The Knowledge Engineering Review, Vol. 20:1, 1–37. © 2005, Cambridge University Press
doi:10.1017/S0269888905000299 Printed in the United Kingdom

Methods and techniques for the evaluation of user-adaptive systems

CRISTINA GENA

Dipartimento di Informatica, Università degli Studi di Torino, Corso Svizzera 185, 10149 Torino, Italy;
E-mail: cgena@di.unito.it

Abstract

This article presents a comprehensive overview of methods and techniques used for the evaluation of user-adaptive systems. It describes the methodologies derived both from the evaluation of human–computer interaction systems and from information retrieval and information filtering systems by giving examples of the application of these methodologies in the user-adaptive systems. The state of the art and the main results in the evaluation of these systems are reported. In particular, empirical evaluation and layered approaches are discussed in detail. Finally, focus on less explored methodologies, such as qualitative approaches (e.g. Grounded Theory), is proposed.

1 Introduction

The evaluation of adaptive and user modeling (UM) systems (Brusilovsky, 1996, 2001; Kobsa *et al.*, 2001; Maybury & Brusilovsky, 2002) is a fundamental stage in their development and it should become common practice. As the application of these systems moves from the research lab to real field usage, the evaluation of the real user–system interaction becomes fundamental. The adaptations generated by user modeling techniques often tend to improve the user–system interaction. Since most of the time the exploitation of such techniques makes the system more complex, it should be evaluated whether the adaptivity really improves the system and whether the user really prefers the adaptive version of it. Moreover, a user of an adaptive system has more expectations and therefore she is more frustrated when the application does not work as she expected. Evaluation studies have shown that the users often have problems with the adaptive features of a system, and thus avoid using them. Therefore, the challenge of adaptive systems becomes to demonstrate that their exploitation can improve the interaction by testing the utility of the adaptation choices.

Another good reason for a deeper user involvement is the closeness with human–computer interaction (HCI) techniques focused on user participation. As HCI systems, adaptive systems should also adopt a user-centered approach because users are both the main source of information, and the main target of the application. As in the evaluation of regular interactive 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.

However, in the evaluation of adaptive systems, difficulties can be caused by the need to distinguish different adaptation aspects and therefore evaluate them separately. Layered approaches have been proposed in the literature in order to break down adaptation in its constituents and to separately evaluate every identified constituent. The origin of this idea can be traced to Totterdell and Boyle (1990), who proposed that different adaptation metrics have to be related to different components of a logical model of adaptive user interfaces. For Totterdell and

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Boyle 'Two types of assessment were made of the user model: an assessment of the accuracy of the model's inferences about user difficulties; and an assessment of the effectiveness of the changes made at the interface'. We can see that this separation is one of the peculiarities that differentiate the evaluation of user-adaptive systems from the evaluation of 'non-adaptive' systems. The evaluation cannot be performed by considering the system as a whole, but at least two layers have to be evaluated separately: the content layer and the interface layer. Moreover, it has to be taken into account that a user-adaptive system will be different for different users, and the evaluation could underline that these differences are useful for some users and for others they are not. Furthermore, the evaluation of an adaptive system can provide feedback that can be used to modify the adaptation strategies of the system itself. Thus the evaluation phases of an adaptive system can be considered as *generative* models since they can contribute during the design of the application and can provide the means of combining design specification and evaluation into the same framework (Dix *et al*., 1998).

This paper presents a comprehensive overview of methods and techniques that can be used for the evaluation of adaptive systems, from the well-known HCI techniques through the selection process evaluation methodologies, to qualitative approaches.

Most techniques are currently used in the evaluation of adaptive systems, even though some of them, such as controlled experiments, sometimes are not properly designed and thus they do not produce significant results. For this reason the correct design of these methodologies will be proposed in a more detailed way. Some other techniques presented in this paper are often disregarded, such as ethnographic studies, task analysis, and systematic observation, but they should be considered since, as it will be shown, they could offer interesting results. Most real examples will be given, however the way in which evaluation results have to be interpreted and linked to the adaptation components that need to be changed will be roughly sketched since this process highly depends on the features of every single application and on how adaptations are implemented.

The paper is organized as follows. Section 2 gives a detailed description of HCI evaluation methodologies, presenting methods and examples, catalogued according to the more accepted classification in the literature. Discussions about usability problems and user-centered approaches to the development and the evaluation of user-adaptive systems are illustrated and solutions proposed. Section 3 presents methodologies for the evaluation of the information selection process, which are borrowed from the evaluation of information retrieval and information filtering systems. These methodologies are also largely applied to user-adaptive systems, in order to assess the relevance of the contents presented to the users. Section 4 sketches the state of the art in the evaluation of these systems, both in quantitative terms (How many papers in the main conferences and journals report evaluations? How many evaluations report significant results?) and in qualitative terms (the main surveys and discussions that have arisen in the last few years). Section 5 illustrates qualitative methodologies of evaluations, which are seldom taken into consideration in the field of user-adaptive systems, but they could lead to effective results. Finally, Section 6 concludes the paper.

Most of the examples reported in the paper refer to works presented at the last User Modeling and Adaptive Hypermedia conferences (Kay, 1999; Brusilovsky *et al*., 2000, 2003; Bauer *et al*., 2001; De Bra *et al*., 2002), the Workshop on Evaluation of Adaptive Systems (Weibelzahl *et al*., 2001; Weibelzahl & Paramythis, 2003) and the journal *User Modeling and User-adaptive Interaction: The Journal of Personalization Research* (UMUAI).

2 HCI-oriented evaluation methodologies

The methodologies for evaluating adaptive systems are generally borrowed from the methodologies used in HCI and by those utilized for the evaluation of the information selection process (see Section 3 for details).

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

The HCI methodologies can be used in the evaluation of adaptive systems mostly to evaluate the interface adaptations, the usability of adaptive systems, to collect users' and experts' opinions, etc. Such methodologies can also be used for the evaluation of the information selection process in order to collect user data important to the analysis of the process.

One of the possible classifications of evaluation methodologies in the HCI is the following (Dix *et al.*, 1998; Preece *et al.*, 1994).

- *Collection of user's opinions*. This group of methodologies can be used to collect real user data that can be further processed in order to obtain both quantitative (questionnaires) and qualitative (interviews, focus groups) information. In user-adaptive systems evaluation, collection of users' opinions is often helpful in gathering information to evaluate the success of content adaptations (see Section 3).
  - *Interviews*, which are used to collect self-reported experiences, opinions, preferences and behavioral motivations. Interviews can be structured, semi-structured, unstructured. Interviews are often carried out after a test session to gather user's opinion, such as the user's satisfaction with the system (for an example, see Gena *et al.*, 2001). Interviews can also be used in interpretative evaluation (see Section 5).
  - *Questionnaires*, wherein both the questions and the answers are fixed. The questionnaire can be filled in by the user or read by the interviewer (e.g. face-to-face interviews, telephone interviews, etc.). However, in the latter case the influence of the interviewer in the completion of questions has to be strictly controlled in order to avoid possible interferences. Questionnaires can be helpful to collect user information important in gathering knowledge of the system the user modeling or system adaptations (requirement gathering phase; for details, see Section 2). For instance, Bental *et al.* (2003) used a range of questionnaires in different periods, given to the target users, to measure the effects of system usage. Since a large number of questionnaires are required in order to generalize results (see Section 4.1) existing user surveys are sometimes utilized to build the knowledge base of a system or to build stereotype-based user-modeling systems (Gena, 2001; Hara *et al.*, 2004). Questionnaires can also be used to compare the system assumptions with the real user choices (Ardissono *et al.*, 2004) in order to evaluate the information selection process performed by the system as follows.
    * *On line questionnaires*, to collect users' opinions (with the same aim as off-line questionnaires) or to gather user preferences in order to generate suggestions. In the latter case, questionnaires are used not only during experimental sessions but also during the real running of the system. They can be used to acquire user interest profiles in collaborative (Malone *et al.*, 1987; Sharnadand & Maes, 1995) and feature-based recommender systems (Resnick & Varian, 1997). At the beginning, the system can use the user's rating to generate recommendations based on the choices of similar neighbors or to produce feature-based recommendations. Then, the collected data can be used to evaluate the effectiveness of the information selection process; see, for example, McNee *et al.* (2003).
    * *Post-test questionnaires*, to collect structured information after an experimental session, a usability test, or after having tried a system for a while. For instance, Matsuo (2003) asked the users questions about the system features using a five-point Likert scale;[1] Alfonseca and Rodríguezi (2003) asked questions about the usability of their system; Bull (2003) administered a post-test questionnaire to students to ascertain the utility of their system. Post-test questionnaires can also be used to compare the assumption in the user model with an external text (Weibelzahl & Weber, 2003).
    * *Pre- and post-test questionnaires*, to collect changes after an experimental session, or after a real user–system interaction. For instance, in student model and adaptive learning systems, pre- and post-test questionnaires are utilized to register any improvement in the student's knowledge after one or a set of interactions. In the easiest design the questionnaires compare

[1] A rating scale measuring the strength of agreement towards a set of clear statements.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

the results of two groups of students: one group that has been interacting with the adaptive version of the system and the other group that has been interacting with the non-adaptive system. If the adaptations are successful the first group of users will obtain higher scores in the post-test questionnaire. Pre-test questionnaires are also administered to group students on the basis of their ability (Mitrovic, 2001; Jackson *et al*., 2003), their knowledge (Specht & Kobsa, 1999), their motivational factors and their learning strategies (Kurhila *et al*., 2002) and then the results of the different groups can be compared with post-test questionnaires.

  * *Focus group*, which is a structured discussion about specific topics moderated by a trained group leader (Greenbaum, 1998). 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. For that reason, they can be used before system implementation, during the requirement analysis phase. Depending on the kind of user involved (e.g. final users or domain experts) they can be utilized to gather functional requirements, data requirements, usability requirements, and environmental requirements (see Section 2.1 for details and Gena & Ardissono (2004) for an example). Focus groups can also be used in interpretative evaluation (see Section 5).

- *Observing and monitoring usage*. This group of methodologies can be used to collect data from real usage that can be further analyzed quantitatively (the systematic observation, logging use) and qualitatively (user observation, verbal protocol and think aloud protocols) in order to obtain information important in the global evaluation of user-adaptive systems.
  - *User observation*, which is aimed at learning from the direct or indirect observations of users (such as video recording; see for example, Stacey *et al*., 2003) in the context of real usage by interviewing and observing users at work, while doing their own work. These methods are most reliable and precise but they are also very expensive. They are related to the ethnographic investigations and interpretative evaluation that will be discussed in detail in Section 5.
  - *The systematic observation*, which 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 & Gottman, 1986, p. 4). The aim is first to define various forms of behavior (behavioral codes) and then to ask observers to record when behavior corresponds to predefined codes. On a more analytic point of view, two types of analysis can be undertaken: task-based analysis and performance-based analysis (Preece *et al*., 1994). The systematic observation can be used during the requirement gathering phase (see Section 2.1) to systematically analyze the user–system interaction.
  - *Verbal protocol and think aloud protocols*, wherein users are asked to think out loud when they are performing a task in an experimental session or a usability test in order to record their spontaneous reactions (Ericsonn & Simon, 1984). See, for example, Pirolli and Fu (2003).
  - *Logging use*, which can be considered a kind of indirect observation and consists of the analysis of log files that register all the actions of the users. Log files analysis shows the real behavior of the users and is one of the most reliable ways to demonstrate the real effectiveness of user modeling and adaptive solutions. See, for example, Baudisch and Brueckner (2002), Kurhila *et al*. (2002), Smyth and Cotter (2002), Bull (2003), Dimitrova (2003), Pirolli and Fu (2003), Beck *et al*. (2003b). Often log data are used to run algorithm with real user data (Scarano *et al*., 2002), for simulation such as to reproduce Web surfing (Kufklik *et al*., 2003), and to simulate e-mail usage (McCreath & Kay, 2003). See Section 3.5 for details on simulation.
- *Predictive evaluation*. This group of low-cost evaluation methods predicts the usability from models, specifications or early prototypes (Preece *et al*., 1994). They are usually carried out by usability or domain experts.
  - *Heuristic evaluation* (Nielsen & Molich, 1990), wherein a small set of evaluators examines a user interface and looks for problems that violate some of the general principles of good

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

interface design. Magoulas *et al.* (2003) proposed an integration of heuristic evaluation in the evaluation of adaptive learning environments. They modified Nielsen's heuristics (1993) to reflect pedagogical consideration and then they collocated their heuristics into the level of adaptation proposed by Weibelzahl and Lauer (2001).

  - *Expert review*, which is an evaluation method wherein experts play the role of less experienced users in order to identify usability problems. Experts can be asked to simulate the system usage by the typical users and to check the correctness of system adaptations at the interface (Console *et al.*, 2003). The experts' choices can also be compared with system classifications (Magnini & Strapparava, 2001; Ardissono *et al.*, 2004; Goren-Bar *et al.*, 2001; Teo, 2001). Experts can also be interviewed during the knowledge gathering process.
  - *Parallel design*, wherein different design alternatives are explored before settling on a single proposal to be further developed (Nielsen, 1993). 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). The design rationale[2] and design space analysis[3] can also be helpful in the context of exploring and reasoning among different design alternatives. For details about design rationale, see Lee and Lai (1991), while for design space analysis, see Bellotti and MacLean (nd).
  - *Cognitive walkthroughs* (Lewis *et al.*, 1990), wherein an expert evaluates the usability of a design. The evaluator constructs task scenarios from a specification, and then role plays the part of a user. Potential problems are evaluated against psychological criteria.

- *Formative evaluation*, which is aimed at checking the first design choices and getting clues for revising the design. Formative evaluation takes place before actual implementation and influences the development of the system. As in the regular interactive system development, formative evaluation can be used in the first design stages of an adaptive system. However, this technique is not satisfactory for a complete evaluation of the system and researchers should use it merely to collect initial feedback and to avoid initial design mistakes.
  - *Wizard of Oz simulation*, which 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.
  - *Scenario-based design*, which is a method used to describe existing activities or to foresee new activities. A scenario is aimed at illustrating a usage situation by showing the user's actions step by step. It can be represented by textual descriptions, images, and videos and it can be employed in different design phases.
  - *Prototypes*, which are artifacts that simulate or animate some, but not all the features of the intended system (Dix *et al.*, 1998), can be divided into two main categories: static, paper-based prototypes and interactive, software-based prototypes. The software prototypes can be further classified into *horizontal prototypes*, when they contain a shallow layer of the whole surface of the user interface, and *vertical prototypes*, when they include a small number of deep paths through the interface, but do not include any part of the remaining paths. See, for example, Ardissono *et al.* (2004).
- *Experiments and tests*
  - *usability testing*, 'which means making sure that people can find and work with the functions to meet their needs' (Dumas & Redish, 1999, p. 22). A usability test is characterized by the following features.
    * the primary goal is to improve the usability of a product,
    * participants represent real users,
    * participants do real tasks,

[2] *Design rationale* 'is the information that explains why a computer system is the way it is, including its structural or architectural description and its functional or behavioral description' (Dix *et al*, 1993, p. 212).
[3] *Design space analysis* is an 'approach to design that encourages the designer to explore alternative design solution' (Preece *et al*, 1994).

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

  * users' performances are recorded,
  * data are analyzed and, as a consequence, changes recommended. One or more usability test of an adaptive system should always be performed, as in regular HCI systems. Usability of the system should be tested taking into account both interface adaptation and general interface solutions. For example, see Scarano *et al*. (2002), Alfonseca and Rodríguezi (2003), Bental *et al.* (2003), Santos *et al*. (2003), while for details on the usability test, see Dumas and Redish (1999) and Rubin (1994).

– *Experimental evaluation*, which refers to the appraisal of a theory by observation in experiments. Experimental evaluation will be discussed in more detail in Section 4.1.

2.1 Evaluation and usability problems in user-adaptive systems

Following the accepted definition in HCI, evaluation can be described as a process through which information about the usability of a system is gathered in order to improve the system or to assess a completed interface, and the evaluation methods are procedures for collecting relevant data about the operation and the usability of a software system (Preece *et al*., 1994). Several discussions have arisen about the usability of adaptive interfaces.

As pointed out by Höök (2000), intelligent user interfaces may violate several major usability principles, such as giving the user control over the system, making it predictable (in the sense that the same input always causes the same response), and making the system transparent so that the user can understand its inner workings. In addition, most adaptive interface developers are more concerned with defining inference strategies than with interface design. For Höök, the usability of intelligent user interfaces sometimes can require a new way of addressing usability, different from the usability principles outlined for direct-manipulation systems.

Benyon (1993) proposed adaptivity as a solution to usability problems, since a single interface cannot be designed to meet the usability requirements of all the groups of users of a system. However, adaptive systems can improve usability only if it can be shown that, without the adaptive capability, the system would have performed less effectively. He discusses five related and interdependent activities that need to be considered when designing adaptive systems:

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

According to Benyon, these five phases are derived from the categories that have to be considered during the requirements analysis phase of a software system. *Requirements analysis* (Preece *et al*., 1994) can be defined as the process of finding out what a user requires from a software system and focuses on what is required but not on how to provide it. Requirement gathering consists of looking up similar software systems, discussing the needs of people who will use the system, analyzing existing systems to discover problems with current design, observing the natural work settings, prototyping.

The above activities presented by Benyon directly correspond to the following stages of the requirement analysis (Preece *et al*., 1994).

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

- *Functional requirements*, which specify what the system must do.
- *Data requirements*, which specify the structure of the system and the data that must be available for processing to be successful.
- *Usability requirements*, which specify the acceptable level of user performance and satisfaction with the system. The activity of gathering usability requirements is often known as usability study and it is closely allied to the evaluation process. In addition to the technique described in Section 2, which can be chosen depending on the status of the system under analysis (running system, new system, etc.), other analyses can be undertaken.
  - *Task analysis*, which is based on the analytic de-composition and the re-composition of users' actions and users' cognitive processes to complete tasks. This method can be used to investigate users' actions and to decide which phase of the interaction has to be adapted,
  - *Cognitive and socio-technical models*, which are similar to task analysis methods, but consider other issues. Cognitive 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, p. 230). Socio-technical models instead consider social and technical issues and recognize that technology is a part of the wider organizational environment. Both goal-oriented cognitive models and socio-technical models can be considered as *generative* models (Dix *et al*., 1998). Moreover, they could make predictions about the behavior of different types of users and therefore be adopted in the construction of the user models and their corresponding system adaptations.

As well as obtaining detailed information about users and their work, in requirements gathering, it is also important to consider the effects of the *physical environment* (Preece *et al*., 1994). The job of system designer should be to achieve a harmony between users, tasks, environments and systems.

As Benyon (1993) underlined, adaptive systems should benefit more than other systems from a requirement analysis before starting any kind of evaluation, because a larger number of features have to be taken into account in the development of these systems. The recognition that an adaptive capability may be desirable leads to improved system analysis and design. As a demonstration, he reported an example of an adaptive system development, wherein he prototyped and evaluated the system with a number of users. Several user characteristics were examined to determine their effects on the interaction. Then, further task knowledge and functional analysis were identified.

A user-centered perspective is also proposed by Oppermann (1994), which 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:

- a problem solvable by adaptivity has to be identified;
- the user characteristics related to the hard problem have to be detected;
- ways of inferring user characteristics from the interaction have to be found;
- adaptation techniques offering the correct 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 associated to specific user actions. Furthermore, the adaptations must be potentially useful to the user. The necessity of an iterative process is due to the fact that the real behavior of users in a given situation is hard to foresee; therefore, some evidence can be shown only by monitoring the users' activity. From the iterative evaluation point of view, the design phases and their evaluation have to be repeated until good results are reached.

Oppermann's iterative process is very close to the user-centered system design approach. The essential principles of user-centered design are to make user issue central in the design process, to carry out early testing and evaluation with users and to design iteratively.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Gould and Lewis (1983) originally phrased this principle as follows:

- focus on users and their tasks early in the design process;
- measure reactions by using prototype, interfaces or other simulation of the system;
- design iteratively.

Similarly, Norman and Draper (1986) discussed a user-centered approach. They assumed that if system designers always take into account features, habits, preferences and behavior of the users, they would be able to produce easier-to-use systems. During the different stages 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 (Preece *et al*., 1994):

- 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 Section 5). In this last case the subjects involved are members 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.

The 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 empirical knowledge of the users' behavior from the first development phases. In the case of a user modeling system, the choice of the 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 users of the system, 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 the case of innovative systems.

Petrelli *et al*. (1999) 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 incremental system design as a way to satisfy more users. 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, after having analyzed the results of a questionnaire they distributed to 250 visitors. 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 do not like going to the museum alone) and context-aware perspective (museum visitor prefers watching paintings than interacting with a device).

Similar to the user-centered perspective is the continuous empirical evaluation approach of Ortigosa and Carro (2003). They proposed continuous evaluation of adaptive Web courses in order to identify possible failures or improvement possibilities and to offer concrete suggestions on how to improve them. In their evaluation prototype they propose continuously analyzing (i) adaptive course description, (ii) users' profile, performances, and actions, (iii) students' opinions. For instance, they evaluate the navigation paths (using a pattern recognition algorithm) that lead to bad and good scores and if significant differences emerge, they adopt possible improvements.

3 The selection process evaluation

The main metrics are derived from the evaluation of information retrieval and information filtering systems, since the problem is quite similar: from a set of contents a sub-set of user-relevant contents

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

has to be extracted. For details on methods and measurements for the evaluation of information filtering systems, see Hanani *et al.* (2001). In user-adaptive systems these metrics are exploited to evaluate content adaptations (e.g. accuracy of recommendations, accuracy of system predictions and/or system preferences, similarity of expert rating and system prediction, inferred domain knowledge in the user model, etc.). The data to be analyzed by these metrics are usually gathered using the methodology described in Section 2 (in particular, questionnaires and log files).

3.1 Precision and recall

Two fundamental measures in the selection process evaluation, based on *true positives* (TPs, the number of relevant contents that the system proposes to users), *false positives* (FPs, the contents that have been suggested to users but they do not like), and *false negatives* (FNs, the contents that have not been suggested to users but they do probably like), are (Salton & McGill, 1983):

- *precision*, which refers to the 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:

$$precision = \frac{TP}{TP + FP}$$

- *recall*, which is the ratio between the user-relevant contents and the contents present in the collection (thus also including the contents the system does not suggest even if they can be relevant to the user):

$$recall = \frac{TP}{TP + TN}$$

Hence, precision indicates how accurate the system is, and recall indicates how thorough it is in finding valuable information. As can be noted, precision and recall can be calculated when the classification of the contents is dichotomic: they are classified as relevant or not relevant to the user.

For examples of evaluation of recommender systems (Resnick & Varian, 1997) with the estimation of precision and recall returned to a human advisor proposal, see Magnini and Strapparava (2001); for a precision/recall evaluation of algorithm performance, see Goren-Bar *et al.* (2001); Teo (2001) and Zhu *et al.* (2003), for an evaluation made with hypothetical user profiles, see Sunderam (2002); and for the precision/recall of different rules for automatic classification, see McCreath and Kay (2003). Ruvini (2003) calculates the predictive accuracy of a system that models user search strategy as a percentage of examples of the test set correctly predicted as positive or negative.

In machine learning systems and in inductive reasoning-based systems, the selection process is evaluated starting from the data of the real system usage (e.g. the log files, the purchase data of an e-commerce site) which are divided into:

- *training set*, usually 80% of the available data;
- *test set*, the remaining 20%.

In particular, the training set is used to form the learned hypothesis, while the test set is used to evaluate the accuracy of this hypothesis over an independent and identically distributed sample of data (for details, see Mitchell, 1997).

In collaborative systems, for instance, the training set data are used to find and select the user's neighbors and then generate the recommendations. The remaining data (test set) are used to evaluate the accuracy of the recommendations. In this case, the precision is calculated as the ratio between the user-relevant contents and the number of recommendations, while the recall is

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

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 correct suggestions among the overall suggestions, while the recall is the ratio between the correct suggestions and the suggestions that should be correct for the selected subjects.

For an example of generation of a training and test set starting from data logs for an automatic user classification, in a system learning interaction model, see Semeraro *et al.* (2001); for an estimate of the precision and recall with different size training sets, see Goren-Bar *et al.* (2001); for an example of a training set used to learn a user model and a test set used to check this model in a system that models a user search goal, see Ruvini (2003); for an example of an iterative training–testing approach, see Beck *et al.* (2003a).

3.2 *Utility metrics*

The utility (Hanani *et al.*, 2001) is a metric able to detect false positives and false negatives. The utility is calculated using the following formula:

$$utility = (A*TP) + (B*TN) + (C*FP) + (D*FN)$$

where A, B, C, D are multiplicative coefficient is used to establish the relative benefit (when they are positive) or the relative cost (when they are negative). The concept of false positives is particularly important in the case of personalized e-commerce Web sites, for instance, and it should be reduced when it is not related to users' purchases (Tasso & Omero, 2002).

Another metric focused on errors is the *error rate* (ε) (Lewis, 1995), which is defined by the following formula (using the above notation):

$$\varepsilon = \frac{FN + FP}{TP + FP + TN + FN}$$

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

3.3 *Coverage*

Sarwar *et al.* (1998) 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.* (1998, p. 6) used 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'.

3.4 *Accuracy*

Accuracy can be measured in different ways. The two general approaches used are the statistical recommendation accuracy and the decision-support accuracy.

Statistical recommendation accuracy (Good *et al.*, 1999; Sarwar *et al.*, 1998) evaluates the accuracy of a filtering system by comparing the numerical prediction values against user ratings for

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

the items that have both predictions and ratings. The main metrics used are mean absolute error (MAE), root mean squared error (RMSE) and correlation.

MAE and RMSE evaluate the distance between the system predictions and the user's opinion using rate vectors. A smaller value means more accurate system prediction.

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

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |u_i - r_i|}{n}$$

where n is the number of the contents, u_i is the user's opinion on the content i, and r_i is the system prediction on the content i.

For instance, in order to assess the effectiveness of different interfaces to elicit information from new users for a recommender system, McNee *et al.* (2003) use:

- a '*30 based on 12*' or a '*30 based on N*' MAE calculation by taking either the first 12 or all N ratings from the signup process and then generating predictions for 30 randomly selected ratings the user entered into the system after finishing signup;
- an '*all but one*' calculation, by removing one of the first 12 user ratings and generating a prediction for it by using the other 11 ratings.

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 notation):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (u_i - r_i)^2}{n}}.$$

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

Good *et al.* (1999) suggest that a good value of MAE should be near to 0.7, on a scale of values ranging from 0 to 5. As can be noted, these metrics support applications in which the system predictions and the user's opinion take non-dichotomic values.

Correlation can be used as a measure of linear agreement between two vectors: user evaluations are correlated with system evaluations. A higher correlation value indicates more accurate recommendations. For instance, Jackson *et al.* (2003) measured the correlation between student ability and the proportion of various dialog moves in a tutoring system that implements conversational dialog as tutoring strategies.

A statistic usually exploited to measure the 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 a linear relationship between two variables. It ranges from −1.0 to +1.0 (perfect negative and perfect positive correlations, respectively). A value of zero represents the complete absence of correlation. The Pearson correlation coefficient is calculated by the following formula:

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

$$r = \frac{\sum (X_i - \overline{X})\,(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2\,(Y_i - \overline{Y})^2}}$$

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

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 ratio of the joint variation of X and Y relative to the variation of X and Y considered separately.

An important competitor of the Pearson correlation coefficient is the *Spearman rank correlation coefficient* (ρ). This latter correlation is exploited when the variables involved in the relationship are not normally distributed and are calculated by applying the Pearson correlation formula to the ranks of the data rather than to the actual data values themselves. For details, see Keppel (1991).

Correlation tells us whether there is a relationship between two variables. If we want to use correlation data to make predictions on a variable on the basis of another variable we can exploit a *regression equation*. For details, see Keppel (1991).

For instance, Beck *et al.* (2003b) measured the prediction accuracy of a reading tutor that listens to the users by calculating the correlation between predictions and actual test scores and by using a linear regression algorithm to make predictions.

Accuracy can also be calculated in other ways. In order to cross-validate learning algorithms for an adaptive educational collaborative environment, Gaudioso and Boticario (2002) compute the accuracy by simply recording a correct prediction for all the instances for which at least one of the algorithms made a correct guess.

Decision support accuracy metrics (Good *et al.*, 1999; Sarwar *et al.*, 1998) evaluate how effective a prediction engine is at helping users select high-quality items from the item set. These metrics are based on the observation that in most of the cases, filtering is a binary operation: users will either view the item, or not. The main metrics are follows.

- *Reversal rate* measures the percentage of reversal recommendations (the contents the user does not like). For more details, see Good *et al.* (1999) and Sarwar *et al.* (1998).
- *ROC sensitivity* is a signal processing measure of the decision-making power of a filtering system. For more details, see Good *et al.* (1999) and Sarwar *et al.* (1998). For instance, Kim *et al.* (2002) used MAE, ROC sensitivity, ROC specificity, ROC accuracy, and ROC error rate to calculate the distances between users' explicit preferences and the recommendations provided by a case-based recommender system that uses implicit rating information (access logs).
- *PRC sensitivity* is a measure of the degree to which the system presents relevant information. For more details, see Sarwar *et al.* (1998).

3.5 The simulation

In the simulation methodology the same data set is utilized in different experimental sessions. Different selection algorithms and different solutions are tested on the same sets of data in order to reach the optimal results (which are often evaluated using the accuracy metrics described above). For instance, in recommender systems, the users can evaluate all the contents, and then the behavior of the system is simulated in different experimental sessions using the collected data with different algorithms. Then accuracy of results can be calculated. Systems are also provided with the real users' evaluations for the learning process, to effect future evaluations, and to simulate the adaptive features. For example, see Shapira *et al.* (1999). The users' data can be collected by real users or simulated by the researchers. The same approach is frequently exploited in the evaluation of adaptive systems, wherein real user data are used to simulate the real user's behavior with a running system exploiting one or more adaptive solutions (Goren-Bar *et al.*, 2001; Teo, 2001; Kim *et al.*, 2002; Scarano *et al.*, 2002; Kufklik *et al.*, 2003; McCreath and Kay, 2003; Zhu *et al.*, 2003). Hypothetical user profiles (Sunderam, 2002; Castillo *et al.*, 2003) or expert proposals (Magnini & Strapparava, 2001) are also used to simulate real user behavior.

The simulation methodology is also proposed by the TREC (Text Retrieval Conference) project, managed by the NIST (National Institute of Standards and Technology). TREC is aimed at defining standards for evaluation of information retrieval systems and it is now considered the main benchmark for information retrieval comparisons (Hull, 1998). In TREC, different systems are

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

being tested with the same very large data collections. The performance of each system is compared to verify the test results.

4 The state of the art in evaluation of user-adapted systems

During the past few years, the empirical research on UM and adaptive systems has been very fragmented. From the analysis (Chin, 2001) of the papers published in the *User Modeling and User-adaptive Interaction* journal (UMUAI) during its first nine years of activity, only one-third of the articles include some type of evaluation. Moreover, most of them reported only preliminary evaluations. Excluding the latter, only one-quarter of the UMUAI papers reported statistically significant evaluation results. Resulting from my own analysis of the 2001 User Modeling (UM01) and the 2002 Adaptive Hypermedia Conferences (AH02), they include one-third and one-quarter of papers with some kind of evaluations, respectively. By excluding preliminary evaluations, these two estimates decrease to one-third and one- fifth. The last User Modeling Conference (UM03) showed an improvement in the number of evaluation studies carried out: one-half of the studies reported some kind of evaluation, but excluding preliminary evaluation, the estimate decreases to one-third.

In his PhD thesis on evaluation of adaptive systems, Weibelzahl (2003) compiled a synopsis of 43 studies published in the UMUAI journal and in the proceedings of User Modeling Conferences (1997–1999), and considered the most important works in the area of adaptive systems. He accepted that about a quarter (11/43) of the studies examine either only a single user, hypothetical users or the sample size is not reported. Actually, only 14 out of the 43 studies are high quality in terms of sample size and statistical analysis. He also reported that the most frequent measures include accuracy, precision and recall, domain knowledge and duration of the interaction. He than classified the studies according to the proposal described in Section 4.3.

Chin (2001) found a strong correlation between topic areas and empirical evaluations. He classified the user-adaptive systems and their evaluation in five areas:

- *student modeling systems*, which are typically evaluated by comparing the system behavior with and without the student models, and by analyzing real data (such as log files);
  - *systems that use machine learning methods to acquire user models*, which 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);
  - *adaptive hypermedia and information filtering systems*, which are typically evaluated by measures developed in library sciences such as recall and precision and similarity/relevance metrics;
  - *plan recognition systems*, which 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;
  - *mixed-initiative interaction systems*, which are typically evaluated by comparing system response choices with human choices or by comparing the efficiency of the dialog needed to achieve information with human–human dialogs or with theoretically minimum dialogs;
  - *user interfaces/help systems*, which are typically evaluated by subjective user satisfaction, task completion speed, error rate, and quality of task achievement.

The main studies investigating the proper evaluation methodologies of evaluation in UM and adaptive systems are concerned with the empirical evaluation (Chin, 2001) and layered evaluations of adaptive systems (Karagiannidis & Sampson, 2000; Brusilovsky *et al.*, 2001; Paramythis *et al.*, 2001, 2003; Weibelzahl, 2001, 2003; Weibelzahl & Lauer, 2001; Weibelzahl & Weber, 2001). While the first study basically reports the correct methodologies that have to be followed to carry out empirical evaluation (which are quite similar to those required for the empirical evaluation of regular HCI systems), the other studies basically propose approaches that have to be taken into

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

account for the evaluation of adaptive systems and for the correct interpretation of the obtained results.

4.1 The empirical evaluation

The empirical evaluation, also known as the controlled experiment, refers to the appraisal of a theory by observation in experiments. This method of evaluation is derived from empirical science and cognitive and experimental psychology. The general idea underlying the empirical evaluation is that by changing one element (the independent variable) in a controlled environment its effects on user behavior can be measured (on the dependent variable).

Controlled experiments are often performed in the evaluation of user-adaptive systems (mostly for the evaluation of interface adaptations), but sometimes experiments are not properly designed and thus they do not produce significant results. Since those results are required, because they can be extended in order to provide generalizations and guidelines for the future works and thus promoting an improvement in the development of user-adaptive systems, it is important to correctly carry out every design step and evaluate results with the required statistics.

According to Keppel *et al*. (1998), the schematic process of an experiment can be summarized asfollows.

- *Identify the issue or question of interest*.
- *Review the relevant theories and research*.
- *Develop a research hypothesis as a succinct statement of the purpose of the study*.
- *Identify the independent and dependent variables*. The *independent variable* comprises the range of treatment conditions under the control of the experimenter (for instance, an application with or without a user model). The *dependent variable* (also known as the response measure) comprises the behavior observed after the manipulation of the independent variable and measured afterwards. Examples of dependent variables are:
  - the task completion time;
  - the number of errors;
  - proportion/qualities of tasks achieved;
  - interaction patterns;
  - learning time/rate;
  - user satisfaction, etc.

  Whiteside *et al*. (1998) provide a list of measurement criteria (*usability metrics*) that can be used to determine the measuring method for a usability attribute. These metrics are also the ones usually considered as dependent variables in HCI controlled experiments:

  - time to complete a task;
  - percentage of task completed;
  - percentage of task completed per unit time;
  - ratio of successes to failures;
  - time spent in errors;
  - percentage or number of errors;
  - percentage or number of competitors better than it;
  - number of commands used;
  - frequency of help and documentation use;
  - percentage of favorable/unfavorable user comments;
  - number of repetitions of failed commands;
  - number of runs of successes and of failures;
  - number of times interface misleads the user;
  - number of good and bad features recalled by users;
  - number of available commands not invoked;

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

– number of regressive behaviors;
– number of users preferring a specific system;
– number of times users need to work around a problem;
– number of times users are disrupted from a work task;
– number of times users lose control of the system;
– number of times users express frustration or satisfaction.

Moreover, other variables can also be measured, such as:

– amount of required help;
– amount of requested material;
– duration of total interaction;
– number of navigation steps.

Some dependent variables can only be measured indirectly, such as:

– cognitive load measured through blood pressure;
– pupil dilatation;
– eye-tracking;
– number of fixations and fixation times;
– heart rate;
– skin temperature;
– electrodermal activity;
– pressure applied to a computer mouse, etc.

See Marshall (2001) for eye-tracking in user modeling systems and Ikeahara *et al.* (2003) for an experimental methodology to evaluate cognitive load in adaptive information filtering.

- *Selecting the subjects*. Even if the *population*[4] of the experiment is a limited population, it is often difficult if not impossible to collect data from all members of that population. Psychologists used 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 from a larger population to make inferences about that population. The best way to have a representative sample is to *randomly select the subjects* from the target population. In this way every member of the population has an equal probability of being selected. Random sampling ensures the greatest *external validity*, which can be defined as the ability to make generalizations about a population from a sample.
  However, much research in psychology is based on samples obtained through non-random selection processes (e.g. *availability sampling*), since most researchers are more concerned with the study of the relationship existing between variables in a controlled situation. In particular, what researchers are concerned about is whether the results of the experiments can be duplicated or replicated by themselves or the others. This is another meaning of the concept of external validity, namely, whether the results of a particular independent variable generalize to different samples of subjects who serve in essentially the same experiment.
  It is important to notice that in this case, in order to detect if there are differences caused by the manipulation of the independent variable(s), the experiment needs as many subjects as are necessary to provide a relatively sensitive test of the research hypothesis. One way to increase the sensitivity in an experiment is to increase the number of subjects.
- *Conduct the experiment*, which consists primarily of collecting data using a particular experimental design.[5] In this phase subjects are assigned to the different treatment conditions. In the

[4] The *population* is here defined as the set of objects that a study investigates, using samples of the population in an experiment.
[5] In statistics books, *experimental design* refers to a general plan for conducting an experiment. The most common design for an experiment in which two or more independent variables are manipulated is called the factorial design.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

simplest procedure, the *between-subjects design*, an *experimental group* of subjects is assigned to the treatment, while another group of subjects, the *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 the user model (experimental group) and another group without it (control group). Usually, this procedure is aimed at detecting differences between the two experimental conditions (e.g. is the application designed with a user model more successful than the one without?). For more details, see Chin (2001), while for an example, see Gena and Ardissono (2004). At the other extreme of between-subjects design is the *within-subjects design* in which each subject serves in all of the treatment conditions, e.g. subjects completing tasks using both the application with UM and the one without. For example, see Paris *et al.* (2001) and Bontcheva (2002).

When two or more independent variables (treatment conditions) are manipulated at the same time, we are dealing with *factorial experiments*. 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 treatments result from the combination of two independent variables. For example, Müller *et al.* (2001), in an experimental condition, simulated a crowded airport terminal where the user asks a mobile assistant system questions via speech. In that situation the two independent variables manipulated were: the navigation into the system and the time pressure of the user.

There may be more than two groups, of course, depending on the number of independent variables and the number of levels each variable can assume. 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*). For example, see Gena (2002, 2003).

- *Use descriptive statistics to describe the data.* These statistics (such as mean, variance, standard deviation) are designed to describe or summarize a set of data. In order to report significant results and make inferences, the descriptive statistics are not sufficient and some inferential statistics measure is required.
- *Use inferential statistics to evaluate the statistical hypothesis.* These statistics are designed to make inferences about larger populations and to assess the significance level of an experiment, which is the probability (p) with which the experimenter is willing to reject the null hypothesis when it is actually correct.[6] The accepted significance levels are 0.05 ($p=5\%$, 'significant') and 0.01 ($p=1\%$, 'very significant').

  The main inferential statistics used in experimental evaluation are as follows.

  – *ANOVA*, the analysis of variance, which is used to determine whether differences in dependent variables are due to the different independent variable treatments or due to chance factors. Therefore, a ratio is calculated, the *treatment index*, dividing the between-groups variability[7] by the within-groups variability:[8]

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

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

[6] The *null hypothesis* is the statistical hypothesis evaluated by hypothesis testing. Usually represented as the absence of a relationship in the population.

[7] 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.

[8] The *within-group variability* takes into consideration the variation of subjects treated alike.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

If the obtained value is close to 1 there are no treatment effects (and therefore the null hypothesis is true); otherwise if the obtained value is greater than 1 there can be treatment effects. In the latter case, we have to know how much the treatment index must be greater than 1 in order to reject the null hypothesis and to assume an alternative hypothesis[9] to be true. The treatment index, also known as the *F ratio*, can be calculated using the statistical procedure called the *F test*. If the obtained value of *F* is equal or exceeds the *critical value*, the null hypothesis is rejected, otherwise it is retained. The critical value of *F* is calculated using the *table of critical values*, constructed for all the possible combinations of *a* (the number of treatment conditions) and *n* (the number of participants) in a relatively small space (see Keppel, 1991; Keppel *et al*., 1998). For examples of the *F* test, see Müller *et al*. (2001) and Brunstein *et al*. (2002).

Another test that analyzes the results of two-group studies and the differences between two means is the *t test*.[10] The *t* test is algebraically equivalent to the *F* test. See, for instance, Martin and Mitrovic (2002) and Santos *et al*. (2003).

The *chi-square test* (χ^2) is the common measure used to evaluate the significant values assumed by *categorical data*[11] (while measures above can be applied when the dependent variables to measure are *continuous*: they can take values such as time or number of errors, etc.). In the chi-square test the observed frequencies with which different classes of response occur are compared with expected frequencies derived from theoretical or empirical considerations. For an example of the chi-square test, see Kurhila *et al*. (2002) and Pirolli and Fu (2003).

– *MANOVA*, the multivariate analysis of variance, is a technique used for assessing group differences across multiple metric dependent variables simultaneously, based on a set of categorical variables acting as independent variables. See, for example, Weibelzahl and Weber (2002).
– The *sensitivity measures*. The sensitivity of an experiment is given by the effect size and the power. The *effect size or treatment magnitude* (ω^2) measures the strength, or the magnitude, of the treatment effects in an experiment. It gives the magnitude of the change in the dependent variable values due to changes in the independent variable as a percentage of the total variability. In social sciences *small, medium, large* effects of ω^2 are, respectively, 0.01/0.06/>0.15. For example, see Gena and Ardissono (2004). The *power of an experiment* 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 (using, for instance, within-subjects designs or analysis of covariance).[12] The power can be used for estimating the sample size. In social sciences, the accepted value of the power is equal to 0.80, which means that 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 it more likely to obtain the desired effects. 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).

[9] The *alternative hypothesis is* the hypothesis that is accepted when the null hypothesis is rejected.
[10] The *t* test is a special case of the *F* test because the two tests are algebraically equivalent: $F=(t)^2$ and $t=\sqrt{F}$.
[11] 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) that cannot be ordered.
[12] The *covariance* is the measure of the degree to which deviations vary together or co-vary. If the deviations in the *X* variable tend to be of the same size as the corresponding deviations in 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, and education, learning styles, previous experience, prior knowledge and aptitudes.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

  - The *non-parametric statistics*, which are statistics that make no assumptions about the distribution of the scores making up a treatment condition. 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 properties such as means. Examples of non-parametric tests are the Mann–Whitney test, Kruskal–Wallis test, Wilcoxon test, and Friedman test. For example, see Dimitrova *et al.* (2001) and Jamenson and Schwarzkopf (2002).
  - The *post-hoc tests*, which are measurement conducted after the data have been examined and are used to limit errors, such as the *type I* (*or α error*), which happens when the null hypothesis is rejected even though it is true (wrongly assumed treatment effect). The most widely used posthoc test in psychology and the behavioural sciences is Tukey's Honestly Significant Difference (Tukey's HSD) test.
- *Draw conclusion regarding the research hypothesis.*
- *Prepare a formal report for publication or presentation.*

For more details on empirical evaluation and statistics, see Keppel (1991) and Keppel *et al.* (1998).

As emphasized by Chin in his survey concerning the empirical evaluation of UM and user-adaptive systems (2001), 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 factors. In an ideal experiment, only the independent variable should vary from condition to condition. In 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 becomes hard to distinguish between the effects of the manipulated variable and the effects due to confounding variables. The presence of confounding variables usually ruins the 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 the 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 time, locations, or other environmental conditions can influence the dependent variables: there may be more distracting noise at certain times; the computer may be slower at certain times; the experimenter may bias the participants by words, body language, appearance and so on. One way to control the potential sources of confounding variables holding them constant, so that they have the same influence on each of the treatment conditions. Unfortunately, not all the potential variables can be handled in this way (for instance, reading speed, intelligence, etc.). For the other remaining nuisance variables, their effects can be neutralized by randomly assigning[13] subjects to the different treatment conditions.

At the end of his overview, Chin proposed that authors publishing in UMUAI should report the following common measures:

- the number, the source and relevant background of participants;
- the independent, dependent and covariant variables;
- the analysis method;
- the post-hoc probabilities;
- the raw data (in a table or appendix) if not too voluminous;
- the effect size and the power (which should be at least 0.8);
- the non-significant results (with corresponding effect size and power).

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

- the data can be contaminated;
- there can be unwarranted assumptions about scales for variables;

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

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

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

4.1.1 How to evaluate the success of an experiment?

The statistics described above are often reinforced by other measures and qualitative considerations in order to correctly evaluate the results obtained by an experiment aimed at assessing the success of an adaptive application. For instance, users' suggestions can be taken into consideration in order to discover the effects generated by the exploitation of adaptive techniques such as reduction of the complexity of the interaction, deeper knowledge of the system functionality, increased satisfaction, reduction of the interaction anxiety, and so on.

Höök (1997) argued that the evaluation of an interactive system is always a hard task, and when the system is adaptive the task becomes even harder because the evaluation must distinguish the adaptive features from the general usability problems. Since most adaptive system evaluations are comparisons between the system with and without adaptations, the problem is clear: most of the time 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 into 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 the relevance of the information provided is more important than the time spent to find it. Furthermore, lots of applications are designed for a long-term interaction and therefore it is hard to correctly evaluate them in a short and controlled test.

Weibelzahl and Lauer (2001) assumed that adaptivity reduces the complexity of the interaction and therefore they propose evaluating the user's behavior using measures of complexity to demonstrate the complexity reduction. They called this new criterion *behavioral complexity*. For instance, the number of clicks to reach a goal can be used as a measure, since an easier interface provides shorter paths to the goals. However, in addition 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. However, as Herder (2003) noted, the browsing activity is expected to produce more complex navigation than goal directed interaction. Therefore, the metrics for the evaluation of user navigation should take into account both the site structure and the kind of user tasks since, depending on these factors, a reduction in the complexity of the interaction is not necessarily caused by the adaptive behavior of the system.

Weibelzahl and Weber (2002) suggested that empirical studies are very good at identifying design errors and wrong assumptions but they do not suggest new theories or approaches directly. Thus, empirical evaluations have to be combined with theoretical grounds to yield fruitful results (see qualitative evaluation, Section 5).

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

- the percentage of repeated visits to the Web site;
- the average length of the visit;
- the average time spent to read a page;
- the conversion ratio (referred to as the percentage of users who become buyers);
- the average user expense;

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

- the ROI (return of investments);
- the usability;
- user satisfaction.

Del Missier and Ricci (2003) suggested adopting a context-matching approach (Payne *et al*., 1999) in the experimental evaluation, in order to cope with the contingencies of user behavior. For recommender systems, for instance, this means that it will be necessary to reproduce the real decision environment and the real system should be tested, with no changes in databases, interface, algorithms and parameters. Furthermore, to assure external validity, also the evaluation setting and the sample of users should be selected to be representative (see discussion in Section 4.1). In this way, it will be possible to manipulate only a few relevant factors in order to understand their impact on the user–system behavior in the real decision environment.

Herder (2003) proposed a utility-based evaluation approach based on Brusilovsky's layers (see Section 4.2). The interesting point here is that the evaluation can be seen as a utility function U that maps a system, *given some user context*, to a quantitative representation of user satisfaction and performance. However, this function is highly dependent on the criteria employed: for different domains, different criteria can be thought of (learning rate for educational hypermedia, increase of sales for e-commerce systems, etc.).

4.2 The layered evaluation

As introduced in Section 1, layered evaluation (Totterdell & Boyle, 1990) is one of the peculiarities that differentiates the evaluation of user-adaptive systems from the evaluation of regular systems. Different frameworks have been proposed in the literature with different levels of granularity (layers). The key point of all these approaches is that, depending on the different layers identified in the adaptation process, different evaluations and analyses have to be taken into account.

Karagiannidis and Sampson, (2000) and Brusilovsky *et al.* (2001) distinguished two high-level phases characterizing the adaptive systems and the two corresponding separated evaluations.

- *The interaction assessment phase*, where the assessment process is being evaluated. Here the question can be stated as: 'are the conclusions drawn by the system concerning the characteristics of the user–computer interaction valid?' or 'are the user's characteristics being successfully detected by the system and maintained in the user model?'
- *The adaptation decision-making phase*, where the adaptation decision-making is being evaluated. Here the question can be stated as 'are the adaptations decisions valid and meaningful for selected assessment results?' or 'is the selected adaptive presentation technique appropriate for the given user goals?'

They suggested that 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, it is not evident whether one or both of the above phases have been unsuccessful. For instance, the conclusion of the interaction assessment phase should be correct, but the adaptive choices should not be meaningful, or the adaptation decision should be reasonable but based on incorrect assessment results.

The effectiveness of such an approach has been demonstrated by the experimental results in adaptive educational systems. A set of experiments reported contradictory results about the effects of adaptive link annotation. Different evaluations (Eklund & Brusilovsky, 1998; Brusilovsky & Eklund, 1998; Weber & Specht, 1997) reported that even if students seem to understand and like adaptive navigation support features, it did not influence their performance on test. Afterwards, a re-analysis (Specht & Kobsa, 1999) of data from previous experiments by Weber & Specht (1997) has 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 methods such as enabling/disabling links or direct

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

guidance with the adaptive 'next' link. Therefore, these results demonstrated that, in order to choose the correct interface adaptation, the previous knowledge of the students has to be taken into account. Indeed, the more knowledge the students have on a subject (interaction assessment layer), the bigger the improvement they gain using the adaptive annotation (adaptation decision-making layer), while less experienced students obtain the best results using the direct guidance.

Jameson (1999), in his overview of types of empirical studies of adaptive systems, 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 require controlled evaluations with users and experience with real-world use. In addition, all the empirical studies should address questions concerning the following.

- *The correctness of assumptions about users relied on inferences techniques*: can the general assumptions about users be shown empirically to be correct?
- *The appropriateness of inference techniques used*: are the techniques used well suited to dealing with the inference tasks faced by the system?
- *The adequacy of available data*: is there typically enough data available about each user to enable the system to make useful inferences about him or her?
- *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?
- *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?

As it can be noted, this is not a layered framework, however the questions addressed clarify how to detected differences in the evaluation of the components of an adaptive system.

Paramythis *et al.* (2001) suggested a modular approach to the evaluation of adaptive user interface. They use a high-level model of adaptation made up of the following components:

- interaction monitoring;
- interpretation/inferences;
- explicitly provided knowledge;
- modeling;
- adaptation decision-making;
- applying adaptations, transparent models and adaptation 'rationale';
- automatic adaptation assessment.

Then, they identified seven evaluation modules (composed of one or more of the adaptation components listed above), which can be evaluated individually and in combination on the basis of the evaluation goals.

- *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 such as focus group, questionnaires, interviews, think aloud protocol, logging use and expert-based evaluation can be used.
- *Module A2* comprises *explicit provided knowledge* and *modeling* and can be used for the same purposes as module A1 and also for evaluating the *transparency* of the process and the possible *overhead* imposed to the user. The main method used is the expert-based evaluation.
- *Module B* comprises *adaptation decision-making* and can be used to evaluate the *necessity*, the *appropriateness*, and the *acceptance* of adaptation. The suggested methods are expert-based evaluation and user involvement.
- *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.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

- *Module D1* comprises *modeling* and *transparent models* and can be used to evaluate the *completeness*, the *coherence*, and the *rationality* of the presentation. End users and experts can be involved in the assessment of the model together with testing of real interactions.
- *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 as for module D1.
- *Module E* comprises *automatic adaptation assessment* and its goal is to ensure that the system shares the same views as the users as regards the success or the failure of adaptations. The feasibility of this approach still has to be investigated.

The possible combinations of modules are the following.

- *Modules A1, A2 and D* modules capture the entire process of constructing models and presenting these models to the users.
- *Modules B and C* capture the process of deciding upon and applying adaptations.
- *Modules C and D2* capture, for examples, how predictable and how controllable the adaptive interface is.
- *Modules A1, A2, B and C* capture the entire traditional adaptation cycle of an adaptive system.

It is important to observe that, in their model, Paramythis *et al.* not only detected different adaptation components and their combination for the evaluation goals, but they also suggested the proper evaluation techniques suitable for those goals.

In their approach they further clarified the following points:

- since the concept of the typical user of a system cannot be applied in adaptive systems, 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);
- the expert evaluations are conducted here by domain experts instead of usability experts as in the usual HCI evaluations.

Similarly, Weibelzahl (Weibelzahl, 2001), proposed an evaluation framework made up of six steps.

- Evaluation of reliability and external validity of input data acquisition used to build the user model.
- Evaluation of the correctness of the inference mechanism and accuracy of user properties.
- 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.
- Change of system behavior when the system adapts (in which way does system behavior change in comparison to the normal division of labor?).
- Change of user behavior when the system adapts (do users change their behavior when the system adapts in comparison to the normal division of labor? In which way?).
- 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 completed already (this is especially important in the case of finding no difference between an adaptive and a nonadaptive system).

Finally, Weibelzahl *et al.* (2003) proposed a merged version of their frameworks (Weibelzahl, 2001; Paramythis *et al.*, 2001). The new framework is based on a new model of the adaptation process based on the 'approximate model' of user action introduced by Norman (1988). Particularly, (i) they introduced the separation between the observation/collection of external stimuli and their interpretation; (ii) they distinguished the high-level goal, or intention to act, from the actual translation of that intention into a series of concrete actions; (iii) they implemented a

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

distinction between modeling and drawing decisions on the basis of the resulting models. The new model for decomposing adaptation is made up of six stages. For each stage they suggested some methods that might be feasible for the evaluation of each layer.

- *Collection of input data*: might be evaluated in terms of reliability, accuracy, precision, latency, sampling rate, etc. In summary, this stage evaluates the technical quality of input data.
- *Interpretation of the collected data*: this stage evaluates the validity of the interpreted input data.
- *Modeling of the current state of the world*: this stage compares the system-maintained dynamics model with the real-world entities to which they correspond.
- *Deciding upon adaptation:* here the downward inference from abstract user-context properties to concrete adaptation decision is evaluated.
- *Applying the adaptation decision*: this stage evaluates the physical level of adaptation.
- *Evaluating adaptation as a whole*: they suggested evaluating both system and user under real world conditions.

4.3 A corpus of study

The recent conferences and publications in the field of user modeling and adaptive systems 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 principles and guidelines. The key point is to carry out evaluations leading to significant results that can be re-used in other research, and promote the development of standard measures that would be able to reasonably evaluate the systems' reliability. If many studies show that certain adaptations work, general principles can be extracted. Some reference guidelines have already emerged and can be taken into account in the development of adaptive systems. Therefore, when a researcher decides to evaluate an (adaptive) system, results of previous evaluation should be taken into account.

For instance, Sears and Shneiderman's (1994) evaluation on menu voices sorting reported that the users were disoriented by the menu voices sorted on usage frequency because of the lack of significance in the adapted menu. 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). Due to this evidence, researchers should be careful in applying this technique (see also Debevc *et al*., 1996).

To assist researchers in the evaluation of adaptive systems and to promote the construction of a corpus of guidelines, Weibelzahl and Weber (2001) proposed the development of an online database[14] for studies of empirical evaluations of adaptive systems called Easy-D. The aim of this proposal is to serve as a reference for researchers in the field of adaptive systems and to guide the planning of new evaluations that fulfill the following methodological requirements.

- *Evaluation layer*: according to the framework proposed in Weibelzahl and 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 takes place?
- *Method of evaluation*: a short description of the evaluation, using one 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,

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

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

    * controlled evaluations with hypothetical (i.e. simulated) users,
    * experience with real world use.
- *Data type*: brief description of the kind of data analyzed.
- *Criteria*: what were the main criteria and what measures were used in study?
- *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)?

5 The qualitative evaluation

Qualitative methods of evaluation are seldom applied in the assessment of user-adaptive systems compared with the methods described above. However, qualitative approaches are related to the real usage of an application, and in real conditions. Therefore, they can offer significant results both in terms of system evaluation and in terms of requirement gathering.

The distinction between quantitative and qualitative research originates from the social sciences. An immediate and meaningful definition of qualitative research comes from Anselm Strauss, one of developers of the Grounded Theory. For Strauss (Strauss & 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.'

Qualitative methods of research often make use of ethnographic investigations, also known as participant observation. In social sciences, and in particular in 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 researchers immerses themselves in a new world with the goal of exploring the members' point of view.

The use of ethnographic materials is common in interactive systems evaluation. Supporters of socially based studies have found that ethnographic approaches can uncover requirements for system design through the detailed observation of the work settings. In contrast, more traditional approaches, such as laboratory-based studies, tend to be disconnected from the lived detail of the work. From an ethnographic perspective, the quantitative methods, such as controlled experiments and usability tests, 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 software system features are part of a set of working practices. On the opposite side, sustainers of quantitative methods of research observed that also in the case of field studies the situation is not completely natural since the subjects are likely to be influenced by the presence of evaluators and/or recording equipment.

On a more analytic level, ethnographic methods are used to analyze work processes and work practices (Dourish, 2001). Work processes are formalized or regularized procedures by which work

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

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.

Preece *et al.* (1994) classified the ethnographic investigations under the umbrella term of 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 by observing users in natural settings can one detect the presence of these factors. For Walsham (1993) 'interpretatism is concerned with approaches to the understanding of reality and asserting that all such knowledge is necessary a social construction and thus subjective'. An interpretative evaluation can be valuable at various stages in the development cycle but particularly for feasibility study, design feedback or post-implementation review.

Interpretative evaluation comes in the following forms (Preece *et al.*, 1994).

- *Contextual inquiry*, which is a semi-structured interview covering whatever interesting aspects are 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 & Beyer, 1998). Particularly, the evaluator is interested in:
  - structure and language used in the work;
  - individual and group actions and intentions;
  - the culture affecting the work;
  - explicit and implicit aspects of the work.
- *Cooperative evaluation*, which includes methods wherein 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 Section 2), which allows the user to ask questions, comments and suggest appropriate alternatives for the evaluators, and for the evaluators to prompt the user (Monk *et al.*, 1993).
- *Participative evaluation*, which 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 groups and prototypes (see Section 2 and Greenbaum & Kyng, 1991).
- *Ethnography*, which is concerned with collecting data about the real work situation. The ethnographic approach in HCI acknowledges the importance of learning more about the way technology is used *in situ*. This implies (Monk *et al.*, 1993):
  - the need to use a range of methods including intensive observation, in-depth interviewing, participation in cultural activities and simply hanging about, watching and learning as events unfold;
  - the holistic perspective, in which everything—belief systems, rituals, institution, artifacts, texts, etc.—is grist for the analytical mill; and immersion in the field situation.

Different kinds of data sources may be collected as part of this practice, including video, annotations in notebooks, snapshots, etc., while video analysis support tools and databases provide a valuable way of categorizing materials.

In the analysis of everyday life, the theory of ethnomethodology by Harold Garfinkel can finally be collocated (Garfinkel, 1967). Preece *et al.* (1994) defines ethnomethodology as a method that

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

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 carrying out everyday 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 breaking 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. The disoriented reactions of the involved subjects confirmed the existence of such rules.

Ethomethodology has been used to study how new technologies have been introduced into various work settings. The interest is in the details of the work of technology as a social achievement and in the social practice through which participants recognize themselves to the technology in course of its design, construction, development and use and in talking and writing about it (see, for example, Dourish, 2001).

5.1 The Grounded Theory

The Grounded Theory is a qualitative research methodology developed by Anselm Corbin 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 & Corbin, 1998, p. 12).'

In the introduction of their book '*Basics of Qualitative Research*' (Strauss & Corbin, 1998), the authors outline the characteristics of grounded theorists. They must have:

- the ability to step back and critically analyze situations;
- the ability to recognize the tendency toward bias;
- the ability to think abstractly;
- the ability to be flexible and open to helpful criticisms;
- sensitivity to the words and actions of respondents;
- 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 researchers do not develop a new way of thinking about data in the world where they live. 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 & Corbin, 1998, p. 8) with the final goal of 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 a 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:

- data (interviews, observations, documents, records, films, etc.);
- 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 questions and making comparisons;
- written and verbal reports.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

In particular, the second point is more developed and further divided into three important steps that make up the core component of the Grounded Theory:

- *open coding*: the analytic process through which concepts are identified and their properties and dimensions are discovered;
- *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;
- *selective coding*: the process of integrating and refining the theory.

5.1.1 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 and user-adaptive systems, qualitative methods of evaluation are seldom exploited. Concerning Grounded Theory, an evaluation study applying this methodology 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).

According to the authors, 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 the following:

- video of learners using the application;
- interview with learners;
- pre-test and post-test results comparison;
- data logging of user navigation and online tests and tasks;
- questionnaires related to user attitude;
- tasks and questions results;
- focus group studies carried out with two small groups of users;
- staff evaluation of the course;
- staff diaries;
- interviews with staff involved;
- a formal staff report of the experience;
- 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 the following:

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

- *the student model*: performance results, components of the student model, the cooperative student model;
- *the learning materials*: subject content, design features, usability, learning presentation strategy;
- *the management of learning*: the tutor, the learning environment.

While some examples of variables identified in subcategories were as follows:

- *performance results*: pre-test, post-test, on/off computer tests, on/off computer tasks;
- *usability*: ease of use;
- *the tutor*: involvement.

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

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

6 Conclusions

This paper presents a review of the methodologies and the related examples for the evaluation of useradaptive systems. Particular emphasis has been given to the HCI methodologies (see Sections 2, 4.1 and 5) by describing all the techniques that can be applied. Another survey (Chin, 2001) concentrates on empirical evaluation (controlled experiments, see Section 4.1), which is largely applied in HCI, and gives a detailed overview of the correct methodologies of evaluation. However, controlled experiments, and other summative methods can only be used with a running system, during the last design phases. If a researcher decides to develop the system following a user-centered approach, other evaluation techniques have to be taken into account in order to avoid expensive design mistakes and usability problems (see Section 2). Moreover, quantitative evaluation is not a unique way to test systems. Qualitative methodologies (see Section 5) are also effective, and sometimes more reliable than quantitative approaches.

The choice between quantitative and qualitative methodologies depends on the point of view of the evaluation: while quantitative research tries to explain the variance of the dependent variable(s) generated through the manipulation of independent variable(s) (variable-based), in qualitative research the object of the study becomes the individual subject (case-based). Qualitative researchers sustain that a subject cannot be reduced to a sum of variables and therefore a deeper knowledge of a smaller group of subjects is more useful than an empirical experiment with a representative sample. Even if the final goals of both approaches are similar (they both want to come up with predictive theories to generalize over individual behaviors), they are carried out in a different way: while quantitative researchers try to explain the cause–effect relationships between variables and make generalizations on the obtained results (*extensive approach*), qualitative researchers want to comprehend the subjects under study by interpreting their points of view and by analyzing the facts in depth (*intensive approach*) in order to propose a new general understanding of the reality.

Indeed, the choice between quantitative and qualitative methodologies is not trivial and depends on the aims and the purpose of the evaluation. If we would like to test the impact of different adaptation techniques in a given interface, for example, a controlled experiment can produce useful results. For instance, I tested the impact of different adaptation techniques applied to a Web site using a controlled experiment (Gena, 2002). In contrast, to discover the features of the interaction useful in modeling the user, observing users in context and interviewing them can provide material to build the user model categories. For instance, in a project concerning the design of a Web site offering technical information at different levels (depending on the different backgrounds of the users), we collaborated with the domain experts and final users in an iterative design–test–redesign

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

process to define the different levels of interaction for the given users (e.g. expert, non-expert, etc.) (Gena & Ardissono, 2004).

To sum up, if we want to discover new theories, a qualitative approach can produce more fruitful results, while if we want to investigate the relationship between variables, or if we want to measure the accuracy of the selection process, the quantitative approach may be preferred. However, this is not a rule and both approaches can offer interesting results.

Considering the state of the art, even though an improvement has been registered in the number of evaluation studies in the last few years, the evaluation of user modeling and adaptive systems needs to reach more rigorous levels in terms of subject sampling, statistical analysis, correctness in procedures, experiment settings, etc. Moreover, evaluation studies should benefit from the application of qualitative methods of research and from a rigorous and complete application of a user-centered design approach in every development phase of these systems. Last but not least, user-adaptive systems always need a layered evaluation that differentiates, at least, problems concerning content adaptation and interface adaptation. Layered approaches have been discussed in detail in this paper (see Section 4.2) and they probably represent one of the peculiarities in the evaluation of user-adaptive systems. However, these approaches have the drawback of proposing evaluation levels without specifying the techniques that have to be used (excluding the framework of Paramythis *et al.*, 2001). In particular, the latest proposals on layered evaluation (Weibelzahl *et al.*, 2003) add new layers of analysis without offering practical suggestions to the researchers who want to discover correct methodologies of testing.

In order to offer some practical suggestions to the researchers interested in user-adaptive system evaluation, Table 1 summarizes all the methods and techniques from this paper by considering the following.

- The basic layered approach that distinguishes content and interface adaptations. In particular, for every evaluation technique a suggestion is made as to which layer it can be used.
- The kind of factors they are most suited to evaluate.
- At which moment of the system development they can be used. In particular, I distinguish three phases:
  - *requirement phase*, which occurs before any system implementation; during this phase it may be useful to gather data about typical users (their features, their behavior, their actions, their needs, etc.), the application domain, the system features, etc.;
  - *preliminary evaluation phase*, which occurs during the system development; it is very important to carry out one or more evaluations during this phase to avoid expensive and complex re-design of the system once it is finished;
  - *final evaluation phase*, which occurs at the end of the system development and it is aimed at the evaluation of the overall functioning of the system.
- The applicability of the evaluation methodologies. This dimension underlines if there are constraints or particular conditions necessary to utilize the methodology.

It is important to point out again that the evaluation of a user-adaptive system is important not only for testing, as in the regular HCI applications, but also as a knowledge source for the development of the adaptive application. As outlined in Section 2.1, most evaluation techniques can be considered as generative models since they can be a source of information for the knowledge base of an adaptive system.

Moreover, the results of the evaluation can have an impact on different features of the system. For instance, if the adaptation of the contents shows bad results the problem can involve the rules of adaptation, the domain knowledge, the user model, etc. If the interface adaptations do not help all the users in task completions the problem can involve bad interface design, incorrect classification of users, instability of the user model, etc.

Thus, the analysis of collected results is a very important stage, in which the layered evaluation can help: the different facets of the system are linked to the evaluation results in order to discover the possible pitfalls.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Table 1 Summary of the proposed methodologies

	Content	Interface	Kind of factors	Phases of the evaluation	Applicability conditions
Interviews	x	x	User opinion, user satisfaction	Requirement phase, preliminary and final evaluation phase	
Questionnaires	x	x	User data, user preferences and opinions	Requirement phase, preliminary and final evaluation phase	
Post-test questionnaires	x	x	User opinion and user satisfaction after a test	Requirement phase, preliminary and final evaluation phase	After a test session of an adaptive system
Pre- and post-test questionnaires	x	x	Improvement in user knowledge	Final evaluation phase	Student-adaptive systems
On-line questionnaires	x	x	User opinions, user satisfaction, user preferences	Final evaluation phase	Web-based adaptive applications
Focus group	x	x	Qualitative data gathered from target users	Requirement phase	
Systematic observation	x	x	Sequences of real user actions	Requirement phase	A running user-adaptive system or similar existing systems
Think out loud protocol		x	Usability of interface adaptations	Preliminary evaluation phase	A running user-adaptive prototype/system
Logging use	x	x	Real usage data, data for simulation	Final evaluation phase	A running user-adaptive system
Heuristic evaluation		x	Usability of interface adaptations	Preliminary evaluation phase	A running user-adaptive prototype/system
Expert review	x	x	User, domain and interface knowledge	Preliminary evaluation phase	
Parallel design		x	Different adaptive interface solutions	Preliminary evaluation phase	
Cognitive walkthrough		x	Usability of interface adaptations	Preliminary evaluation phase	A running user-adaptive prototype/system
Wizard of Oz simulation	x	x	Early prototype evaluation	Preliminary evaluation phase	
Scenario-based design	x	x	Evaluation before implementation	Preliminary evaluation phase	
Prototypes	x	x	Evaluation of vertical or horizontal prototype	Preliminary evaluation phase	A running user-adaptive prototype
Usability test		x	Usability of interface adaptations	Final evaluation phase	A running user-adaptive prototype/system
Experimental evaluation	x	x	Interface (and content) adaptations	Final evaluation phase	A running user-adaptive prototype/system

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Table 1 *Continued*

	Content	Interface	Kind of factors	Phases of the evaluation	Applicability conditions
Task analysis			Real user actions	Requirement phase	A running user-adaptive system or similar existing systems
Cognitive and socio-technical models			Data to model the user, to establish stakeholder requirements, . . .	Requirement phase	A running user-adaptive system; representative users collaboration
Precision and recall	x		Accuracy of recommendations	Preliminary and final evaluation phase	Dichotomic classification of user-relevant content
Training set and test set	x		Generate recommendations—evaluate the accuracy of generated recommendations	Preliminary and final evaluation phase	Learning-based adaptive systems
Coverage	x		Percentage of correct recommendation	Preliminary and final evaluation phase	Classification of user-relevant content
MAE and RMSE, correlation, regression	x		Statistical recommendation accuracy	Preliminary and final evaluation phase	Contents classified in a range of values
Reversal rate	x		Decision support accuracy	Preliminary and final evaluation phase	Dichotomic classification of user-relevant content
ROC sensitivity	x		Decision support accuracy	Preliminary and final evaluation phase	Dichotomic classification of user-relevant content
PRC sensitivity	x		Decision support accuracy	Preliminary and final evaluation phase	Dichotomic classification of user-relevant content
Utility metrics	x		False positive, false negative	Preliminary and final evaluation phase	Dichotomic classification of user-relevant content
Simulation	x		To test different algorithms and solutions	Preliminary evaluation	Log files, usage data
Contextual inquiry	x	x	Focus on real user actions	Requirement phase	A running user-adaptive system
Cooperative evaluation	x	x	Collaboration with real users during final evaluation step	Final evaluation phase	Final user collaborating during the design-redesign phase
Participative evaluation	x	x	Real users collaboration during the system implementation	Requirement phase and preliminary evaluation phase	Final user participating during the system implementation
Ethnography	x	x	Collection of data in real work situation	Requirement phase and final evaluation phase	A running user-adaptive system or similar systems
Grounded theory	x	x	To combine qualitative and quantitative evaluation, to discover new theories	Preliminary and final evaluation phase	A running user-adaptive prototype/system

To conclude, I advocate the importance of evaluation and testing in every design phase of a useradaptive system and at different layers of analysis. Significant testing results can lead to more appropriate and successful systems. From my point of view, both quantitative and qualitative methodologies of research can offer fruitful contributions and their correct application has to be carried out by the researchers working in this area.

Acknowledgements

I want to thank the anonymous reviewers and the editor, Simon Parsons, of *The Knowledge Engineering Review*, who helped me considerably to improve the style and the content of this article. I am also very grateful to Roberto Albano who helped me to revise the statistical part of the paper, Roberto Esposito for the suggestions concerning the machine learning metrics and Silvia Likavec for proofreading. Last but not least, I am very grateful to Luca Console for his advice and his careful reading of the paper.

References

Ardissono, L, Gena, C, Torasso, P, Bellifemine, F, Chiarotto, A, Difino, A and Negro, B, 2004, User modeling and recommendation techniques for personalized electronic program guides in L Ardissono, A Kobsa and M Maybury (eds) *Personalization and User-adaptive Interaction in Digital TV*. Dordrecht: Kluwer Academic Publishers, 30–26.

Alfonseca, E and Rodríguezi, P, 2003, Modelling users' interests and needs for an adaptive on-line information system in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 76–82.

Bakeman, R and Gottman, JM, 1986, *Observing Behavior: An Introduction to Sequential Analysis*. Cambridge: Cambridge University.

Barker, T, Jones, S, Britton, C and Messer, D, 2002, The use of a co-operative student model of learner characteristics to configure a multimedia application. *User Modeling and User-adaptive Interaction* **12**(2), 207–241.

Baudisch, P and Brueckner, L, 2002, TV Scout: lowering the entry barrier to personalized TV program recommendation in P DeBra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and AdaptiveWeb-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, 58–68.

Bauer, M, Gmytrasiewicz, PJ and Vassileva, J (eds), 2001, *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer.

Beck, JE, Jia, P, Sison, J and Mostow, J, 2003a, Assessing student proficiency in a reading tutor that listens in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 323–327.

Beck, JE, Jia, P, Sison, J and Mostow, J, 2003b, Predicting student help-request behavior in an intelligent tutor for reading in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 303–312.

Bellotti, V and MacLean, A. nd *Design Space Analysis (DSA)*. http://www.mrc-cbu.cam.ac.uk/amodeus/summaries/DSAsummary.html

Bental, D, Cawsey, A, Pearson, J and Jones, R, 2003, Does adapted information help patients with cancer? in P Brusilowsky, A Corbett and F De Rosis (eds) *User Modelling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin, Springer Verlag, pp. 288–291.

Benyon, D, 1993, Adaptive systems: a solution to usability problems. *User Modeling and User-adaptive Interaction (3)*, 65–87.

Beyer, H and Holtzblatt, K, 1998, *Contextual Design: Defining Customer-Centered Systems*. San Francisco, CA: Morgan Kaufmann.

Bontcheva, K, 2002, Adaptivity, adaptability, and reading behaviour: some results from the evaluation of a dynamic hypertext system in P De Bra, P Brusilovsky and R Conejo (eds), *Adaptive Hypermedia and Adaptive Web- Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, 69–78.

Brunstein, A, Jaqueline, W, Naumann, A, Krems, J F, 2002, Learning grammar with adaptive hypertexts: Reading or searching? in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web Systems (Lecture Notes in Computer Science, 2347)*. Berlin, Springer Verlag.

Brusilowsky, P, 1996, Methods and techniques of adaptive hypermedia. *User Modelling and User Adapted Interaction* **6**(2–3), 87–129.

Brusilowsky, P, 2001, Adaptive hypermedia. *User Modeling and User-adapted Interaction* **11**(1–2), 87–110.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Brusilovsky, P, Corbett, A and De Rosis, F (eds), 2003, *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer.

Brusilovsky, P and Eklund, J, 1998, A study of user model based link annotation. *Educational Hypermedia, Journal of Universal Computer Science* **4**, 429–448.

Brusilovsky, P, Karagiannidis, C and Sampson, D, 2001, The benefits of layered evaluation of adaptive applications and services in S Weibelzahl, D Chin and G Weber (eds), *Proceedings of the First Workshop on Empirical Evaluation of Adaptive Systems, Sonthofen, Germany*, pp. 1–8.

Brusilovsky, P, Stock, O and Strapparava, C (eds), 2000, *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 1892)*. Berlin: Springer.

Bull, S, 2003, User modelling and mobile learning in P Brusilowsky, A Corbett and F De Rosis (eds) *User Modelling 2003 (Lecture Notes in Computer Science 2702)*, Berlin, Springer Verlag, pp. 383–387.

Castillo, G, Gama, J and Breda, AM, 2003, Adaptive Bayes for a student modeling prediction task based on learning styles in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 328–332.

Chin, DN, 2001, Empirical evaluation of user models and user-adaptive systems. *User Modeling and User-adaptive Interaction* **11**(1/2), 181–194.

Console, L, Gena, C and Torre, I, 2003, Evaluation of an on-vehicle adaptive tourist service in S Weibelzahl and A Paramythis (eds) *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems (at the 9th International Conference on User Modeling UM2003), Pittsburgh*, pp. 51–60.

Corbetta, P, 1999, *Metodologie e tecniche della ricerca sociale*. Bologna: Il Mulino.

De Bra, P, Brusilovsky, P and Conejo, R (eds), 2002, *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer.

Debevc, M, Meyer, B, Donlagic, D and Svecko, R, 1996, Design and evaluation of an adaptive icon toolbar. *User Modeling and User-Adapted Interactaction* **6**(1), 1–21.

Del Missier, F and Ricci, F, 2003, Understanding recommender systems: experimental evaluation challenges in S Weibelzahl and A Paramythis (eds) *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems (at the 9th International Conference on User Modeling, UM2003)*, pp. 31–40.

Dimitrova, V, 2003, Using dialogue games to maintain diagnostic interactions in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 117–121.

Dimitrova, V, Self, J and Brna, P, 2001, Applying interactive open learner models to learning technical terminology in M Bauer, PJ Gmytrasiewicz and J Vassileva (eds) *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer, pp. 148–157.

Dix, A, Finlay, J, Abowd, G and Beale, R, 1998, *Human Computer Interaction*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Dourish, P, 2001, *Where the Action is: The Foundations of Embodied Interaction*. Cambridge, MA: MIT Press.

Dumas, JS and Redish, JC, 1999, *A Practical Guide To Usability Testing*. Norwood, NJ: Ablex Publishing Corp.

Eklund, J and Brusilovsky, P, 1998, The value of adaptivity in hypermedia learning environments: a short review of empirical evidence in P Brusilovsky and P De Bra (eds) *Proceedings of Second Adaptive Hypertext and Hypermedia Workshop (at the Ninth ACM International Hypertext Conference—Hypertext '98)*, pp. 13–19.

Ericsonn, K A and Simon, H A, 1984, *Protocol analysis: Verbal reports as data*, MIT Press, Cambridge, MA.

Feinman, A and Alterman, R, 2003, Discourse analysis techniques for modeling group interaction in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 228–237.

Garfinkel, H, 1967, *Studies in Ethnomethodology*. Polity Press.

Gaudioso, E and Boticario, JG, 2002, User data management and usage model acquisition in an adaptive educational collaborative environment in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 143–152.

Gena, C, 2001, Designing TV viewer stereotypes for an electronic program guide in M Bauer, PJ Gmytrasiewicz and J Vassileva (eds) *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer, pp. 247– 276.

Gena, C, 2002, An empirical evaluation of an adaptive Web site in *Proceedings of the 2002 International Conference on Intelligent User Interfaces*. New York: ACM Press, pp. 192–193.

Gena, C, 2003, Evaluation methodologies and user involvement in user modeling and adaptive systems. PhD thesis, Università degli Studi di Torino, Italy.

Gena, C and Ardissono, L, 2004, Intelligent support to the retrieval of information about hydric resources. *Proceedings of the Adaptive Hypermedia Conference 2004, Eindhoven, The Netherlands (Lecture Notes in Computer Science)*, pp. 126–135.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Gena, C, Perna, A and Ravazzi, M, 2001, E-tool: a personalized prototype for Web based applications in D de Waard, K Brookhuis, J Moraal and A Toffetti (eds) *Human Factors in Transportation, Communication, Health, and the Workplace*. Shaker Publishing, 485–496.

Good, N, Schafer, JB, Konstan, JA, Botchers, A, Sarwar, BM, Herlocker, JL and Ricdl, J, 1999, Combining collaborative filtering with personal agents for better recommendations. *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pp. 439–446.

Goren-Bar, D, Kuflik, T, Lev, D and Shoval, P, 2001, Automating personal categorization using artificial neural networks, in M Bauer, P Gmytrasiewicz and J Vassileva (eds) *User Modelling 2001 (Lecture Notes in Computer Science 2109)*. Berlin, Springer Verlag.

Gould, JD and Lewis, C, 1983, Designing for usability—key principles and what designers think in *Human Factors in Computing Systems, CHI '83 Proceedings*. New York: ACM, pp. 50–53.

Greenbaum, TL, 1998, *The Handbook of Focus Group Research*, 2nd edn. New York: Lexington Books.

Greenbaum, J and Kyng, M (eds), 1991, *Design at Work: Cooperative Design of Computer Systems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Habieb-Mammar, H, Tarpin-Bernard, F and Prévôt, P, 2003, Adaptive presentation of multimedia interface case study: Brain Story Course in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702). Berlin: Springer, pp. 15–24.*

Hanani, U, Shapira, B and Shoval, P, 2001, Information filtering: overview of issues, research and systems. *User Modeling and User-adaptive Interaction* **11**, 203–259.

Hara, Y, Tomomune, Y and Shigemori, M, 2004 Categorization of Japanese TV viewers based on program genres they watch in L Ardissono, A Kobsa and M Maybury (eds) *Personalization and User-adaptive Interaction in Digital TV*. Dordrecht: Kluwer Academic Publishers.

Herder, E, 2003, Utility-based evaluation of adaptive systems in S Weibelzahl and A Paramythis (eds) *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems (at the 9th International Conference on User Modeling, UM2003), Pittsburgh*, pp. 25–30.

Holtzblatt and Beyer, 1998.

Höök, K, 1997 Evaluating the utility and usability of an adaptive hypermedia system in *Proceedings of 1997 International Conference on Intelligent User Interfaces*. Orlando, FL: ACM, pp. 179–186.

Höök, K, 2000, Steps to take before IUIs become real. *Journal of Interacting with Computers* **12**(4), 409–426.

Hull, DA, 1998 The TREC-7 filtering track: description and analysis. In *E. M. International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 246–254.

Ikehara, S, Chin, DN and Crosby, ME, 2003, A model for integrating an adaptive information filter utilizing biosensor data to assess cognitive LoadCurtis in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 208–212.

Jackson, T, Mathews, E, Lin, D, Olney, A and Graesser, A, 2003, Modeling student performance to enhance the pedagogy of autoTutor in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 368–372.

Jameson, A, 1999, 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.

Jameson, A and Schwarzkopf, E, 2002, Pros and cons of controllability in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and AdaptiveWeb-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 193–202.

Karagiannidis, C and Sampson, D, 2000, Layered evaluation of adaptive applications and services in P Brusilovsky, O Stock and C Strapparava (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 1892)*, pp. 343–346.

Kay, J, 1999, *UM99 User Modeling: Proceedings of the Seventh International Conference*. New York: Springer.

Keppel, G, 1991, *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, NJ: Prentice-Hall.

Keppel, G, Saufley, WH and Tokunaga, H, 1998, *Introduction to Design and Analysis, A Student's Handbook*, 2nd edn.

Kim, Y, Ok, S and Woo, Y, 2002, A case-based recommender system using implicit rating technique in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 522–526.

Kobsa, A, 1994 User modeling and user-adaptive interaction. *Proceedings of the Conference on Human Factors in Computing Systems 1994, Boston, MA*, pp. 415–416.

Kobsa, A, 2002, ICS 105: project in human-computer interaction and interfaces. Unpublished notes. http://www.ics.uci.edu/~kobsa/.

Kobsa, A, Koenemann, J and Pohl, W, 2001, Personalized hypermedia presentation techniques for improving online customer relationships. *The Knowledge Engineering Review* **16**(2), 111–155.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Koychev, I, 2002, Tracking changing user interests through prior-leaning of context in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 223–232.

Kufklik, T, Shapira, B, Elovici, Y, and Maschiach, A, 2003, Provacy preservation improvement by learning optimal profile generation rate in P Brusilowsky, A Corbett and F De Rosis (eds) *User modelling 2003 (Lecture Notes in Computer Science 2702)*. Berlin, Springer Verlag, pp. 168–177.

Kurhila, J, Miettinen, M, Nokelainen, P and Tirri, H, 2002, EDUCO—A collaborative learning environment based on social navigation in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web- Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 242–252.

Langford, J, 2005 Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research* **6**, 273–306.

Lee and Lai, 1991.

Lewis, D, 1995, Evaluating and optimizing autonomous text classification systems. *Proceedings of the 18th Annual International ACM Special Interest Group on Information Retrieval*, New York: ACM, pp. 264–254.

Lewis, C, Polson, PG, Wharton, C and Rieman, C, 1990, Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces in JC Chew and J Whiteside (eds) *Proceedings of CHI'90*. New York: ACM, pp. 235–242.

Magnini, B and Strapparava, C, 2001. Improving user modelling content-based techniques in M Bauer, P Gmytrasiewicz and J Vassileva (eds) *User Modelling 2001 (Lecture Notes in Computer Science 2109)*. Berlin, Springer Verlag.

Magoulas, GD, Chen, SY and Papanikolaou, KA, 2003, Integrating layered and heuristic evaluation for adaptive learning environments in S Weibelzahl and A Paramythis (eds) *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems (at the 9th International Conference on User Modeling, UM2003), Pittsburgh*, pp. 5–14.

Malone, T W, Grant, K R, Turbak, F A, Brobst, S A, and Cohen, M D, 1987, *Intelligent information-sharing systems*. Communications of the ACM, **30**(5) 390–402, May 1987.

Marshall, S, 2001, Eye tracking: a rich source of information for user modeling. Invited talk at UM 2001, Sonthofen, Germany in M Bauer, PJ Gmytrasiewicz and J Vassileva (eds) *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer, p. 315.

Martin, B and Mitrovic, T, 2002, WETAS: a web-based authoring system for constraint-based ITS in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 396–305.

Matsuo, Y, 2003, Word weighting based on user's browsing history in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 35–44.

Maybury, M and Brusilovsky, P (eds), 2002, The adaptive Web. Communications of the ACM, 45.

McCreath, E and Kay, J, 2003, IEMS: helping users manage email in P. Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 263–272.

McNee, SM, Lam, SK, Konstan, JA and Riedl, J, 2003, Interfaces for eliciting new user preferences in recommender systems in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 178–187.

Mitchell, T, 1997, *Machine Learning*. New York: McGraw-Hill.

Mitrovic, A, 2001, Investigating students' self-assessment skills in M Bauer, PJ Gmytrasiewicz and J Vassileva (eds) *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer, pp. 247–250.

Mitrovic, A, Koedinger, K and Martin, B, 2003, A comparative analysis of cognitive tutoring and constraint-based modeling in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 313–322.

Monk, A, Wright, P, Haber, J and Davenport, L, 1993, *Improving Your Human Computer Interface: A Practical Approach*. BCS Practitioner Series. Hemel Hempstead: Prentice-Hall International.

Müller, C, Großmann-Hutter, B, Jameson, A, Rummer, R and Wittig, F, 2001, Recognizing time pressure and cognitive load on the basis of speech: an experimental study in M Bauer, PJ Gmytrasiewicz and J Vassileva (eds) *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer, pp. 24–33.

Ng, MH, Hall, W, Maier, P and Armstrong, R, 2002, Using effective reading speed to integrate adaptivity into Webbased learning in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 428–431.

Nielsen, J, 1993, *Usability Engineering*. Boston, MA: Academic Press.

Nielsen, J and Molich, R, 1990, Heuristic evaluation of user interfaces in *Proceedings of CHI'90, Seattle, Washington*, pp. 249–256.

Norman, D, 1998, *The Psychology of Everyday Things*. New York: Basic Books.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Norman, DA and Draper, S.W, 1986, *User Centered System Design: New Perspective on HCI*. Hillsdale NJ: Lawrence Erlbaum Associates.

Oppermann, R, 1994, Adaptively supported adaptivity. *International Journal of Human–Computer Studies* **40**, 455– 472.

Ortigosa, A and Carro, R M, 2003, The continuous empirical evaluation approach: Evaluating adaptive web-based courses in P Brusilowsky, A Corbett and F De Rosis (eds) *User modelling 2003 (Lecture Notes in Computer Science 2702)*. Berlin, Springer Verlag, pp. 163–167.

Paramythis, A, Totter, A and Stephanidis, C, 2001, A modular approach to the evaluation of adaptive user interfaces in S Weibelzahl, D Chin and G Weber (eds) *Proceedings of the First Workshop on Empirical Evaluation of Adaptive Systems, Sonthofen, Germany*, pp. 9–24.

Paramythis, A, Totter, A and Weibelzahl, S, 2003, A framework for the evaluation of adaptive systems. Unpublished Work presented at *the Second Workshop on Empirical Evaluation of Adaptive Systems (in conjunction with UM2003), Johnstown, PA*.

Paris, C, Wan, S, Wilkinson, R and Wu, M, 2001, Generating personal travel guides—and who wants them? In M Bauer, PJ Gmytrasiewicz and J Vassileva (eds) *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer, pp. 251–253.

Payne, JW, Bettman, JR and Schkade, DA, 1999, Measuring constructed preferences: towards a building code. *Journal of Risk and Uncertainty* **19**, 243–270.

Petrelli, D, De Angeli, A and Convertino, G, 1999, A user centered approach to user modeling in J Kay (ed.) *UM99 User Modeling: Proceedings of the Seventh International Conference*. New York: Springer, pp. 255–264.

Pirolli, P and Fu, WT, 2003, SNIF-ACT: a model of information foraging on the world wide web in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 45–54.

Preece, J, Rogers, Y, Sharp, H and Benyon, D, 1994, *Human–Computer Interaction*. Reading, MA: Addison-Wesley.

Resnick, P and Varian, H R, 1997, *Special issue on recommender systems*, Communications of the ACM, **40** 1997.

Rubin, J, 1994 *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, 1st edn. New York: John Wiley & Sons.

Ruvini, JD, 2003, Adapting to the user's Internet search strategy in P . Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 55–64.

Salton, G and McGill, M, 1984, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Santos Jr, E, Nguyen, H, Zhao, Q and Pukinskis, E, 2003, Empirical evaluation of adaptive user modeling in a medical information retrieval application in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 292–296.

Sarwar, BM, Konstan, JA, Borchers, A, Herlocker, J, Miller, B and Riedl, J, 1998, Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system in *Proceeding of the ACM Conference on Computer Supported Cooperative Work (CSCW), Seattle, WA*, pp. 439–446.

Scarano, V, Barra, M, Maglio, P and Negro, A, 2002, GAS: Group Adaptive Systems in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 47–57.

Sears, A and Shneiderman, B, 1994, Split menus: effectively using selection frequency to organize menus. *ACM Transactions on Computer–Human Interaction* **1**(1), 27–51.

Semeraro, G, Ferilli, S, Fanizzi, N, Abbattista, F, *et al*, 2001, Learning interaction models in a digital library service in M Bauer, P Gmytrasiewicz and J Vassileva (eds) *User Modelling 2001 (Lecture Notes in Computer Science 2109)*. Berlin, Springer Verlag.

Shapira, B, Shoval, P and Hanani, U, 1997, Stereotypes in information filtering. *Information Processing & Management* **33**(3), 273–287.

Shardanand, U and Maes, P, 1995, Social information filtering for automating 'Word of Mouth' in *Proceedings of CHI-95, Denver, CO*, pp. 210–217.

Smyth, B and Cotter, P, 2002, Personalized adaptive navigation for mobile portals, *Proceedings of the 15th European Conference on Artificial Intelligence – Prestigious Applications of Intelligent Systems*. Lyon, France, 2002.

Specht, M and Kobsa, A, 1999, 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 *Banff, Canada*, 1999.

Stacey, K, Sonenberg, E, Nicholson, A, Boneh, T and Steinle, V, 2003, A teaching model exploiting cognitive conflict driven by a Bayesian network in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 352–362.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press

Strauss, AL and Corbin, JM, 1998, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks: Sage.

Sunderam, AV, 2002, Automated personalization of Internet news in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and AdaptiveWeb-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 348–357.

Tasso, C and Omero, P, 2002, *La Personalizzazione dei Contenuti Web*. Franco Angeli.

Teo, 2001.

Totterdell, P and Boyle, E, 1990, The evaluation of adaptive systems in D Browne, P Totterdell and M Norman (eds) *Adaptive User Interfaces*. London: Academic Press, pp. 161–194.

Walsham G, 1993, *Interpreting Information Systems in Organizations*. Chichester: Wiley.

Weber, G and Specht, M, 1997, User modeling and adaptive navigation support in WWW-based tutoring systems in A Jameson, C Paris and C Tasso (eds) *User Modeling: Proceedings of the Sixth International Conference*. New York: Springer, pp. 289–300.

Weibelzahl, S, 2001, Evaluation of adaptive systems in M Bauer, P Gmytrasiewicz and J Vassileva (eds) *User Modeling 2001 (Lecture Notes in Computer Science, 2109)*. Berlin: Springer, pp. 292–294.

Weibelzahl, S, 2003, Evaluation of adaptive systems. Dissertation, University of Trier, Germany.

Weibelzahl, *et al*, 2003.

Weibelzahl, S, Chin, DN and Weber, G (eds), 2001, *Proceedings of the First Workshop on Empirical Evaluation of Adaptive Systems (in conjunction with UM2003), Sonthofen, Germany*. http://art.ph-freiburg.de/um2001/prodeedings.html.

Weibelzahl, S and Lauer, CU, 2001, Framework for the evaluation of adaptive CBR-systems in U Reimer, S Schmitt and I Vollrath (eds) *Proceedings of the 9th German Workshop on Case-Based Reasoning (GWCBR01)*. Aachen: Shakerm, pp. 254–263.

Weibelzahl, S and Paramythis, A (eds), 2003, *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems (in conjunction with UM2003), Johnstown, PA*. http://art.ph-freiburg.de/um2003/proceedings.html

Weibelzahl, S and Weber, G, 2001, A database of empirical evaluations of adaptive systems in R Klinkenberg, S Rüping, A Fick, N Henze, C Herzog, R Molitor and O Schröder (eds) *Proceedings of Workshop Lernen–Lehren–Wissen–Adaptivität (LLWA 01)*. Research Report in Computer Science no. 763, University of Dortmund, pp. 302– 306.

Weibelzahl, S. &Weber, G, 2002, Adapting to prior knowledge of learners in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and AdaptiveWeb-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 448–451.

Whiteside, J, Bennett, J and Holtzblatt, K, 1988, Usability engineering: our experience and evolution in M Helander (ed.) *Handbook of Human–Computer Interaction* New York: North-Holland, 791–817.

Woo, Y, Ok, S and Mun, Y, 2002, An automatic rating technique based on XML document in P De Bra, P Brusilovsky and R Conejo (eds) *Adaptive Hypermedia and Adaptive Web-Based Systems (Lecture Notes in Computer Science, 2347)*. Berlin: Springer, pp. 424–427.

Zhu, T, Greiner, R and Haeubl, G, 2003, Learning a model of a Web user's interests in P Brusilovsky, A Corbett and F De Rosis *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 65–75.

Zukerman, I, George, S and George, M, 2003, Incorporating a user model into an information theoretic framework for argument interpretation in P Brusilovsky, A Corbett and F De Rosis (eds) *User Modeling 2003 (Lecture Notes in Computer Science, 2702)*. Berlin: Springer, pp. 106–116.

https://doi.org/10.1017/S0269888905000299 Published online by Cambridge University Press