing that most of them are accomplished by adaptation and personalization:

Paying attention to the risk of interference with the main task of driving by causing a dual-tasks [Green, 2000].

Choosing the modality of interaction. The kinds of interaction in HMI can vary from hand-held remote control to voice control, from controls and buttons near to the display or touch-screen interfaces to advanced HMI exploiting the potential of multimodal interaction such as haptic (tactile and kinaesthetic) and acoustic interaction [Mariani, 2002]. Multimodal interaction in particular has been found as positively affecting driver's safety, resulting in faster reaction times and fewer errors for emergency response displays when compared with simple visual or auditory

display [Liu et al, 1999]. Indeed, the exploitation of non-visual interactions can lighten the driver's overload and can be effective for attracting driver' s attention (e.g. auditory interaction for alerts, warning). However, these kinds of interaction can be chosen only in particular situations (e.g., auditory modality is effective if output is simple) and for located actions [Summerskill et al., 2002].

Readability of the content to be displayed must be taken into account. Besides the usability of controls and display, the items presented on the screen have to be easily readable and immediate. For instance, a study [Burnett and Porter, 2002] showed that the exploitation of landmarks vs. distance cues in vehicle navigation systems decreases the number of glances and workload was perceived to be lower.

Tuning the interaction. At the beginning the interaction should be easier since the unexpected system responses are handled in a more problematic way and the user workload could increase with undesired effects. Then, more experienced the user becomes less decision she has to take about the next actions to perform since she developed an interaction strategy [Jahn, et al. 2002]. In MASTROCARONTE the layout is selected on the basis of the user's receptivity and familiarity with technological devices. But, as a set of rules periodically revises and updates the User Model, after a number of interactions (which change according with the level of receptivity itself) the layout selected becomes richer in amount of information and services.

The interface must be adequate to each specific user. Even if interfaces that make the choice of responses as obvious as possible are beneficial [Norman, 1998], in this particular context also individual differences concerning receptivity and cognitive load should be taken into account. For instance, the maximum quantity of information to be presented to old and young drivers should be differing. Empirical studies [Labiale and Galliano, 2002] showed that by increasing the number of pictograms displayed on an in-car system the number and the duration of visual fixation increase for both old and young people. However, older drivers require longer visual fixation than young drivers according to the explanatory hypothesis of a perceptual and cognitive slowing down of elderly subjects. Labiale and Galliano proposed no more than 9 pictograms for young drivers and no more than 6 for older drivers. However, the exploitation of Intelligent User Interface taking into account other factors can modify the complexity of each in car display. For instance, people having more familiarity towards technology and computers seem to be faster at learning new interfaces. Visual diseases can affect the amount of information to be read particularly in a visual-centered task such as driving. The speed of the car, the type of the roads, the traffic and the driving experience may change the way in which drivers manage the switching between primary and secondary task.

The last dimension (driving experience) has not been considered in MASTROCARONTE (it could be added in the next release) but all the other ones are exactly the features taken in consideration by the system for loading the right stylesheet.

Conclusions

To conclude, I advocate, of course, the importance of a correct evaluation methodology during the development and the testing of a user-adapted system. Significant testing results can lead to more suitable and successful systems. According to my point of view, both in case of qualitative and quantitative methods of research can gain fruitful contributions.

First of all, I want reaffirm the importance of a "correctly-carried-out" evaluation of user-adapted systems, as sustained in Chapter 1. In fact, the problem of most evaluation is the non-significance of the results and therefore the absence of generalisations. Moreover, I want to underline again the importance of considering the other evaluation techniques derived from HCI during the development of adaptive systems, as depicted in Table 3.1, in addiction to empirical evaluation. Moreover, I advocate the involvement of the users in every design phases and the importance of an embodied perspective in the development of user-adapted systems.

Concerning the choice between quantitative and qualitative methodologies, I advocate the importance of both techniques in evaluation of adaptive systems. This implies a different point of view in the evaluation: while the quantitative research tries to explain the variance of the dependent variable generated through the manipulation of independent variables (*variable-based*), in qualitative research the object of study becomes the individual subjects (*case-based*). Qualitative researchers sustain that a subject cannot be reduced to a sum of variables and therefore a deeper knowledge of a fewer group of subjects is more fruitful than an empirical experiment with a representative sample.

The goals of the analyses are also different: while quantitative researchers try to explain the cause-effect relationships, qualitative researchers want to comprehend the subjects under study by interpreting their point of views. While quantitative research tries to explain why subjects behave in a particular way, qualitative research tries to explain how subjects behave in that particular way.

Concerning the methodologies of analysis, quantitative researchers try to validate a theory by falsification, while qualitative researchers try to individuate the so-called *ideal types* through the description and the classification of the collected empirical data and the individuation of typology dimensions. The *ideal types* are conceptual categories useful to interpret the reality under observation. In case of user modeling systems, these categories can be used or can offer suggestions to model the features of the users and then adapt the system to these features.

The choice between quantitative and qualitative methodologies is not trivial and depends from the aims and the purpose of the evaluation.

For instance, if we want to test the impact of the exploitation of different adaptation techniques in a given interface, a controlled experiment can gain useful results. If we want to discover the aspects of an interaction that we have to take into account to model this interaction, observing users in context and interview them can offer material useful to build the user model categories.

In general, if we want to discover new categories useful to model the interaction, a qualitative approach can gain more fruitful results, while if we want to investigate the impact of already known variables a quantitative approach can be preferred. However, this is not a rule and both methods can take interesting results.

From a methodological point of view, at the moment my interest is directed towards qualitative evaluations, in particular ethnographic studies, since I have always carried out quantitative evaluations. In my future activity I hope to carry out qualitative analyses in order to put in practice the ideas I have developed during writing of the thesis. However, I will go on also with quantitative research and I will try to follow the rules I have here described in order to contribute, by means of both the methodologies, to the construction of a corpus of guidelines from which other researchers can obtain insightful information as Weibelzahl and Lauer sustain [1.5].

Bibliography – References

[Ahlberg and Wistrand, 1995] C. Ahlberg and E. Wistrand, *IVEE: An Information Visualization and Exploration Environment,* InfoVis'95, New York, NY, 1995, pp. 66-73.

[André and Rist, 2000] E. André, T. Rist, *Presenting through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems,* Proceedings of the International Conference on Intelligent User Interfaces, IUI ' 2000, New Orleans, 2000

[Ardissono et al., 1999] L. Ardissono, A. Goy: *Tailoring the interaction with users of electronic shops,* in: J. Proceedings of 7th International Conference on User Modeling, 1999. pp. 35-44.

[Ardissono and Goy, 2000] L. Ardissono and A. Goy, *Tailoring the Interaction with Users in Web Stores,* User Modeling and User-Adapted Interaction 10(4), 2000, 251-303.

[Ardissono et al., 2001] L. Ardissono, F. Portis, P. Torasso, F. Bellifemine, A. Chiarotto, A. Difino, *Architecture of a system for the generation of personalized Electronic Program Guides,* Proc. UM' 2001 Workshop on Personalization in Future TV, Sonthofen, Germany, 2000.

[Ardissono and Faihe, 2001] L. Ardissono and Y. Faihe, *UM' 2001 Workshop on Personalization in Future TV,* Sonthofen, Germany, 2001, http://www.di.unito.it/~liliana/UM01/TV.html.

[Ardissono and Buczak, 2002] L. Ardissono and A. Buczak, *TV' 02: the 2nd Workshop on Personalization in Future TV,* Malaga, Spain, 2002, http://www.di.unito.it/~liliana/TV02/.

[Auditel, 2000] Auditel, 2000, http://www.auditel.it

[Bakeman, R. & Gottman, 1986] R. Bakeman and J. M. Gottman, *Observing behavior: An introduction to sequential analysis* Cambridge: Cambridge University, 1986.

[Barbieri et al., 2001] M. Barbieri, M.Ceccarelli, G. Mekenkamp, J. Nesvadba, *A personal TV receiver with storage and retrieval capabilities,* Proc. UM' 2001 Workshop on Personalization in Future TV, Sonthofen, Germany, 2001.

[Barker et al., 2002] T. Barker, S. Jones, C. Britton, D. Messer, *The Use of a Co-operative Student Model of Learner Characteristics to Configure a Multimedia Application,* User Modeling and User-Adapted Interaction 12(2), 2002, 207-241.

[Baudisch and Brueckner, 2002] P. Baudisch and L. Brueckner, *TV Scout: Guiding Users from Printed TV Program Guides to Personalized TV Recommendation,* Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[Baus et al., 2002] J. Baus, A. Krüger, W Wahlster, *A Resource-Adaptive Mobile navigation System,* Proc. of the 2002 International Conference on Intelligent User Interfaces, ACM Press, 2002, pp. 15-22.

Bibliography

[Bellifemine et al, 1999] F. Bellifemine, et al.: 1999, *Deliverable A12D1, Agent-system software for the AVEB Phase 1 Demonstrator*, FACTS (FIPA Agent Communication Technologies and Services) is the ACTS project number AC317 of the European Commission, 1999.

[Bellotti and MacLean] V. Bellotti and A. MacLean, *Design Space Analysis (DSA)*, viewed at: http://www.mrc-cbu.cam.ac.uk/amodeus/summaries/DSAsummary.html

[Benyon, 1993] D. Benyon, *Adaptive Systems: A Solution to Usability Problems*. User Modeling and User-Adapted Interaction (3), 1993, 65-87.

[Beyer and Holtzblatt, 1998] H. Beyer and K. Holtzblatt, *Contextual Design: Defining Customer-Centered Systems*, Morgan Kaufmann Publishers, Inc., San Francisco CA , 1998.

[Billsus and Pazzani, 1999] D. Billsus and M. Pazzani, *A Personal News Agent that Talks, Learns and Explains*, Proc. 3rd Int. Conf. on Autonomous Agents (Agents '99), Seattle, WA, 1999, 268-275.

[Billsus and Pazzani., 2000] D. Billsus and M. Pazzani, *User Modeling for Adaptive News Access*, User Modeling and User-Adapted Interaction. 10:2/3, 2000, pp. 147-180.

[Blumer, 1969] H. Blumer, *Symbolic interactionism: perspective and methods,* Prentice-Hall (Englewood Cliffs, N J), 1969.

[Brusilowsky, 1996] P. Brusilowsky, *Methods and Techniques of adaptive hypermedia*, User Modelling and User Adapted Interaction 6 (2-3): 1996, pp 87-129.

[Brusilovsky and Eklund, 1998] P. Brusilovsky and J. Eklund, *A Study of User Model Based Link Annotation in Educational Hypermedia*, Journal of Universal Computer Science, 4, 1998, 429-448.

[Brusilovsky et al., 1998] P. Brusilovsky, A. Kobsa, J. Vassileva editors, *Adaptive Hypertext and Hypermedia*, Kluwer, 1998.

[Brusilovsky, 1999] P. Brusilovsky, *Methods and Techniques of Adaptive Hypermedia.*, User Modeling and user Adapted Interaction, 6 (2-3): 87-129, 1996.

[Brusilovsky et al., 2001] P. Brusilovsky, C. Karagiannidis, and D. Sampson, *The benefits of layered evaluation of adaptive applications and services*, S. Weibelzahl, D. N. Chin, & G. [Weber, 2001] Weber (Eds.), Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001, 2001, 1-8.

[Brusilowsky, 2001] P. Brusilowsky, *Adaptive Hypermedia*, User Modelling and User Adapted Interaction 11 (1-2): 2001, pp 87-110.

Bibliography

[Brusilovsky and Maybury, 2002] P. Brusilovsky and M. Maybury, *Communications of the ACM: Special Issue "The Adaptive Web"*, 45, 2002.

[Buczak, 2002] A. Buczak, J. Zimmerman and K. Kurapati, *Personalization: Improving Ease-of-Use, Trust and Accuracy of a TV show Recommender*, Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[Burattini and Cordeschi, 2001] E. Burattini e R. Cordeschi (a cura di), *Intelligenza Artificiale: la storia e le idee, in Intelligenza Artificiale. Manuale per le discipline della comunicazione*, Carocci, Roma, 2001.

[Burnett and Porter, 2002] G. E. Burnett and J. M. Porter, *An empirical comparison of the use of distance versus landmark information within the Human-Machine Interface for vehicle navigation system*, Human Factors In Transportations, Communication, Health and the Workplace, Shaker Publishing, 2002, pp. 49-64.

[Calvi, 1986] G. Calvi, *Indagine sociale italiana*, Rapporto Eurisko, FrancoAngeli, 1986.

[Campbell et al., 1998] J. L. Campbell, C. Carney, T. H. Kantowitz, *Human Factors Design Guidelines for Advanced Traveler Information Syetems (ATIS) and Commercial Vehicle Operations (CVO)*, Office of Safety and Traffic Operations R&D Federal Highway Administration FHWA-RD-98-057, 1998.

[Card et al., 1983] S. K. Card, T. P. Moran, T and A. Newell, *The Psychology of Human-Computer Interaction,* Lawrence Erlbaum Associates Inc., Hillsdale, New Jersey, USA, 1983.

[Card et al., 1994] S. K. Card, J. D. MacKinlay, B. Shneiderman, *Readings in Information Visualization*, Morgan Kaufmann, 1999.

[Carroll and Rosson, 1987] J. Carroll and M. B. Rosson, *The Paradox of the Active User*, J.M. (Ed.), Interfacing Thought: Cognitive Aspects of Human-Computer Interaction, Cambridge, MA: MIT Press, 1987.

[Casetti and Di Chio, 1998] Casetti F., Di Chio F., *Analisi della televisione*, Strumenti Bompiani, 1998.

[Cassel, 2000] J. Cassel, *Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents*, Cassell, J. et al. (eds.), Embodied Conversational Agents, pp. 1-27. Cambridge, MA: MIT Press, 2000.

[Chen, 1999] C. Chen, *Information Visualization and Environments*, Springer Verlag, 1999.

[Cheverest et al., 2000] K. Cheverest, N. Davies, K. Mitchell, P. Smyth, *Providing tailored context-aware information to city visitors,* Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2000), Lecture Notes in Computer Science 1892, Springer Verlag 2000, pp. 73-85..

Bibliography

[Chin, 2001] D. Chin, *Empirical Evaluation of User Models and User-Adapted Systems*, User Modeling and User-Adapted Interaction Vol. 11, No. 1/2, Kluwer Academic Publishers, 2001, 181-194.

[Chin and Crosby, 2002] D. Chin and M. Crosby editors, *Empirical Evaluation of User Models and user Modeling Systems*, User Modeling and User-Adapted Interaction, vol. 12 (2-3), 2002.

[Cohen, 1960] J. Cohen, *A coefficient agreement for nominal scales,* Educational and Psychological Meas. 20,1960, 37-46.

[Conklin, 1987] J. Conklin, *Hypertext: An Introduction and Survey*, IEEE Computer, 20, 1987, 17-41.

[Console et al., 2001] L. Console, I. Lombardi, R. Montanari, M. Guagliumi, M. Salvoni, L. Tonelli, I. Torre, *MASTROcarONTE a Multiagent Adaptive System for Tourist Recommendations Onboard the car, which Observes the Needs and Tailors the Helps*, In Proc. IJCAI Workshop on Airtificial Intelligence in Mobile Systems, AAAI Press, Seattle, WA, Agosto 2001, pp. 19-25.AIMS 2001.

[Console et al., 2002] L. Console, S. Gioria, I. Lombardi, V. Surano, I. Torre, *Adaptation and personalization on board cars, a framework and its application to tourist services*, Adaptive Hypermedia and Adaptive Web-Based Systems 2002, Lecture Notes in Computer Science, Springer Verlag, 2002, 112-121.

[Console et al., 2003] L. Console, C. Gena and I. Torre, *Personalization and adaptation for on-board information system: a prototype and its evaluationm,* Workshop on Automotive User Interface, HAAMAHA 2003.

[Corbetta, 1999] P. Corbetta, *Metodologie e tecniche della ricerca sociale*, Il Mulino, Bologna, 1999.

[Cotter and Smyth, 2000] P. Cotter and B. Smyth, *A Personalized Television Listing Service*, Communications of the ACM 43 (8), 2000.

[Debevc, 1993] M. Debevc, *Adaptive bar*, INTERACT '93 and CHI '93 conference companion on Human factors in computing systems, 1993.

[Debevc et al., 1997] M. Debevc, B. Meyer, and R. Svecko, *An adaptive short list for documents on the World Wide Web*, Proceedings of the 1997 international conference on Intelligent user interfaces, 1997.

[Diaper, 1989] D. Diaper (Ed.), *Task analysis for human-computer interaction*, Chicester, U.K.: Ellis Horwood, 1989.

Bibliography

[Difino et al., 2002] A. Difino, B. Negro, A. Chiarotto, *A Multi-Agent System for a Personalized Electronic Program Guide*, Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[Dilthey, 1954] W. Dilthey, Critica della ragione storica, Torino, Einaudi, 1954.

[Dix et al., 1998] A. Dix, J. Finlay, G. Abowd and R. Beale, *Human Computer Interaction*, Second Edition, Prentice Hall, 1998.

[Dourish and Button, 1998] P. Dourish and G. Button, *On "Technomethodology": Foundational Relationships between Ethnomethodology and System Design*, Human-Computer Interaction, 13(4), 1998, 395-432.

[Dourish, 2001a] P. Dourish, *Where The Action Is: The Foundations of Embodied Interaction,*. MIT Press, 2001.

[Dourish, 2001b] P. Dourish, *Seeking a Foundation for Context-Aware Computing*, Human-Computer Interaction, 16(2-3), 2001.

[Dumas and Redish, 1999] J. S. Dumas and J. C. Redish, *A Practical Guide To Usability Testing*, Norwood, N.J. Ablex Publishing Corp, 1999.

[Eklund and Brusilovsky, 1998] J. Eklund and P. Brusilovsky, *The value of adaptivity in hypermedia learning environments:A short review of empirical evidence*, P. Brusilovsky & P. De Bra (Eds.), Proceedings of Second Adaptive Hypertext and Hypermedia Workshop (at the Ninth ACM International Hypertext Conference - Hypertext'98), 1998, 13-19.

[Espinoza and Höök, 1996] F. Espinoza and K. Höök, *A WWW Interface to an Adaptive Hypermedia System*, Conference on Practical Application of Agent Methodology (PAAM'96), London, 1996.

[Eurisko, 2000] Eurisko, Sinottica, 2000, http://www.eurisko.it .

[Fink et al., 1998] J. Fink, A. Kobsa, A. Nill, *Adaptable and adaptive information for all users, including disabled and elderly people,* New review of Hypermedia and Multimedia, Vol. 4, 1998, pp. 163-188.

[Fucs, 2001] R. Fucs, *Personality Traits and their Impact on Graphical User Interface Design: Lesson Learned from the Design of a real Estate Website*, Proc PC-HCI 2001, Patras, Greece, 2001.

[Garfinkel, 1967] H. Garfinkel, *Studies in Ethnomethodology,* Poli1y Press, 1967.

[Geertz, 1973] C. Geertz, *The interpretation of cultures,* Basic Books, New York, 1973.

[Gena, 2001]

Bibliography

Gena, C. (2001). Designing TV Viewer Stereotypes for an Electronic Program Guide. In: Bauer, M., Gmytrasiewicz, P.J., Vassileva, J. (eds) User Modeling 2001. UM 2001. Lecture Notes in Computer Science(), vol 2109. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44566-8_43

[Gena and Ardissono, 2001]
Gena, C., & Ardissono, L. (2001, July). On the construction of TV viewer stereotypes starting from lifestyles surveys. In *Workshop on Personalization in Future TV*.

[Gena et al, 2001a] C. Gena, A. Perna, F. Cena, R. Morisano, *A personalized hypermedia application for web-based tools*, Poster and Demonstrations Proceedings of the 12th Acm Conference on Hypertext and Hypermedia, ACM Press, Arhus, Denmark, 2000.

[Gena et al., 2001b] C. Gena, A. Perna, M. Ravazzi, *E-tool: a personalized prototype for web based applications*, Europe Chapter of the Human Factors and Ergonomics Society Annual Conference 2001, Fiat Research Center, Torino 7-9 November , 2001.

[Gena, 2002]
Cristina Gena. 2002. An empirical evaluation of an adaptive web site. In Proceedings of the 7th international conference on Intelligent user interfaces (IUI '02). Association for Computing Machinery, New York, NY, USA, 192–193. https://doi.org/10.1145/502716.502752
[Goffman, 1959] E. Goffman, *The presentation of self in everyday life,* Doubleday, Garden City, N.Y., 1959.

[Good et al., 1999] N. Good, J. B. Schafer, J. A. Konstan, A. Botchers, B. M. Sarwar, J. L. Herlocker, and J. Ricdl, *Combining collaborative filtering with personal agents br better recommendations,* Proceedings of the Sixteenth National Conference on Artificial Intelligence, 1999, 439-446.

[Goren-Bar and Glinansky, 2002] D. Goren-Bar and D. Glinansky, *Family Stereotyping - A Model to Filter TV Programs for Multiple Viewers*, Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[Green, 2000] P. Green, *Crashes induced by driver information systems and what can be done to reduce them* (SAE paper 2000- 2001-C008), Convergence 2000 Conference Proceedings, SAE Publication, 2000, pp. 26-36.

[Greenbaum and Kyng] J. Greenbaum and M. Kyng (Eds.), *Design at Work: Cooperative Design of Computer Systems*, Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1991.

[Greening, 2000] D. R. Greening, *LikeMinds: building consumer trust with accurate product recommendations*, 2000, http:www.likeminds.com/technology/white_papers/websell/

[GroupLens, 2002] GroupLens, 2002, http://www.cs.umn.edu/Research/GroupLens.

Bibliography

[Hanani et al., 2001] U. Hanani, B. Shapira and P. Shoval, 2001, *Information Filtering: Overview of issues, research and systems*, User Modeling and User-Adapted Interaction 11, 2001, 203-259.

[Hankey et al., 2000] J. M. Hankey, T. A. Dingus, R. J. Hanowsky, W. W. Wierwille & C. Andrews, *In-vehicle information systems behavioral model and design support*, Final report., (FHWA-RD- 2000-135), McLean, VA: Federal Highway Administration, U.S. Department of Transportation, 2000.

[Healey et al., 2002] J. Healey, R. Hosn and S. H. Maes, Adaptive content for device independent multi-modal borwser application, *Adaptive Hypermedia and Adaptive Web-Based Systems 2002*, Lecture Notes in Computer Science, Springer Verlag, 2002, pp. 401-405.

[Heidegger, 1927] M. Heidegger, *Being and time,* english translation 1962, Harper & Row, New York, 1927.

[Herder and van Dijk, 2002] E. Herder and B. van Dijk, *Personalized adaptation to device characteristic*, Adaptive Hypermedia and Adaptive Web-Based Systems 2002, Lecture Notes in Computer Science, Springer Verlag, 2002, pp. 598-602.

[Höök, 1997] K. Höök, *Evaluating the Utility and Usability of an Adaptive Hypermedia System*, Proceedings of 1997 International Conference on Intelligent User Interfaces, ACM, Orlando, Florida, 1997.

[Höök, 2000] K. Höök, *Steps to take before IUIs become real*, Journal of Interacting with Computers, vol. 12, no. 4, 2000, pp. 409 426.

[Hughes et al., 1995] J. Hughes, J. O'Brien, T. Rodden, M. Rouncefield and I. Sommerville, *Presenting Ethnography in the Requirements Process*, Proc. RE'95, York, 1995 27-35.

[Hull, 1998] D. A. Hull, *The TREC-7 Filtering Track: Description and Analysis*, E. M. International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, 246–254.

[Ishizuka et al, 2001] M. Ishizuka, S. Descamps, H. Prendinger, *A Multimodal Presentation Mark-up Language for Enhanced Affective Presentation*, Advances in Education Tecnologies: Multimedia, WWW and Distant Education, Proc. Int'l Conf. on Intelligent Multimedia and Distant Learning (ICIMADE- 2001), Fargo, North Dakota, 2001, 9-16.

[Jahn, et al. 2002] G. Jahn, J. F. Krems, C. Gelau, *Skill-develoment when interacting with in-vehicle information systems: a training study on the learnability of different MMI concepts*, Human Factors In Transportations, Communication, Health and the Workplace, Shaker Publishing, 2002, pp. 35-47.

[Jameson, 1999] A. Jameson, *User-Adaptive Systems, An Integrative Overview*, Presented at UM99, the Seventh International Conference on User Modeling, Banff, June 1999;

and at IJCAI99, the Sixteenth International Joint Conference on Artificial Intelligence, August 1999.

[Kay, 1990] A. Kay, *User interfaces: a personal view*, B. Laurel, The Art Of Human-Computer Interface Design, Addison-Wesley, New York, 1990.

[Keppel, G., 1991] G. Keppel, *Design and Analysis: A Researcher's Handbook*, Englewood Cliffs, NJ: Prentice-Hall, 1991.

[Keppel et al., 1998] G. Keppel, W. H. Saufley and H. Tokunaga, *Introduction to Design and Analysis*, A Student's Handbook. Second Edition, 1998.

[Kirby, 1991] M.A.R. Kirby*, CUSTOM Manual Dpo/std/1.0.*, Huddersfield: HCI Research Centre, University of Huddersfield. 1991.

[Kobsa et al., 2001] A. Kobsa, J. Koenemann, W. Pohl, *Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships*, The Knowledge Engineering Review 16(2), 2001, 111-155.

[Kobsa, 2001] A. Kobsa, "An Empirical Comparison of Three Commercial Information Visualization Systems", IEEE Symposium on Information Visualization, San Diego, CA, 2001, pp. 123-130.

[Kobsa, 2002] A. Kobsa, ICS 105: Project in Human-Computer Interaction and Interfaces, notes for the course, 2002, http://www.ics.uci.edu/~kobsa/.

[Kurapati and Gutta, 2002] K. Kurapati and S. Gutta, *TV Personalization through Stereotypes*, Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[Labiale and Galliano, 2002] G. Labiale and M. Galliano, *Effects of complexity of in-car information on visual fixations and driving performance,* Human Factors In Transportations, Communication, Health and the Workplace, Shaker Publishing, 2002, pp. 147-154.

[Lee and Lai, 1991] J. Lee and K.-Y. Lai, *What's in Design Rationale?,* Human-Computer Interaction special issue on design rationale 6(3-4), 1991, pp. 251-280 .

[Lester et al, 2000] J. C. Lester, S. Towns, C. B. Callaway, J. L. Voerman and P. J. FitzGerald, *Deictic and Emotive Communication in Animated Pedagogical Agents*, Embodied Conversational Agents, Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. (eds.), MIT Press, Boston, 2000..

[Lewis and al., 1990] C. Lewis, P. G. Polson, C. Wharton and C. Rieman, *Testing a Walkthrough Methodology for Theory-Based Design of Walk-Up-and-Use Interfaces*, J. C. Chew & J. Whiteside (Eds.), Proceedings of CHI'90: New York: ACM, 1990.

Bibliography

[Lewis, 1995] D. Lewis, *Evaluating and optimizing autonomous text classification systems*, Proceedings of the 18th Annual International ACM Special Interest Group on Information Retrieval, ACM Press, New York, 1995, 264-254.

[Lieberman, 1997] H. Lieberman, *Autonomous Interface Agents*, ACM Conference on Human-Computer Interface, CHI'97, Atlanta, March, 1997.

[Macaulay et al., 1990] L. A. Macaulay, M.A.R. Kirby, C.J.H. Fowler and A.F.T. Hutt, *USTM: a new approach to Requirements Specification*. Interacting with Computers, vol. 2 no. 1. 1990.

[Maes, 1994] P. Maes, *Agents that Reduce Work and Information Overload*, Communications of the ACM, Vol. 37, No.7, 146, ACM Press, July, 1994, 31-40.

[Malinowsky, 1922], B. Malinowsky, *Argonauts of The Western Pacific*, London, Routledge, 1922.

[Manber et al., 2000] U. Manber, A. Patel, and J. Robison, *Experience with personalization of Yahoo!,* Communication of ACM 43, 8, 2000, 35 - 39.

[Mariani, 2002] M. Mariani, *COMUNICAR : Designing multimodal interaction for advanced in-vechicle interfaces*, Human Factors In Transportations, Communication, Health and the Workplace, Shaker Publishing, 2002, pp. 113-120.

[Mark et al., 2002] G. Mark, A. Kobsa, V. Gonzalez, *Do Four Eyes See Better than Two? Collaborative versus Individual Discovery in Data Visualization Systems*, Proceedings of the Fifth International Conference on Information Visualisation (IV02), London, U.K. Los Alamitos, CA, IEEE Press, 2002, 249-255

[Mc Donald et al., 2001] K. Mc Donald, B. Smyth, A. F. Smeaton,, P. Browne, P. Cotter.: 2001, *Use of the Fischlar video library system*, Proc. UM' 2001 Workshop on Personalization in Future TV, Sonthofen, Germany.

[McTear, 1993] M.F. McTear, *User modeling for adaptive computer systems: a survey of recent developments*, in Artificial Intelligence Review, Vol. 7, 1993, pp. 157-184.

[Microsoft, 1999] Microsoft Corporation, *Microsoft Agent Software Development Kit*, Microsoft Press, Redmond Washington, 1999.

[Miller, 1956] Miller, G.A., *The magical number seven plus or minus two*, Psicol. Rev. 63, 1956: 81-87.

[Milosavljevic et al., 1998] M. Milosavljevic, J. Oberlander, *Dynamic hypertext catalogues: Helping users to help themselves*, in Proc. 9th ACM Conf. on Hypertext and Hypermedia, 1998.

[Mladenic, 1999] D. Mladenic, *Machine learning for better Web browsing*, Proceedings of the Eight Electrotechnical and Computer Sc. Conference ERK'99, Ljubljana, Slovenia: IEEE section, 1999.

[Mobasher et al., 2000] R. Mobasher, R. Cooley and J. Srivastava, *Automatic personalization based on Web usage mining*, Communications of the ACM, (43) 8, 2000.

[Mohageg and Wagner, 2000] M. Mohageg and A. Wagner, *Design Considerations for Information Appliances*, Eric Bergman (eds.), Information Appliances and Beyond, Morgan Kaufmann Publishers, 2000.

[Monk et al., 1993] A. Monk, P. Wright, J. Haber and L. Davenport, *Improving Your Human Computer Interface: A Practical Approach*, BCS Practitioner Series, Prentice-Hall International, Hemel Hempstead, 1993

[MovieLens, 2002] *MovieLens*, 2002, http://www.movielens.umn.edu/

[MPEG-7, 2001] MPEG-7, ISO/IEC 15938-5, *Multimedia Content Description Interface – Part 5,* Multimedia Description Schemes, Version 1, 2001.

[[Netica, 2001] NETICA, *Application for Belief Networks and Influence Diagrams*, User's Guide, 2001, http://www.norsys.com.

[NetPerceptions, 2002] NetPerceptions, Inc, Net Perceptions, 2002, http://www.netperceptions.com

[Nielsen, 1990] J. Nielsen, *Paper versus computer implementations as mockup scenarios for heuristic evaluation*, Proceedings of INTERACT '90, 1990, 315-320.

[Nielsen and Molich, 1990] J. Nielsen and R. Molich, *Heuristic evaluation of user interfaces*, CHI '90, Seattle, Washington, Apr. 1-5, 1990, pp. 249-256.

[Nielsen, 1993] J. Nielsen, *Usability Engineering*, Boston, MA.Academic Press, 1993.

[Nielsen and Mack, 1994] J. Nielsen and R. L. Mack, *Usability Inspection Methods*, New York: John Wiley & Sons, 1994.

[Nielsen, 1997a] J. Nielsen, *WebTV Usability Review*, 1997, http://www.useit.com/alertbox/9702a.html .

[Nielsen, 1997b] J. Nielsen, *TV Meets the Web*, 1997, http://www.useit.com/alertbox/9702b.html.

[Nielsen, 2000] J. Nielsen, *Web Usability*, Milano, Apogeo, 2000.

Bibliography

[Norman and Draper, 1986] D.A. Norman and S.W. Draper, *User centered system design: new perspective on HCI*, Hillsdale NJ, Lawrence Erlbaum, 1986.

[Norman, 1998] D. Norman, *The invisible computer*, MIT Press, Cambridge, 1998.

[Norman, 2001] D. Norman, Advanced TV Standard, *Into the Future with Jaunty Air and an Anchor Around our Necks,* 2001, http://www.jnd.org/dn.mss/tv.html

[O'Connor et al., 2001] M. O'Connor, D. Cosley, J. A. Konstan and J. Riedl, *PolyLens: a Recommender System for Groups of Users*, Proc. European Conference on Computer Supported Cooperative Work, ECSCW' 2001, Bonn, Germany 2001.

[Oppermann, 1994] R. Oppermann, *Adaptively supported adaptivity*, International Journal of Human-Computer Studies 40, 1994, 455-472.

[Paramythis et al., 2001] A. Paramythis, A. Totter, and C. Stephanidis, *A Modular Approach to the Evaluation of Adaptive User Interfaces*, S. Weibelzahl, D. Chin, and G. Weber (Eds.) Empirical Evaluation of Adaptive Systems. Proceedings of workshop held at the Eighth International Conference on User Modeling in Sonthofen, Germany, July 13th, 2001, 9-24.

[Pearson, 2000] J. Pearson, *Speech Enhancement for DTV*, Slides presented at UM' 2001 Workshop on Personalization in Future TV, Sonthofen, Germany, 2001.

[Perkowitz and Etzioni, 2000] M. Perkowitz and O. Etzioni, *Adaptive Web Sites*. Communications of the ACM, 43(8), 2000, 152-158.

[Petrelli et al., 1099] D. Petrelli, A. De Angeli, G. Convertino*, A user centered approach to user modeling*, Proc. 7th Int. Conf. on User Modeling, 1999, pp. 255-264.

[Preece et al. 1994] J. Preece, Y. Rogers, H. Sharp, D. Benyon, *Human-computer interaction,*. Addison-Wesley Pub, 1994.

[Preece et al., 2002] J. Preece, Y. Rogers, and H. Sharp, *Interaction Design: Beyond Human-Computer Interaction*, New York, NY:Wiley, 2002.

[Reeves and Nass, 1996] B. Reeves, C. Nass, *The Media Equation: How People Treat Computers, Televisions, and New Media as Real People and Places*, Cambridge University Press, New York, 1996.

[ReplayTV, 2000] *ReplayTV* Inc, 2000, http://www.replaytv.com/flat.html

[Resnick and Varian, 1997] P. Resnick and H. R. Varian, *Communications of the ACM: Special Issue on Recommender Systems*, 40, 1997.

[Rich, 1989] E. Rich, *Stereotypes and User Modeling*, A. Kobsa and W. Wahlster, editors, User Models in Dialog Systems, Springer Verlag, Berlin, 1989,31-51.

[Riecken, 2000] D. Riecken, *Communications of the ACM: Special Issue on Personalization*, 43, 2000.

[Rittel and Webber, 1973] H. Rittel, and M. Webber, *Dilemmas in a General Theory of Planning*, Policy Sciences, Vol. 4, Elsevier Scientific Publishing Company, Inc. Amsterdam, 1973, 155-169.

[Rossi, 1958] P. Rossi, introduction and notes in M. Weber, *Il metodo delle scienze storico-sociali*, Torino, Einaudi, 1958.

[Rossi, 1984] P. Rossi, *Spiegazione e comprensione da Dilthey a Max Weber,* Rivista di Filosofia, LXXV, n. 1, 1984, 63-90.

[Rubin, 1994] J. Rubin, *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, John Wiley & Sons; 1 edition, 1994.

[Salton and McGill, 1983] G. Salton, and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1984.

[Sarwar, 1998] B. M. Sarwar, J. A. Konstan, A. Borchers, J. Herlocker, B. Miller, and J. Riedl, *Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System*, Proceeding of the ACM Conference on Computer Supported Cooperative Work (CSCW), 1998.

[Schwab et al., 2000a] I. Schwab, W. Pohl and I. Koychev, *Learning to recommend from positive evidence*, Proceedings of 2000 International Conference on Intelligent User Interfaces, New Orleans, LA, 2000, pp. 241-247.

[Schwab et al., 2000b] I. Schwab, A. Kobsa and I. Koychev, Learning about Users from Observation. In Adaptive User Interfaces: Papers from the 2000 AAAI Spring Symposium. Menlo Park, CA: AAAI Press.

[Schwarzkopf, 2001] E. Schwarzkopf, *An Adaptive Web Site for the UM 2001 Conference*, The UM2001 Workshop on User Modeling, Machine Learning and Information Retrieval, Sonthofen, Germany, 2001.

[Sears and Shneiderman] Sears, A. & Shneiderman, B. (1994), *Split menus: Effectively using selection frequency to organize menus*, ACM Transactions on Computer-Human Interaction, 1, 1, 1994, 27-51.

[Serco, 2000] Serco, 2000, http://www.usability.serco.com

[Shardanand and Maes, 1995] U. Shardanand and P. Maes, *Social Information Filtering for Automating "Word of Mouth",* In Proceedings of CHI-95, Denver, CO., 1995.

Bibliography

[Shneiderman, 1997] B. Shneiderman, *Direct Manipulation for Comprehensible, Predictable and Controllable User lntcrfaces*, Proceedings of 1997 International Conference on Intelligent User Interfaces, (eds.) J. Moore, E. Edmonds, and A. Puerta, ACM, Orlando, Florida, 1997.

[Shneiderman, 1998] B. Shneiderman, *Designing the User Interface*, Addison-Wesley, 3rd revised edition, 1998.

[Singularis, 2000] Singularis, Singularis Personalization and Profiling Platform (S.3P), 2000, http://www.singularis.ch

[Smyth and Cotter] B. Smyth, and P. Cotter, *Personalized Adaptive Navigation for Mobile Portals*, Proceedings of the 15th European Conference on Artificial Intelligence – Prestigious Applications of Intelligent Systems, Lyons, France, 2002.

[Smyth et al., 2002a] B. Smyth, P. Cotter, J. Ryan, *Evolving the Personalized EPG - An Alternative Architecture for the Delivery of DTV Services*, Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[Smyth et al., 2002b] B. Smyth, D. Wilson and D. O' Sullivan, *Improving the Quality of the Personalized Electronic Programme Guide*, Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[SONICBlue, 2002] SONICBlue, ReplayTV, 2002, http://www.replaytv.com

[Specht and Kobsa, 1999] M. Specht, and A. Kobsa, *Interaction of Domain Expertise and Interface Design in Adaptive Educational Hypermedia.*, Workshops on Adaptive Systems and User Modeling on the World Wide Web at WWW-8, Toronto, Canada, and UM99, Banff, Canada, 1999.

[Spenke and Beilken, 1999] M. Spenke and C. Beilken, *Discovery Challenge: Visual, Interactive Data Mining with InfoZoom – the Financial Data Set*, Workshop Notes on "Discovery Challenge", 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'99, 1999, pp. 33-38.

[Spenke et al., 1996] M. Spenke, C. Beilken, and T. Berlage, *The Interactive Table for Product Comparison and Selection*, UIST 96 Ninth Annual Symposium on User Interface Software and Technology, Seattle, 1996, pp. 41-50.

[Spiliopoulou, 2000] M. Spiliopoulou, *Web usage mining for Web site evaluation*, Communications of the ACM, (43) 8, 2000.

[Stone and Lester, 1999] B. A. Stone and J. C. Lester, *Dynamically sequencing an animated pedagogical agent*, Proceedings of the Thirteenth National Conference on Artificial Intelligence, 424-431, 1996.

Bibliography

[Strauss and Corbin, 1998] A. L. Strauss and J. M. Corbin, *Basics of qualitative research: techniques and procedures for developing grounded theory*, SAGE, Thousand Oaks, 1998.

[Suchman, 1987] L. Suchman, *Plans and Situated Actions*, Cambridge University Press, Cambridge, UK, 1987.

[Summerskill et al., 2002] S. J. Summerskill, G. E. Burnett and J. M. Porter: *Feeling your way home: the design of haptic control interfaces within car*, Human Factors In Transportations, Communication, Health and the Workplace, Shaker Publishing, 2002, 189-194.

[Thompson et al., 2000] C.A Thompson, M. Göker, *Learning to Suggest: The adaptive place advisor*, AAAI Spring Symposium on Adaptive User Interfaces, Stanford March 2000, AAAI, 130-135.

[TiVo, 2002] TiVo, Inc., TiVo: TV your way, 2002, http://www.tivo.com

[Torasso and Console, 1989] P. Torasso and L. Console, *Diagnostic Problem Solving*, North Oxford Academic.

[Torre 2001] I. Torre, *Goals, tasks and Application domains as the guidelines for defining a framework for User modeling*", Proc. User Modeling 2001, Lecture Notes in Computer Science, Springer Verlag, 2001, 260-262.

[van Setten et al., 2002] M. van Setten, M. Veenstra, M. and A. Nijholt, *Prediction Strategies: Combining Prediction Techniques to Optimize Personalization*, Proc. AH' 2002 Workshop on Personalization in Future TV, Malaga, Spain, 2002.

[Vassileva and Okonkwo, 2001] J. Vassileva, C. Okonkwo, *Affective Pedagogical Agents and User Persuasion*, C. Stephanidis (ed.) Proc. "Universal Access in Human - Computer Interaction (UAHCI)", held jointly with the 9th International Conference on Human-Computer Interaction, New Orleans, USA, 397-401, 2001.

[Visciola, 2000] M. Visciola, *Usabilità dei siti Web*, Milano, Apogeo, 2000.

[Weber et and M. Specht, 1997] G. Weber and M. Specht, *User modelling and adaptive navigation support in WWW-based tutoring systems*, Proc. of 6th International Conference on User Modelling, 1997, 289-300.

[Weibelzahl and Lauer, 2001] S. Weibelzahl and C. U. Lauer, *Framework for the Evaluation of Adaptive CBRSystems*, U. Reimer, S. Schmitt, & I. Vollrath (Eds.), Proceedings of the 9th German Workshop on Case-Based Reasoning (GWCBR01), Aachen: Shakerm, 2001, 254-263.

[Weibelzahl and Weber, 2001] S. Weibelzahl and G. Weber, *A database of empirical evaluations of adaptive systems,* R. Klinkenberg, S. Rüping, A. Fick, N. Henze, C. Herzog, R. Molitor, & O.

Schröder (Eds.), Proceedings of Workshop Lernen - Lehren - Wissen - Adaptivität (LLWA 2001), research report in computer science nr. 763, University of Dortmund, 2001, 302-306.

[Weibelzahl et al., 2001] S. Weibelzahl, D. N. Chin, & G. Weber (Eds.), Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001, 2001, http://art.ph-freiburg.de/um2001/prodeedings.html.

[Weiser, 1991] M. Weiser, *The Computer for the 21st Century*, Scientific American (September), 1991, 66-75.

[Wenger, 1998] E. Wenger, *Communities of Practice: Learning, Meaning and Identity*, Cambridge: Cambridge University Press, 1998.

[Whiteside et a.l., 1988] J. Whiteside, J. Bennett, and K. Holtzblatt, *Usability Engineering: Our Experience and Evolution*, Handbook of Human-Computer Interaction (M. Helander, ed.), New York: North-Holland, 1988, 791-817.

[Wilkinson et al., 2000] R. Wilkinson, S. Lu, F. Paradis, C. Paris, S. Wan, Mi Wu., *Generating personal travel guides from discourse plans*, Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2000), Lecture Notes in Computer Science 1892, Springer Verlag, 2000, 392-85.

[Winograd and Flores, 1986] T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design*, Norwood, NJ: Ablex, Paperback issued by Addison-Wesley, 1986.

[Wittgenstein, 1953] L. Wittgenstein, *Philisophical investigations,* Blackwell, Oxford, 1953.