The evaluation of adaptive and personalised information retrieval systems: a review

Catherine Mulwa*, Seamus Lawless,
Mary Sharp and Vincent Wade

Centre for Next Generation Localisation,
Knowledge and Data Engineering Group,
School of Computer Science and Statistics,
Trinity College,
Dublin, Ireland
Email: mulwac@scss.tcd.ie
Email: seamus.lawless@scss.tcd.ie
Email: mary.sharp@scss.tcd.ie
Email: vincent.wade@scss.tcd.ie
*Corresponding author

Abstract: A current problem with the research of adaptive systems is the inconsistency of evaluation applied to the adaptive systems. However, evaluating an adaptive system is a difficult task due to the complexity of such systems. Evaluators need to ensure correct evaluation methods and measurement metrics are used. This paper reviews a variety of evaluation techniques applied in adaptive and user-adaptive systems. More specifically, it focuses on the user-centred evaluation of adaptive systems such as personalised recommender systems and adaptive information retrieval systems. The review tackles the question of 'How have user-centred evaluations of adaptive and user-adaptive systems been conducted and how can these evaluation practices be improved?' Based on the analysed results of the: (a) evaluation approaches, (b) user-centred evaluation techniques, and (c) evaluation metrics, we propose an evaluation framework for end-user experience in evaluating adaptive systems (EFEx).

Keywords: adaptive information retrieval systems; user-adaptive systems; user-centred evaluations; personalised information retrieval systems.

Reference to this paper should be made as follows: Mulwa, C., Lawless, S., Sharp, M. and Wade, V. (2011) 'The evaluation of adaptive and personalised information retrieval systems: a review', *Int. J. Knowledge and Web Intelligence*, Vol. 2, Nos. 2/3, pp.138–156.

Biographical notes: Catherine Mulwa received an Hons BSc degree in Computer Science in 2007 from DIT; in 2008 she received an MSc in Computing (Knowledge Management) from the same institution. Currently she is pursuing her PhD, conducting research in the area of user-centred evaluations of adaptive systems as part of research being carried out at the Centre for Next Generation Localisation (CNGL).

Seamus Lawless is a post-doctoral research fellow in the Centre for Next Generation Localisation (CNGL). The Project is funded by Science Foundation Ireland (SFI) in the category 'Centre for Science Engineering and Technology (CSET)'. He also works on the 1641 Depositions Project.

Mary Sharp graduated from Trinity College Dublin (1988) in Computer Science. She is a Fellow of the Institution of Engineers of Ireland, Irish Computer Society and the Royal Academy of Medicine in Ireland. She is a Chartered Engineer. She is also a lecturer in the SCSS, TCD.

Vincent Wade holds the position of a Professor in the School of Computer Science and Statistics, Trinity College Dublin. Having graduated from University College Dublin with a BSc (Hons) in Computer Science in 1988, he completed his postgraduate studies (MSc and PhD) in Trinity College Dublin. He is Director of the Intelligent Systems Laboratory and also Deputy Director of the SFI Centre for Next Generational Localisation and Personalisation (CNGL). His research is focused on development of innovative digital content management and localisation.

1 Introduction

The research field of adaptive systems has grown rapidly during the past 15 years and this has resulted in terms, models, methodologies, and a plethora of new systems. Adaptive systems in general are becoming more popular as tools for user-driven access to information (Knutov et al., 2009). This has led to the challenge of catering to a wide variety of users in differing environments and user trust issues. This literature review tackles the question of: '*How have user-centred evaluations of adaptive and user-adaptive systems been conducted and how can these evaluation practices be improved?*'

In this paper we distinguish between adaptive systems *as systems that adapts to their environment* and user-adaptive systems *as systems that adapt to their users.* Although the two types of systems overlap, the paper focuses more on the user-adaptive systems.

Evaluation is defined as *the process of examining the product, system components, or design, to determine its usability, functionality and acceptability* (Weibelzahl, 2003), which is measured in terms of a number of criteria essential for any software development project. Evaluation of all systems is important. It is important to not only evaluate but also to ensure that the evaluation uses the correct method (Brusilovsky et al., 2004). This is emphasised in our earlier research on the UCE of adaptive systems (Lawless et al., 2010; Mulwa et al., 2010b). However, the evaluation of adaptive systems is a difficult task due to the complexity of such systems, as shown by many studies (Missier Del and Ricci, 2003; Lavie et al., 2005; Weibelzahl and Weber, 2002; Markham et al., 2003). It is of crucial importance that the adaptive features of the system can be easily distinguished from the general usability of the designed tool. Evaluation of adaptive systems is a crucial stage in their development. Several authors (Höök, 1998; Höök, 2000; Chin, 2001; Masthoff, 2002; Weibelzahl, 2005) have underlined the significance and the difficulties of this task, as well as the lack of user-centred studies and strong models to be followed. All computer systems require comprehensive evaluation; however, this is particularly important in the case of adaptive systems due to their inherent usability problems (Gena, 2006). It is crucial to evaluate these systems both for usability problems at the interface and for the correctness of adaptive solutions. In order to produce effective results, evaluation should occur throughout the entire design cycle and provide feedback for design modification (Gena and Weibelzahl, 2007). As Brusilovsky (2004) argues, given the large set of existing evaluation techniques and systems, the evaluation of adaptive systems and the improvement of such systems is becoming more important than inventing new techniques with questionable benefits. Several forms of evaluation are being conducted in the adaptive e-learning field, both theoretical and empirical. Results of these reviews are quite subjective. Currently there is much debate on how adaptive applications should be evaluated since there is no standard agreed measurement framework for assessing the value and effectiveness of the adaptation yielded by adaptive systems.

The rest of the paper is structured as follows. Section 2 presents an overview of current evaluation practices for adaptive systems. More specifically it focuses on: (a) current evaluation approaches (empirical, layered, utility, heuristic and the user-centred evaluations), (b) the user-centred adaptive variables assessed in the analysed studies, (c) user-centred evaluation methods, (d) the evaluation metrics for adaptive systems, (e) an overview of the evaluation of user-adaptive systems specifically focusing on personalised recommender systems and adaptive information retrieval systems. Section 3 presents pitfalls and problems in the evaluation of adaptive systems and user-adaptive systems. Section 4 presents the proposed evaluation framework for end user experience in evaluating adaptive systems. Finally Section 5 concludes and recommends future work.

2 The evaluations of adaptive systems

The researchers acknowledge that evaluation of adaptive systems is of utmost importance and should become a common practice. It is important to not only evaluate but also to ensure that the evaluation uses the correct methods since an incorrect method can lead to wrong conclusions (De Jong and Schellens, 1997; Gena and Weibelzahl, 2007). The evaluation of these systems is a fundamental stage in their development, yet standardised, comprehensive and recursive evaluation is not a common practice. It is the contention of these authors that such evaluation should become common practice.

Recently, reviews on the state of the art in evaluation practices for adaptive systems have been presented by several researchers (Gena, 2005; Gena and Weibelzahl, 2007; Van Velsen et al., 2008; Mulwa et al., 2010b). Van Velsen et al. (2008) conducted an extensive review on user-centred evaluation studies of adaptive and adaptable systems. The researchers took a descriptive approach which involved: mapping the current user-centred evaluation practice, reflecting on its weaknesses and providing suggestions for improvement (e.g. the need to report think-aloud protocols in more detail than current practice). Most of the methods discussed would be categorised as data collection methods (as identified previously in Gena and Weibelzahl, 2007), rather than evaluation methods; of these, questionnaires were identified as the most popular method in user-centred studies of adaptive systems (Van Velsen et al., 2008). These reviews have identified the most commonly used evaluation approaches, which we'll discuss in following sub-sections.

2.1 Overview of evaluation approaches for adaptive systems

This section introduces the main approaches for evaluation of adaptive systems. It focuses on the user-centred evaluation (UCE) approach of adaptive systems. Key potential benefits of this approach are: (a) savings in terms of time and cost, (b) ensuring the completeness of system functionality, (c) minimising required repair efforts, and (d) improving user satisfaction (Nielsen, 1993). The UCE can serve three goals: (a) verifying the quality of an adaptive system, (b) detecting problems in the system functionality or interface, and (c) supporting adaptivity decisions (De Jong and Schellens, 1997). These functions make UCE a valuable tool for developers of all kinds of systems, because they can justify their efforts, improve upon a system or help developers to decide which version of a system to release.

2.1.1 Empirical evaluation approach

Weibelzahl (2003) acknowledges that empirical research is absolutely necessary for an estimation of the effectiveness, efficiency, and usability of a system that applies artificial intelligent techniques in real-world scenarios. Empirical evaluations (also known as controlled experiments) refer to the appraisal of a theory by observation in experiments. These evaluations help to estimate the effectiveness, efficiency and usability of a system and may uncover certain types of errors in the system that would remain otherwise undiscovered. The researchers acknowledge that the key to good empirical evaluation is the proper design and execution of the experiments so that the particular factors to be tested can be easily separated from other confounding factors. This method of evaluation is derived from empirical science and cognitive and experimental psychology (Gena, 2005). Empirical studies are very good at identifying design errors and false assumptions but they do not suggest new theories or approaches directly. Evaluators are faced with the problem of defining control groups for those systems that either cannot switch off the adaptivity, or where a non-adaptive version appears to be absurd because adaptivity is an inherent feature of these systems (Höök, 2000).

2.1.2 Layered evaluation approach

In the past, several researchers have attempted tackling the problem of evaluating adaptivity by 'Decomposing' and evaluating it in a 'Piece-wise' manner. Previously it has been proved that separating the evaluation of different aspects can help to identify problems in the adaptation process. Paramythis et al. (2010) study on layered evaluations of interactive adaptive systems provided detailed analysis of the layered evaluation approach. The researchers concluded that the main postulation of layered evaluation of these systems was that adaptation needs to be decomposed and assessed in layers in order to be evaluated effectively. Since the first introduction of the term in 2000, the scientific community has adopted this concept in planning and conducting empirical studies. The researcher acknowledge that many authors explicitly refer back to the foundational papers published on the topic to justify experimental designs, to provide rationale for goals or structure of their evaluation studies (Ortigosa and Carro, 2003; Gena, 2005; Goren-Bar et al., 2005; Petrelli and Not, 2005; Arruabarrena et al., 2006; Glahn et al., 2007; Kobsa, 2007; Nguyen and Santos Jr, 2007; Stock et al., 2007; Ley et al., 2009; Limongelli et al., 2008; Popescu, 2009; Santos and Boticario, 2009) or to demonstrate methodological shortcomings of existing studies (Masthoff, 2002; Gena, 2005; Brusilovsky et al., 2006; Yang and Huo, 2008; Brown et al., 2009). The fact that layered evaluation received such a high level of attention in the literature reaffirms the claim that the evaluation of adaptive systems implicates some inherent difficulties. The researchers also accept that the benefits of layered evaluation are perhaps representatively illustrated by a set of studies of a mobile adaptive multimedia guide system for museums called PEACH (Stock and Zancanaro, 2007)

2.1.3 Utility-based evaluation approach

Herder (2003) proposed a utility-based approach which uses layered evaluation techniques and theories on uncertainty and utility from the field of artificial intelligence. The researchers pointed out how interpretable user models would facilitate evaluation; and suggested when choosing a modelling technique that produced implicit representations, researchers should weigh its advantages against the loss of interpretability and also indicated why researchers should decide on common sets of evaluation criteria and methods that are used by researchers in some domain (Herder, 2003).

2.1.4 Heuristic evaluation approach

A heuristic is a general principle or rule of thumb that can be used to critique existing decisions or guide a design decision. An approach which integrates layered evaluation and heuristic evaluation has been proposed (Magoulas et al., 2003). The use of heuristics ensures that the entire system can be evaluated in-depth and specific problems can be discovered at an early design stage before releasing a running prototype of a system. This approach can help evaluators by improving the detection and diagnosis of potential usability problems. Heuristic evaluation identifies usability problems without indicating how they are to be fixed. It is difficult to expect it to address all usability issues when evaluators are not domain experts.

2.1.5 User-centred evaluation approach

Several researchers accept that evaluators of adaptive and user-adaptive systems should adopt a user-centred evaluation approach because users are both the main source of information and the main target of the application. Potential benefits of the user-centred design (UCD) include: (a) the provision of a better understanding of the problem; (b) the rapid testing and validation of story concepts before time consuming coding; (c) the provision of a clear, sociable visual representation of the project vision; (d) the provision of usability by stealth; (e) the engagement of the end-user as a customer; (f) the improvement of the basis for estimation; and (g) the mitigation of project risk etc. User-centred system design brings together task analysis which tells us how people currently accomplish a task, requirement analysis which tells us what a system should do and usability testing which tell us whether a system performs acceptably when a user tries to carry out certain tasks.

Studies have shown that existing evaluation approaches such as the layered approach, empirical approach, utility approach and heuristic approach have not managed to solve the usability issues and that users still encounter inherent usability problems (Tintarev and Masthoff, 2009). In order to address these issues and problems, we propose a user-centred evaluation (UCE) approach when evaluating adaptive systems. The following section provides a detailed review of the UCE evaluation approach. It also introduces the UCE evaluation methods and adaptive variables assessed in the analysed UCE studies.

2.2 User-centred evaluations of adaptive systems

2.2.1 User-centred adaptive variables assessed in studies

The analysis of the survey identified a total of 21 adaptive variables (also known as concepts) that can prompt adaptivity (Mulwa et al., 2010b). These variables are classified under dependent and independent variables and refer to the features of the user that are used as a source of the adaptation, i.e. to what features of the user the system can adapt its behaviour. From the analysed studies, although different names were used by the authors, the concepts being measured were often identical. Brusilovsky et al. (2006) identified adaptive features such as following features which are currently used by existing adaptive hypermedia systems: (i.e. user's goals, knowledge, background and hyperspace, experience, and preferences). Furthermore, the same researcher in 2001 added two more variables: user's interest and individual traits. Furthermore, Kobsa (2007) reviewed the techniques used for personalised hypermedia presentation and described several categories of user data which formed the basis for adaptation in a number of systems developed since 2000, such as demographic data, user's knowledge, user's skills and capabilities, user's interests and preferences, and user's goals and plans. In addition the researchers underlined the significance of computer usage (i.e. interaction behaviour, current task, and interaction history) that can be taken into account when adapting hypermedia pages to the needs of the end user (Kobsa, 2007). On the other hand, Magoulas and Dimakopoulos (2005) explored the dimensions of individual differences that should be included in a student model specification to meet personalisation services requirements and create personalised information access. Van Velsen et al. (2008) identified 13 variables concerning UCE of adaptive and adaptable systems. The researchers grouped them under the following categories: (a) variables concerning attitude and experience (i.e. appreciation, trust and privacy issues, user experience and user satisfaction; (b) variables concerning actual use (i.e. usability, user behaviour and user performance; (c) variables concerning system adoption (intention to use, perceived usefulness); and (d) variables concerning system output (appropriateness of adaptation, comprehensibility and unobtrusiveness) (Van Velsen et al., 2008). The researchers provide a summary of how often each variable was addressed in the 63 studies they analysed and accept that wordings of most variables spoke for themselves.

In the studies selected for our review, 'usability' proved to be the most frequently measured, followed by 'perceived usefulness' and 'appropriateness' of adaptation. Very little research has been conducted to investigate and identify the potential benefit of these variables in evaluating the end-user experiences of adaptive systems. The authors are undertaking research in an attempt to address this issue. It is very important that extensive research be conducted and proper reporting be done. The next section introduces existing UCE methods for adaptive systems. Evaluation of any system should ensure it uses the correct methods, in order to yield significant results.

2.2.2 User-centred evaluation methods

Today, industry is in need of user experience evaluation methods for adaptive systems, examples of such systems are: adaptive e-Learning systems, adaptive information retrieval systems and personalised recommender systems. User-centred design is still the key to designing for good user experience. User experience is conducted in order to improve these systems although in most cases it is tied to context. Several methods exist for understanding users and generating solutions (i.e. probes). Evaluation techniques are concrete methods to carry out the validation of the system. However, in most cases it is often neglected, due to many problems associated with the evaluation of adaptive systems. A critical review is provided for these selected publications. This survey is motivated by the idea that it can indicate what are the techniques used and also provide insightful information on potential benefits of UCE to evaluating adaptive and user-adaptive systems.

The methodologies for evaluating adaptive systems are generally borrowed from the methodologies used in human-computer interaction (HCI) and those utilised for the evaluation of the information selection process. 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 be also used for the evaluation of the information selection process in order to collect user data important to the analysis of the process (Gena, 2005). A complete analysis of HCI methodologies can be found in Gena (2005). The researcher also discusses data collection methods and evaluation metrics for the evaluation of user-adapted systems. The methods and techniques collected and analysed from the investigated studies are listed in Table 1. It is worth noting that every study uses multiple evaluation methods and some of the methods (e.g. data log analysis) are not 'user-centred' according to our definition of UCE, because they do not collect subjective feedback from, or about, (potential) users.

Table 1 Overview of UCE methods (techniques) summarised in the studies ($x = 56$)

Classification	*Evaluation method/instrument*	*Phase of evaluation*	*Variables most frequently assessed*	*Publication examples*
Collection of user's opinions	Interviews, questionnaires (online, post-test, pre/post-test, verbal), focus group, discussion groups	Preliminary	Usability, perceived usefulness, intention to use, trust and privacy issues, appropriateness of adaptation	Gena, 2005; Van Velsen et al., 2008; Masthoff, 2006; Díaz et al., 2008
Observing and monitoring usage	User observation, the systematic observation, verbal protocol	Requirement	Usability, user behaviour	Gena, 2005; Van Velsen et al., 2008
	Think out loud protocols	Preliminary	Usability of interface adaptation	Gena, 2005
Predictive evaluation	Heuristic evaluation, expert review, parallel design, cognitive walkthroughs, social-technical models	Requirement & preliminary	Usability of interface adaptation & user, domain and interface knowledge, privacy, transparency, appropriateness	Gena, 2005; Van Velsen et al., 2008
Formative evaluation	Wizard of Oz simulation, scenario-based design, prototypes	Preliminary	Early prototype evaluations, evaluation before implementation	Van Velsen et al., 2008; Gena, 2005; Masthoff, 2006

Table 1 Overview of UCE methods (techniques) summarised in the studies ($x = 56$) (continued)

Classification	*Evaluation method/instrument*	*Phase of evaluation*	*Variables most frequently assessed*	*Publication examples*
Experiments and tests	Usability testing, experimental evaluation	Final	Interface (and content) adaptation	Gena, 2005; Van Velsen et al., 2008
	Cultural probes	Requirement	–	Masthoff, 2006
	Creative brainstorming sessions	–	–	Van Velsen et al., 2008; Masthoff, 2006
	Empirical observations	–	–	Díaz et al., 2008
Task analysis	Questionnaire, interviews, ethnographic observation, verbal protocols	Requirement	Real user actions	Gena, 2005; Díaz et al., 2008
	Quantitative, Grounded Theory	Final	To combine qualitative evaluation, to discover new theories	Díaz et al. 2008; Gena, 2005
	Prototyping	Preliminary	Evaluation of vertical or horizontal prototype	Gena, 2005
	Cooperative evaluation	Final	Collaboration with real users during the final evaluation step	Gena, 2005

Paramythis et al. (2010) have shown how the traditional layered evaluation methods need to be tailored to suit the particular requirements of adaptivity in the user-system interaction. It has also described some methods (e.g. User-as-Wizard) that are specific to the adaptive systems field (Paramythis et al., 2010). The researchers suggest the best method to employ at any one time primarily depend on when the evaluation takes place (with respect to the system's development lifecycle) and the characteristics of the layer under consideration. These systems can clearly benefit from the many methods available in the field of HCI, to involve users in system design and evaluation. Results from the analysed studies showed that questionnaires, interviews, focus groups and discussions, think-aloud protocols and expert reviews were the most frequently used methods.

In most of the analysed UCE studies, the results of the user-centred evaluations were poorly reported and 'sloppy'. Most of the studies did not report how specific systems were evaluated and presented no evaluation results. The EFEx framework proposed in this paper will address this issue (see Section 4).

2.3 Evaluation metrics for adaptive systems

In the scientific literature, several metrics have been proposed for evaluation of frameworks/systems which enable the development of adaptive frameworks (Raibulet and Masciadri, 2009; Mulwa et al., 2010a). Table 2 presents a summary of evaluation metrics proposed by several researchers. However, it was not explained in the paper whether these metrics were just being proposed or whether they were actually applied to the adaptive systems themselves. Potential benefits and advantages of these metrics include the specification of a common vocabulary for different design, implementation and performance issues of adaptivity. The metrics are known to provide a common means for the evaluation of adaptive systems by considering both the quality of their design (e.g. through the architectural and structural metrics) and their performance (through the interaction and performance metrics). In adaptive systems measurements metrics such as performance, amount of requested material, duration of interaction, number of navigation steps, task success, usability (e.g. effectiveness, efficiency and user satisfaction) has been conducted. The measurement criteria for evaluating the usability of the user interface identified in the analysed studies included: aesthetic, consistency, self-evidence, and naturalness of metaphors, predictability, richness, completeness, motivation, hypertext structure, autonomy, competence and flexibility. The main gap in the state of the art is that UCE metrics for adaptive systems have not been thoroughly investigated and very few studies have been conducted. Although adaptivity, the ability to adapt is an important property of adaptive systems; so far little thought has been given to its evaluation and there are no specific adaptation metrics.

Table 2 Summary of evaluation metrics for adaptive systems

Metrics category	*Name of metric*	*Purpose of metric*
Interaction metrics	AiAI: Administrator interaction Adaptivity index	This metric compares the actions performed by an administrator to manage the system before and after adding the adaptive part. Whenever an action differs, misses or is added, this index increases by one.
	UiAI: User interaction Adaptivity index	It compares the actions performed by a user to access a functionality of a system before and after adding the adaptive part. Whenever an action differs, an additional one is needed or one is missing this index increases by one.
	pLatency: performance Latency	It is used to indicate the delay of the system's responses in the presence of Adaptivity with respect to the response in the absence of adaptivity.
	pQoR: performance Quality of Response	This metric indicates the increase of the quality of the system's responses in the presence of adaptivity.
	pIA: performance Influence on Adaptivity	It indicates if the adaptive strategies are influenced by the other performance metrics.

Table 2 Summary of evaluation metrics for adaptive systems (continued)

Metrics category	*Name of metric*	*Purpose of metric*
Personalisation metric	MpAC: Minimum personalisation adaptive cost	This metric indicates the percentage of entities which are personalised in a framework considering only the minimum number of entities necessary to make a system adaptive.
	AvgpACF: Average personalisation adaptive cost per functionality	This metric indicates the average cost for introduction of adaptive mechanisms per functionality.
	MpOCF: Minimum personalisation Overall Cost	It indicates the percentage of entities which are personalised in the entire system (functional and adaptive) with respect to the minimum number of entities in the adaptive part necessary to make a system adaptive.
	pOCF: personalisation Overall Cost per Functionality	This metric indicates the percentage of entities which are personalised in the entire system with respect to the total number of adaptive entities necessary to provide functionality.
	AvgpACF: Average personalisation overall cost per functionality	This metric indicates the average overall cost for introduction of adaptive mechanisms per functionality. If the result is equal to the one obtained for the AvgpACF, then the functional part is not modified.
	ApOC: Adaptive personalisation Overall Cost	This metric indicates the percentage of the personalisation of the adaptive part with respect to the personalisation of the entire system
	DSAI: Domain specific Adaptivity index	This metric indicates the percentage of the factors specific to the application domain which influences the adaptive part of a system.
Information retrieval	Accuracy of recommendations, accuracy of retrieval	This metrics have been exploited to evaluate adaptation

Gupta and Grover (2004) suggest that the design phase can be evaluated by using metrics such as structural complexity metrics, navigational metrics and usability methods. The researchers acknowledge some metrics, e.g. behavioural complexity, reliability metrics, precision, software size and length metrics help in evaluation of the system as a whole (Gupta and Grover, 2004).

2.4 The evaluations of user-adaptive systems

In the previous section we reviewed the UCE of adaptive systems. However, in the introduction we noted that there is a particular subset of user-adaptive systems whose focus is entirely based on personalisation based on the user such as recommender systems and personalised information retrieval systems (or personalised information retrieval). This section provides a review of evaluation techniques for this subset of adaptive systems. Other examples include systems such as (adaptive user interfaces, recommender systems, reconnaissance agents, adaptive information retrieval, user modelling, personal assistants, personalisation, information filtering, ambient intelligence, adaptive hypertext systems, intelligent tutoring systems and online help systems, etc.).

2.4.1 Overview of evaluations of personalised recommender systems

Personalised recommender systems learn about a user's needs, and identify and suggest information items (news articles, images, videos, etc.) that meet those needs. User needs can be explicitly or implicitly defined either in the form of user tastes, interests and goals, or by system parameters and configurations. Most research efforts in the Recommender Systems field can be said to have been directed towards either defining and improving techniques that provide item recommendations from available preference data, or defining techniques for learning the latter.

Recently, there has been a vast amount of research in the field of evaluation of recommender systems, mostly focusing on designing new algorithms for recommendations. Recommender systems are software applications that aim to support users in their decision making, while interacting with information systems by pre-selecting the information a user might be interested in. Several researchers have discussed various topics relevant to the evaluation of recommender systems in technology enhanced learning environments (TELE). Recommendation techniques aim to estimate ratings for items that have not yet been consumed by users, based on the known ratings users provided in the past. Recommender systems suggest items of interest to users based on available information such as previous usage patterns, the usage patterns of other users and features of the items themselves (Montaner et al., 2003). Main types of recommender systems include 'collaborative filtering techniques' which provide recommendations to a user by using the preferences of other users that have similar preferences to him (Sarwar et al., 2001; Linden et al., 2003). Herlocker et al. (2004) provide an extensive survey of possible metrics for evaluation of recommender systems. The researchers compared a set of metrics, concluding that for some pairs of metrics using both together. Identifying the appropriate criteria for evaluating the true benefits of a recommender framework is challenging issue. After analysis of evaluation metrics, it is clear there is a lack of uniformity in the current metrics for the evaluation of recommender systems, which perhaps is due to the large number of them. In order to address current issues and problems encountered while evaluating recommender systems, a hybrid of UCE methods (techniques) and metrics is proposed. These methods have proved to be very effective in evaluating recommender adaptive systems (Table 1 presents an overview of proposed UCE methods); these techniques have previously been used in evaluations of adaptive systems and proved to be effective. Personalisation is a significant issue which tailors and customises learning experience to individual learners, based on an analysis of the learners objectives, current status of skills and knowledge, and learning style preferences (Cheung et al., 2010; Lawless et al., 2010).

2.4.2 Overview of evaluations of adaptive information retrieval systems

Recently, research has been undertaken exploring how to enhance and combine key aspects of adaptive hypermedia (AH) research with information retrieval (IR) research to provide advanced annotation, slicing, retrieval and composition of multilingual digital content drawn from corporate documents repositories as well as open corpus sources(Jones and Wade, 2006; Steichen et al., 2009). We call such systems, which combine AH and IR approaches to deliver personalised information seeking and access, adaptive information retrieval systems (AIRS). Adaptive information retrieval (AIR) is defined as a search process that adapts toward the user's needs and context. The goal of AIR research is to develop retrieval technology that can predict what information a searcher will need and decide how and when to present that information to the user.

3 Pitfalls and problems in evaluation of adaptive systems

The evaluation of an adaptive system is a difficult task due to the complexity of such systems, as shown by many studies (Missier Del and Ricci, 2003; Lavie et al., 2005). It is of crucial importance that the adaptive features of the system can be easily distinguished from the general usability of the designed tool. Issues arise in the selection of applicable criteria for the evaluation of adaptivity. The evaluation of adaptive systems is not easy, and several researchers have pointed out potential pitfalls and challenges when evaluating adaptive systems (see Table 3). Several reasons have been proposed as being responsible for this shortcoming (Missier Del and Ricci, 2003). Several researchers and evaluators of adaptive systems have identified some pitfalls encountered by developers of these systems (Gena and Weibelzahl, 2007; Tintarev and Masthoff, 2009; Mulwa et al., 2010b). It is crucial that evaluators evade well-known pitfalls and that writers of future evaluation reports increase their empirical value, by reporting the used methodology and results in such a fashion that replication of the study is possible. In order to address these problems, we propose a new methodology. A hybrid evaluation approach which combines the layered evaluation methods identified by Paramythis et al. (2010) (see Table 4) and the user-centred evaluations methods (see Table 1). Previously the researchers have emphasised the significance of applying this methodology (Lawless et al., 2010; Mulwa et al., 2010c).

Table 3 Pitfalls and problems identified in the analysed studies

Pitfalls in evaluation of adaptive systems
• Statistically insignificant results: Adaptivity is typically used when individual users differ. However, differences in approach and preferences are likely to lead to a large variance in performance results, which makes it more difficult to produce statistically comparable results. In order to produce significant results, large volumes of queries and users are required. There are few general guidelines for the selection of these measurements.
• Difficulty in defining the effectiveness of adaptation: It can be difficult to define what constitutes a useful or helpful adaptation.
• Insufficient resources: To fully evaluate an adaptive system it is often necessary to have a large number of individuals interacting with the system. This is in part due to the expected variance between participants mentioned above.
• Too much emphasis on summative rather than formative evaluation: Evaluations often measure only how good or bad a system is rather than providing information on where the problems are and how a system can be improved.

Table 3 Pitfalls and problems identified in the analysed studies (continued)

Problems in evaluation of adaptive systems
• Avoiding confounding factors, sometimes the measured impact of explanations on effectiveness may be confounded with the impact of the accuracy of the recommender system. For example in a recommender system if the recommended items are meant to be liked by the user, and the recommender system has bad accuracy, it would be hard to distinguish between the effects of bad explanations and bad accuracy.
• Reporting the results: Even a perfect experimental design will be worthless if the results are not reported in a proper way. In particular statistical data require special care, as the finding might be not interpretable for other researchers if relevant information is skipped. This problem obviously occurs in other disciplines and research areas that deal with empirical findings, too.
• Specification of adequate control conditions: Another problem, that is inherent to the evaluation of adaptive systems, occurs when the control conditions of experimental settings are defined. In many studies the adaptive system is compared to a non-adaptive version of the system with the adaptation mechanism switched off.
• Approximation of experience verses the real one. Sometimes it can be really difficult and time consuming for participants to really experience the recommendation items
• Allocation of sufficient resources: The fact that evaluations are usually scheduled for the end of a project often results in a radical constriction or even total cancellation of the evaluation phase, because the required resources have been underestimated or are depleted. Empirical work, in particular the data assessment and analysis, require a high amount of personnel, organisational and sometimes even financial resources (Masthoff, 2002). Experiments and real world studies require a considerable amount of time for planning, finding participants, performing the actual data assessment, coding the raw data and statistical analysis.
• Criteria to use (i.e. generalisation of problem). Sometimes when evaluating adaptive systems, there is shortage of information about what exactly they are being evaluated on (McNee et al., 2006)
• Too much emphasis on summative rather than formative evaluation.
• Appropriate measurement to measure true effectiveness.
• Difficulty in attributing cause: is the adaptation causing the measured effect or another aspect of system functionality or design (e.g. system usability).

Table 4 Overview of layers, related criteria, along with methods that can be used for their evaluation

Layer	*Goal*	*Evaluation criteria*	*Evaluation methods*
Collection of input data (CID)	Check quality of raw input data	Accuracy, latency, sampling rate	Data mining, play with layer, simulated users, cross-validation
Interpretation of the collected data (ID)	Check that input data is interpreted correctly	Validity of interpretations, predictability, scrutability	Data mining ,heuristic evaluation, play with layer, simulated users, cross-validation
Modelling the current state of the 'world' (MW)	Check that constructed models represent real world	Primary criteria: validity of interpretations or inferences, scrutability, predictability; secondary criteria: conciseness, comprehensiveness, precision, sensitivity	Focus group, user-as-wizard, data mining, heuristic evaluation, play with layer, simulated users, cross-validation

Table 4 Overview of layers, related criteria, along with methods that can be used for their evaluation (continued)

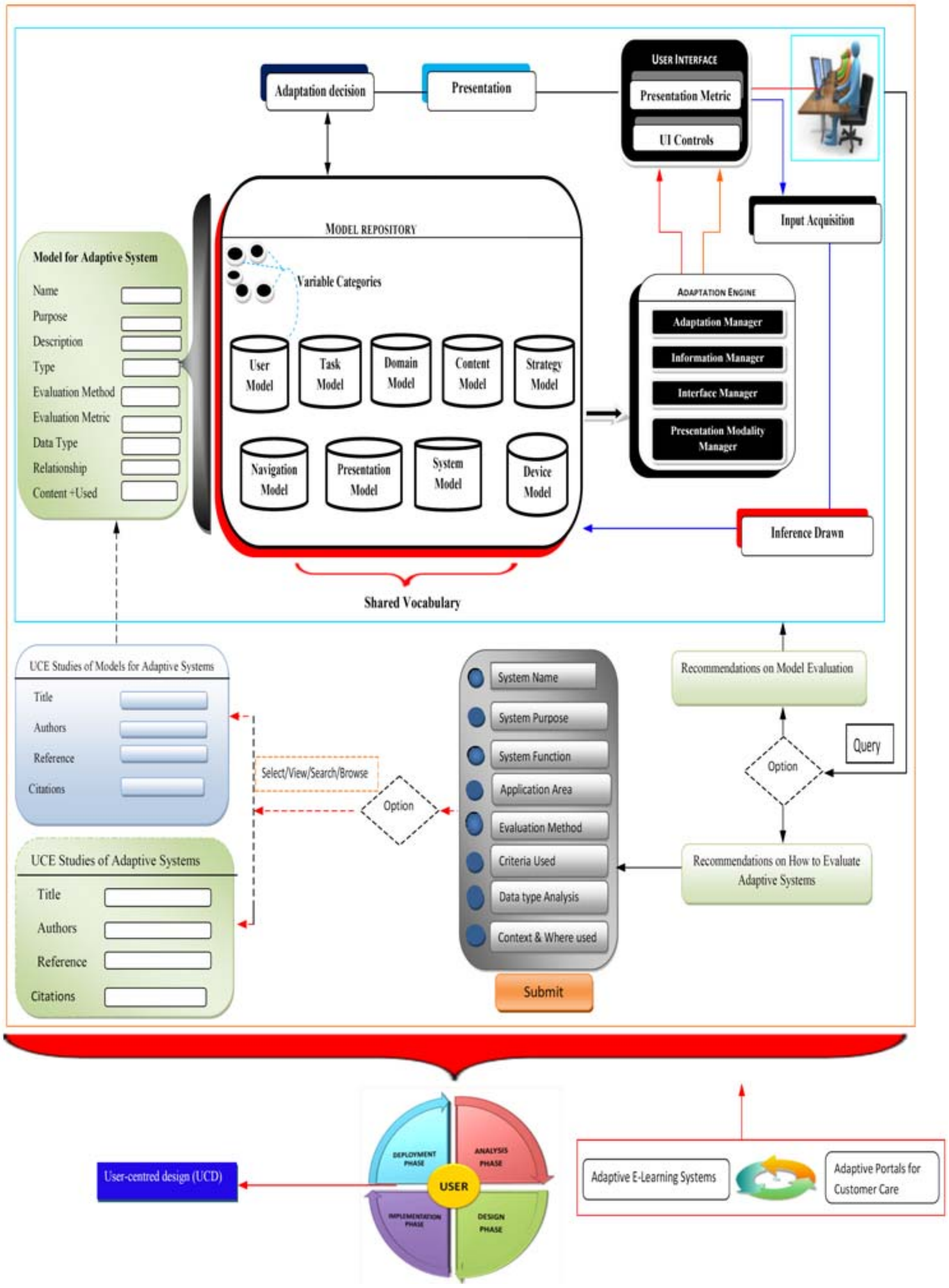
Layer	*Goal*	*Evaluation criteria*	*Evaluation methods*
Deciding upon adaptation (DA)	Determine whether the adaptation decisions made are the optimal ones	Necessity of adaptation, appropriateness of adaptation, subjective acceptance of adaptation, predictability, scrutability, breadth of experience	Focus group, user-as-wizard heuristic evaluation, cognitive walk through, simulated users, play with layer, user test
Applying adaptation decisions (AA)	Determine whether the implementation of the adaptation decisions made is optimal	Usability criteria, timeliness, unobtrusiveness, controllability, acceptance by user, predictability, breadth of experience	Focus group, user-as-wizard, heuristic evaluation, cognitive walkthrough, user test, play with layer
Evaluating adaptation as a whole	Evaluate the overall adaptation theory, may be either formative or summative	Specific for system's objectives or underlying theory	Heuristic evaluation, cognitive walkthrough, user test, play with layer
All layers	–	Privacy, transparency, controllability	Focus group, cognitive walkthrough, heuristic evaluation, user test

4 The proposed evaluation framework

Recommender evaluation frameworks provide personalised services in the adaptive technology enhanced learning systems, and can provide personalised information according to individual information needs. The proposed evaluation framework for end user experience in evaluating adaptive systems (EFEx) was designed and currently is being developed as part of research being carried out at the Centre for Next Generation Localisation (CNGL), which is involved in building interactive adaptive systems which combine adaptive web techniques and technologies, information retrieval (IR) and Adaptive Hypermedia (AH).

A review of UCE approaches, methodologies and techniques adopted by existing systems and frameworks has been conducted and the results analysed. From these results and interviews with domain experts, an architectural design for the EFEx framework was specified and designed as a typical 3-tier architecture which has an interactive and collaborative user interface and consists of: (a) *the presentation layer* which is the topmost level of the application which displays information related to services such as browsing. It communicates with other tiers by outputting results to the browser/client tier and all other tiers in the network. (b) *The business logic layer* which is pulled out from the presentation tier and, has its own layer, it controls an application's functionality by performing detailed processing and (c) *the data persistence layer* which keeps data neutral and independent from application servers or business logic (refer to Figure 1). Giving data its own tier improves scalability and performance.

Figure 1 Architectural design for the recommender section of the EFEx framework (see online version for colours)



Users of EFEx framework are provided with: (a) a repository containing current user-centred evaluation (UCE) studies for adaptive systems; (b) recommendations to users for the identification and application of the most appropriate evaluation methods (techniques), metrics and criteria; (c) translator component which translates the user interface into different languages based on users choice of language; (d) specification of adaptive systems design and structure and (e) a set of components which enable the implementation of UCE.

5 Conclusions and future work

Evaluating adaptive systems is not easy. However, it is of significant importance to ensure scientific progress and to provide convincing arguments that adaptation really does help. It is significant in any evaluation to: (a) decide which criteria to use, (b) avoid confounding factors, (c) take into account domain effects, (d) consider if a metric really measures what the evaluators wants, (e) build up the experiment gradually, and consider limited resources and also (f) take into account the effects of the material you select. The evaluation of these systems is a difficult task due to the complexity of such systems. Evaluators should ensure correct evaluation methods and measurements metrics are used. The results of the analysis of the evaluation approaches, methodologies and metrics adopted by existing systems were used to design an evaluation framework for end-user experience of adaptive systems (EFEx).

The implementation of the framework will be completed and the evaluation techniques identified in this review will be used to validate and evaluate the framework. An extensive review will be conducted specifically focusing on user-centred evaluations of adaptive e-learning systems and adaptive portals for customer care.

Acknowledgements

This research is based upon works supported by Science Foundation Ireland (Grant Number: 07/CE/I1142) as part of the Centre for Next Generation Localization (www.cngl.ie).

References

Arruabarrena, R., Pérez, T., López-Cuadrado, J., Gutiérrez, J. and Vadillo, J. (2006) *On Evaluating Adaptive Systems for Education*, Springer, pp.363–367.

Brown, E., Brailsford, T., Fisher, T. and Moore, A. (2009) 'Evaluating learning style personalization in adaptive systems: quantitative methods and approaches', *IEEE Transactions on Learning Technologies*, Vol. 2, pp.10–22.

Brusilovsky, P. (2004) 'Knowledge-Tree: a distributed architecture for adaptive e-learning', *Proceedings of the 13th International World Wide Web Conference on Alternate Track*, pp.104–113.

Brusilovsky, P., Farzan, R. and Ahn, J. (2006) 'Layered evaluation of adaptive search', *Workshop on Evaluating Exploratory Search Systems at SIGIR06*, Seattle, WA.

Brusilovsky, P., Karagiannidis, P. and Sampson, C. (2004) 'Layered evaluations of adaptive learning systems', *International Journal of Continuing Engineering Education and Lifelong Learning*, Vol. 14, Nos. 4/5, pp.402–421.

Cheung, R., Wan, C. and Cheng, C. (2010) 'An ontology-based framework for personalized adaptive learning', *Advances in Web-Based Learning–ICWL 2010*, pp.52–61.

Chin, D. (2001) 'Empirical evaluation of user models and user-adapted systems', *User Modeling and User-Adapted Interaction*, Vol. 11, pp.181–194.

De Jong, M. and Schellens, P.J. (1997) 'Reader-focused text evaluation. An overview of goals and methods', *Journal of Business and Technical Communication*, Vol. 11, pp.402–432.

Díaz, A., García, A. and Gervás, P. (2008) 'User-centred versus system-centred evaluation of a personalization system', *Information Processing & Management*, Vol. 44, pp.1293–1307.

Gena, C. (2005) 'Methods and techniques for the evaluation of user-adaptive systems', *The Knowledge Engineering Review*, Vol. 20, pp.1–37.

Gena, C. (2006) 'A user-centered approach for adaptive systems evaluation', *5th Workshop on User-Centred Design and Evaluation of Adaptive Systems*, 20–23 June, Dublin, Ireland.

Gena, C. and Weibelzahl, S. (2007) 'Usability engineering for the adaptive web', *The Adaptive Web*, pp.720–762.

Glahn, C., Specht, M. and Koper, R. (2007) 'Smart indicators on learning interactions', *2nd European Conference on Technology Enhanced Learning*, Crete, Greece.

Goren-Bar, D., Graziola, I., Rocchi, C., Pianesi, F., Stock, O. and Zancanaro, M. (2005) 'Designing and redesigning an affective interface for an adaptive museum guide', *Affective Computing and Intelligent Interaction*, Vol. 3784, pp.939–946.

Gupta, A. and Grover, P. (2004) 'Proposed evaluation framework for adaptive hypermedia systems', *3rd Workshop on Empirical Evaluation of Adaptive Systems*, 23–26 August, Eindhoven University of Technology, The Netherlands.

Herder, E. (2003) 'Utility-based evaluation of adaptive systems', *Proceedings of the 2nd Workshop on Empirical Evaluation of Adaptive Systems, at the 9th International conference on User Modelling*, Pittsburgh, USA.

Herlocker, J., Konstan, J., Terveen, L. and Riedl, J. (2004) 'Evaluating collaborative filtering recommender systems', *ACM Transactions on Information Systems*, Vol. 22, No. 1, pp.5–53.

Höök, K. (1998) 'Evaluating the utility and usability of an adaptive hypermedia system', *Knowledge-Based Systems*, Vol. 10, pp.311–319.

Höök, K. (2000) 'Steps to take before IUIs become real', *Journal of Interaction with Computers*, Vol. 12, pp.409–426.

Jones, G. and Wade, V. (2006) 'Integrated content presentation for multilingual and multimedia information access', *New Directions in Multilingual Information Access*, Vol. 40, pp.31–39.

Knutov, E., De Bra, P. and Pechenizkiy, M. (2009) 'AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques', *New Review of Hypermedia and Multimedia*, Vol. 15, pp.5–38.

Kobsa, A. (2007) 'Privacy-enhanced personalization', in Brusilovsky, P., Kobsa, A. and Nejdl, W. (Eds): *Communications of the ACM*, Vol. 50, pp.24–33.

Lavie, T., Meyer, J., Beugler, K. and Coughlin, J. (2005) 'The evaluation of in-vehicle adaptive systems', *User Modeling: Work on the EAS*, pp.9–18.

Lawless, S., Mulwa, C. and O'Connor, A. (2010) 'A proposal for the evaluation of adaptive personalised information retrieval', *Proceedings of the 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*, 28 March, Milton Keynes, UK.

Ley, T., Kump, B., Maas, A., Maiden, N. and Albert, D. (2009) 'Evaluating the adaptation of a learning system before the prototype is ready: a paper-based lab study', *1st International Conference on User Modeling, Adaptation, and Personalization*, Springer, Trento, Italy.

Limongelli, C., Sciarrone, F. and Vaste, G. (2008) 'LS-Plan: an effective combination of dynamic courseware generation and learning styles in web-based education', *5th International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*, Hannover, Berlin, Germany.

Linden, G., Smith, B. and York, J. (2003) 'Amazon.com Recommendations: Item-to-Item Collaborative Filtering', *IEEE Computer Society*, Vol. 7, No. 1, pp.1089–7801.

Magoulas, G., Chen, S. and Papanikolaou, K. (2003) 'Integrating layered and heuristic evaluation for adaptive learning environments', *Proceedings of the 2nd Workshop on Empirical Evaluation of Adaptive Systems, at the 9th International Conference on User Modelling*, Pittsburgh, USA.

Magoulas, G.D. and Dimakopoulos, D.N. (2005) 'Designing personalised information access to structured information spaces', *Proceedings of the 1st International Workshop on New Technologies for Personalized Information Access*, Edinburgh, Scotland, UK.

Markham, S., Ceddia, J., Sheard, J., Burvill, C., Weir, J., Field, B., Sterling, L. and Stern, L. (2003) 'Applying agent technology to evaluation tasks in e-learning environments', *Exploring Educational Technologies Conference*, pp.16–17.

Masthoff, J. (2002) 'The evaluation of adaptive systems', in Patel, N. (Ed.): *Adaptive Evolutionary Information Systems*, Idea Group Publishing, London, pp.329–347.

McNee, S., Riedl, J. and Konstan, J. (2006) 'Making recommendations better: an analytic model for human-recommender interaction', *Proceeding of CHI EA'06 Extended Abstracts on Human Factors in Computing Systems*, 22–27 April, Montreal, Canada.

Missier Del, F. and Ricci, F. (2003) 'Understanding recommender systems: experimental evaluation challenges', *Challenges*, pp.31–40.

Montaner, M., López, B. and Rosa, J.L.D. (2003) 'A taxonomy of recommender agents on the internet', *Artificial Intelligence Review*, Vol. 19, pp.285–330.

Mulwa, C., Lawless, S., Li, W. and Jones, G. (2010a) 'A proposal for the evaluation of simulated interactive information retrieval in customer support', *SIGIR Workshop on the Automated Evaluation of Interactive Information Retrieval*, Geneva, Switzerland.

Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I. and Wade, V. (2010b) 'Adaptive educational hypermedia systems in technology enhanced learning: a literature review', *Proceedings of the 2010 ACM Conference on Information Technology Education*, 7–9 October, Midland, USA.

Mulwa, C., Li, W., Lawless, S. and Jones, G. (2010c) 'A proposal for the evaluation of adaptive information retrieval systems using simulated interaction', *Workshop on the Simulation of Interaction in Automated Evaluation of Interactive Information Retrieval, at the 33rd Annual ACM SIGIR Conference*, Geneva, Switzerland.

Nguyen, H. and Santos Jr, E. (2007) 'An evaluation of the accuracy of capturing user intent for information retrieval', *International Conference on Artificial Intelligence*, Las Vegas, NV, pp.341–350.

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

Ortigosa, A. and Carro, R. (2003) 'The continuous empirical evaluation approach: evaluating adaptive web-based courses', *9th International Conference on User Modelling*, Springer, Berlin.

Paramythis, A., Weibelzahl, S. and Masthoff, J. (2010) 'Layered evaluation of interactive adaptive systems: framework and formative methods', *User Modeling and User-Adapted Interaction*, pp.1–71.

Petrelli, D. and Not, E. (2005) 'User-centred design of flexible hypermedia for a mobile guide: reflections on the hyperaudio experience', *User Modeling and User-Adapted Interaction*, Vol. 15, pp.303–338.

Popescu, E. (2009) 'Evaluating the impact of adaptation to learning styles in a Web-based educational system', *8th International Conference on Web-based Learning*, Springer, Aachen, Germany.

Raibulet, C. and Masciadri, L. (2009) 'Evaluation of dynamic adaptivity through metrics: an achievable target?', *WICSA 2009 WS8 – Adaptive Architectures*, 16 September.

Santos, O. and Boticario, J. (2009) 'Guiding learners in learning management systems through recommendations', *4th European Conference on Technology Enhanced Learning*, Nice, France.

Sarwar, B.M., Karypis, G., Konstan, J.A. and Riedl, J.T. (2001) 'Collaborative filtering for information recommendation systems', *Proceedings of the 10th International World Wide Web Conference, WWW10*, Hong Kong.

Steichen, B., Lawless, S., O'Connor, A. and Wade, V. (2009) 'Dynamic hypertext generation for reusing open corpus content', *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pp.119–128.

Stock, O. and Zancanaro, M. (2007) *PEACH: Intelligent Interfaces for Museum Visits*, Springer-Verlag, New York.

Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., Kuflik, T., Not, E. and Rocchi, C. (2007) 'Adaptive, intelligent presentation of information for the museum visitor in PEACH', *User Modeling and User-Adapted Interaction*, Vol. 17, pp.257–304.

Tintarev, N. and Masthoff, J. (2009) 'Evaluating recommender explanations: problems experienced and lessons learned for the evaluation of adaptive systems', *User Modeling, Adaptation and Personalization*, Trento, Italy.

Van Velsen, L., Vander Geest, T., Klaasen, R. and Steehounder, M. (2008) 'User-centered evaluation of adaptive and adaptable systems: a literature review', *The Knowledge Engineering Review*, Vol. 23, pp.261–281.

Weibelzahl, S. (2003) *Evaluation of Adaptive Systems*, PhD Thesis, University of Trier.

Weibelzahl, S. (2005) 'Problems and pitfalls in evaluating adaptive systems', in Chen, S.Y. and Magoulas, G.D. (Eds): *Adaptable and Adaptive Hypermedia Systems*, IRM Press, Hershey, PA, pp.285–299.

Weibelzahl, S. and Weber, G. (2002) 'Advantages, opportunities and limits of empirical evaluations: evaluating adaptive systems', *Künstliche Intelligenz*, Vol. 16, pp.17–20.

Yang, D. and Huo, H. (2008) 'Assessment on the adaptivity of adaptive systems', *International Conference on Management of e-Commerce and e-Government*, Nanchang, China, pp.437–440.