A recommender framework for the evaluation of end user experience in adaptive technology enhanced learning

Catherine Mulwa*, Seamus Lawless, Ian O'Keeffe, 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: Ian.OKeeffe@scss.tcd.ie
Email: mary.sharp@scss.tcd.ie
Email: vincent.wade@scss.tcd.ie
*Corresponding author

Abstract: Adaptive Technology Enhanced Learning (TEL) has attracted significant interest with the promise of supporting individual learning tailored to the unique circumstances, preferences and prior knowledge of a learner. However, the evaluation of the overall performance of such systems is a major challenge, as the adaptive TEL system reacts differently for each individual user and context of use. Evaluation of such systems is significant but very complex area of research in itself since depending on the aspect of personalisation that needs to be evaluated. Several evaluation techniques need to be combined and executed differently. This paper proposed a novel recommender framework built upon an evaluation educational data set using a hybrid recommend approach to identify appropriate procedures. Recommendations are to software developers and users of adaptive TEL systems. A review and analyses of evaluation studies on adaptive TEL systems was conducted. Based on the analysed results, an educational evaluation data set was created.

Keywords: education evaluation data set; evaluation approaches for personalised TEL; adaptive TEL systems; recommender evaluation framework.


Biographical notes: Catherine Mulwa received a BSc (Hons) in Computer Science in 2007 from DIT; in 2008 she received an MSc in Computing (Knowledge Management) from the same institution. Currently she is in her 4th year of 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 an Assistant Professor works as a Post-Doctoral Research Fellow in the Centre for Next Generation Localisation (CNGL). He also works on the 1641 Depositions Project.
Mary Sharp is an Assistant Professor. She 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 and a Chartered Engineer and a lecturer in the SCSS, TCD.

Ian O’Keeffe is a Research Fellow on the Personalised and Collaborative Informal Learning across Connected Environments (PERCOLATE) project at TCD.

Vincent Wade 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 Director of the SFI Centre for Next Generational Localisation and Personalisation (CNGL).

1 Introduction

Adaptive Technology Enhanced Learning (TEL) has attracted significant interest with the promise of supporting individual learning tailored to the unique circumstances, preferences and prior knowledge of a learner. Evaluation of the overall performance of adaptive TEL systems is difficult, complex and major challenge, as the adaptive system reacts differently for each individual user and context of use. Evaluation of these system has become a significant but very complex area of research in itself since depending on the aspect of personalisation that needs to be evaluated (e.g. quality of the user modelling, performance of different adaptation approaches, knowledge gain from using the adaptive system or overall end user experience); several evaluation techniques need to be combined and executed differently. Evaluation of such systems include: (a) an evaluation of learner knowledge level at the training session; (b) an evaluation of learner satisfaction level; (c) completion date of training session and (d) faster results.

The main contributions of this paper are: (a) the proposed novel recommender framework for evaluating adaptive TEL systems built upon an evaluation educational data set using a hybrid (case-based and knowledge-based approach) approach to identify appropriate evaluation methods, metrics and criteria and (b) an educational evaluation data set for TEL systems. This approach overcomes the limitations of case-based and knowledge-based. Based on a user’s description of their adaptive system and evaluation purpose, it recommends evaluation approaches, evaluation methods, measurement criteria and metrics. The data set is based on peer reviewed evaluation cases. Thus, rather than being a large data set based on many users behaviour. It is based on a smaller data set that has been quality reviewed. Moreover, the data set has the ability to grow overtime as the framework itself provides a mechanism for published authors to add their evaluation cases to the data set; thus we believe the data set is already and in the future could become a very valuable dataset for TEL evaluation choices. Running multiple recommender algorithms and systems over the data set could provide a means of comparison of recommender systems accuracy.

We use recommendation technology to enhance the appropriateness of suggestions of evaluations procedures (evaluation approaches, techniques, methods, metrics and criteria) of adaptive TEL systems. The data set to our knowledge is the first harvested data set of
A recommender framework for the evaluation of end user experience

selections of adaptive evaluation approaches, methods, metrics and criteria for personalised TEL. The key aspect being this information is not arbitrary selections from novice end users but peer reviewed informed choices from published researchers. Although recommendation systems can be applied to large data sets, this does not mean that recommendation systems are inappropriate or not needed to solve complex information problems. In particular, the multi-attribute relationships which need to be traversed by humans to work out what are the most appropriate evaluation procedures (i.e. evaluation approaches, methods/techniques, metrics and criteria are not easily navigated using typical database techniques).

The research described in this paper is attempting to show how a recommender that use case-based and knowledge-based technique to provide such answers. We aim to address the question of: “What are the techniques used in evaluating personalised adaptive technology enhanced learning systems: Can a recommender framework be used to appropriately support the evaluation of such systems”. Evaluation is defined as the process of examining the product, system components or design, to determine its usability, functionality and acceptability (Mulwa et al., 2010b), which is measured in terms of a number of criteria essential for any software development project. Evaluation is central to ensuring the quality of academic practice. In addition to the quality issues, personalised TEL involves high levels of investment that need to be justified. Evaluation has, therefore become important across the sector as a means to demonstrate effectiveness and cost-effectiveness of learning technologies thus the significance of the proposed framework and data set.

The rest of the paper is structured as follows. In Section 2, we discuss evaluation approaches for personalised TEL and dimensions of personalisation. Section 3 presents an evaluation educational data set. Section 4 introduces the proposed hybrid web-based recommender framework for evaluating end user experience in adaptive systems and educational benefits of the framework. Finally, Section 5 concludes the paper and recommends future work.

2 Evaluation approaches for technology enhanced learning and dimensions of personalisation

Evaluation can give valuable hints for improvements by uncovering unexpected behaviour of the learning system and by identifying incongruence between user expectations and system design. The diverse problems associated with Information Systems (IS) evaluation have been widely reported in the normative literature (Irani, 2002). It is significant to ensure the correct evaluation procedures (i.e. instruments, approaches, techniques, metrics and criteria) are used when evaluating such systems. Evaluation determines the merit or importance of artefacts and therefore requires profound evaluation approach.

An evaluation approach for learning resource is considered as any procedure, method, set of criteria, tool, checklist or any other evaluation/verification instrument and mechanism which has the purpose of evaluating the quality of learning resources. Vourikari et al. (2008) conducted a review of evaluation approaches for learning resources. The researchers performed a tentative classification of these approaches and discovered that a plethora of evaluation approaches for digital learning resources existed (Vourikari et al., 2008). The authors noted in some cases that these approaches relied on
a national educational requirement, whereas in other cases the repository had its row quality requirement. Furthermore, Manouselis and Costopoulou (2006) acknowledge a diversity of evaluation approaches for learning resources exists (such as models, methods, criteria and instruments) that are applied to ensure the quality of the learning resources (Manouselis and Costopoulou, 2006).

2.1 Evaluation approaches for personalised technology enhanced learning

Technology enhanced learning investigates ‘how information and communication technologies can be used to support learning and teaching and competence development throughout life’ (Svetsky et al., 2010). Evaluation is a significant and critical activity in evaluation of personalised TEL. The reasons for evaluating such systems are similar to the reasons for evaluating any type of learning provisions. These might include: (a) to determine whether the TEL solution is accomplishing its objectives; (b) to identify who benefited the most or the least from the TEL programme and (c) to identify areas for improvement. Evaluations provide valuable feedback about potential users’ perceptions of the TEL system, how well the software is written and the extent to which the system really does support decision making (Jiang and Klein, 1999).

Personalisation can be performed on an individualised, collaborative or aggregate scope. Individualised personalisation is when the system’s adaptive decisions are taken according to the interests of each individual user as inferred from their user model (Speretta and Gauch, 2005; Teevan et al., 2005). Collaborative personalisation is when information from several user models is used to determine or alter the weights of interests in other user models (Sugiyama et al., 2004). It can be implemented on an aggregate scope when the system does not make use of user models; in which case personalisation is guided by aggregate usage data as exhibited in search logs (i.e. implicitly inferred general users’ interests from aggregate history information) (Agichtein et al., 2006; Smyth and Balfe, 2006). In TEL personalisation refers to the dynamic nature of content, support and presentation in the Technology Enhanced Learning Environment (TELE), which is altered to suit each learner. The type of personalisation which the learning environment aims to accommodate has to be considered. Personalised TEL has attracted significant interest with the promise of supporting individual learning tailored to the unique circumstances, preferences and prior knowledge of a learner. However, the evaluation of the overall performance of such personalised TEL system is a major challenge, as the personalised system reacts differently for each individual user and context of use. The evaluation of such systems is significant in order to determine whether the TEL solution is accomplishing its objectives. However, TEL’s technology aspect brings in other demands such as accountability, including measuring return on investment.

Several evaluation approaches have been used in evaluating personalised TEL (Vavoula and Sharples, 2009). Examples include:

- The *quality approach* used in European e-Learning which investigates the current state of e-Learning quality in Europe. It is based on a survey by the European Quality Observatory (EQA), the European platform for quality in e-Learning involving 1700 participants from all European countries (Ehlers et al., 2005).

- The *lifecycle approach* to educational technology evaluation which places evaluation at the centre of the development process, from the early stages of design to a final
assessment of deployed technology in use. This approach draws on evaluation method and ideas from software engineering educational evaluation and models for evaluating learning.

- The combined and a layered evaluation approach (Drachsler et al., 2010) used to measure the impact of TEL recommendations. The layered evaluation approach separates the ‘interaction assessment’ and the ‘adaptation decision’. Both layers should be evaluated separately in order to effectively interpret the evaluation results (Paramythis et al., 2010). This has a number of advantages over other approaches, such as useful insight into the success or failure of each separate adaptation stage, facilitation of improvements, generalisation of evaluation results and reuse of successful practices.

- The combined four-level by Kirkpatrick Schenkel and six-level approach by Breitner and Hoppe (2005) which focuses mainly on pedagogical objectives.

- The User-Centred Evaluation (UCE) Approach, which has proved to be significant in: verifying the quality of the framework; detecting problems in the system functionality and user interface; and supporting adaptivity decisions (De Jong and Schellen, 1997). These functions make UCE a valuable tool for software developers of all kinds of systems, because they are capable of justifying their efforts, improving upon a system or assisting developers in deciding which version of a system to release. Benefits of UCE approach include: (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 empirical approach derived from empirical science and cognitive and experimental psychology 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.

- In the utility approach the evaluation can be seen as a utility function X that maps a system, given some user context, to a quantitative representation of user satisfaction or performance (Mulwa et al., 2011).

- The recommender approaches such collaborative filtering, content-based, demographic, the knowledge-based and hybrid.

Majority of recommender educational adaptive systems apply collaborative filtering (CF) approach although a few apply the hybrid approach. The underlying assumption of the CF approach is that those who agreed in the past tend to agree again in the future but in reality that is not always the case. Developers encounter problems such as issues in obtaining context information. Using hybrid approaches can avoid some limitations and problems of pure recommender systems, like the cold-start and user trust problem.

### 2.2 Dimensions of personalisation

Several dimensions of personalisation exist (user modelling, personalisation and adaptation mechanism, knowledge gain and user experience) which can be categorised along a number of dimensions. Hybrid personalisation can be seen in many TEL systems, whereby a range of approaches are taken along one of the dimensions. Following is a brief discussion of these dimensions:
User modelling: Macarthur and Conlan (2010) acknowledge that user models are sets of structured stores of information that represent the competencies and tasks carried out by the learner (Macarthur and Conlan, 2010). The researchers emphasise that the development and integration of successful models requires the combination of technological approaches such as AI or rule-based programming with pedagogical and psychological approaches. User models in conjunction with reasoning processes are used to identify, model, track and foster learners’ abilities. This reasoning is carried out with the user model as well as other types of model such as the domain, narrative or pedagogical model. Aspects of the learner that are modelled are their knowledge or skills, cognitive abilities, prior experience, objectives, motivation, interests, needs, attitudes, learning style, preferences, tasks and abilities. The user model can be categorised as either dynamic or static. These models represent the inferred state of the learner or expected state of the learner.

Adaptation mechanism: Adaptation process aims to change the student’s learning path during his navigation according to his learning progresses. To be effective in the learning process, it is necessary to adapt the system, to the specific student’s needs, as opposite to the traditional ‘one-size-fits-all’ approach (Brusilovsky, 2001). In adaptable systems, the learner is given control over their TEL environment; in this case they have an option to customise or tailor the systems resources to match their own personal requirement. Furthermore, in adaptive personalisation it is the system that dynamically creates and takes control of the stored data, selection of learning resources or support to present to the user.

Knowledge gain and user experience: Personalisation strategies aid knowledge gain and support learner’s decision making, which leads to improved user experience. Personalised TEL systems facilitate the progression towards a target skill, knowledge or cognitive strategy. Since learners display individual differences in their learning competencies, adaptation and personalisation has an important role to play in providing increased learning experience.

3 An educational evaluation data set for personalised learning

Several researchers have noted the need for data sets that can be used as benchmarks to compare different recommendation approaches in TEL (Drachsler et al., 2010; Verbert et al., 2011). The researchers investigated a number of steps that may be followed in order to develop referenced data sets that can be adopted and reused by the scientific community. Data sets for educational TEL are many-folded as TEL takes place in the whole spectrum of learning roughly distinguished between formal and non-formal learning settings. Although recommender systems are increasingly applied in TEL, it is still an application area that lacks publically available, comparable, interoperable and reusable data sets that cover the spectrum of formal and informal learning. So although a large volume of research has been conducted on recommender systems in TEL, the community lacks data sets that would allow experimental comparative evaluation of the performance of different recommendation algorithms.
3.1 Educational evaluation data set for adaptive systems

Recently, an educational evaluation data set has been created as part of the research described in this paper which is significant to both the developers and end users of adaptive systems such as recommender adaptive educational systems (Mulwa et al., 2011). A review and analyses of a range of evaluation studies of personalised technology enhanced systems was conducted over a period of 2 years. Based on this analysis, an evaluation data set (for personalised e-Learning) was created (see Table 1). The data set is a characterised, structured and interlinked list of evaluation approaches, methods, metrics and measurement criteria extracted from over 90 papers from the literature of adaptive systems. The data set being used in this paper is based on peer reviewed evaluation cases. Thus, rather than being a large data set based on many users behaviour. It is based on a smaller data set that has been quality reviewed. Moreover, the data set has the ability to grow overtime as the framework itself provides a mechanism for published authors to add their evaluation cases to the data set; thus we believe the data set is already and in the future could become a very valuable dataset for TEL evaluation choices. Running multiple recommender algorithms and systems over the data set could provide a means of comparison of recommender systems accuracy. This data set is stored in a MySQL database, the schema of which is presented in Figure 1. The data set is broken down into four distinct interlinked components which are described in more detailed below:

- **Evaluation Approaches**: In this case there are only five approaches (User-centred evaluation approach, layered, utility-based, heuristic and empirical) discussed in section 2.

- **Evaluation Method/Instrument**: The methodologies for evaluating adaptive recommender systems are generally borrowed from the methodologies used in human computer interaction (HCI) and those utilised for the evaluation of the information selection process (Gena, 2005). From the analysed studies: (i.e. questionnaires, experimental evaluation, interviews, user observations, usability testing and user text) were the most commonly used evaluation methods respectively. Questionnaires are used to collect data from respondents by allowing them to answer a set of questions either on paper or online.

### Table 1 Evaluation data set for evaluating adaptive recommender systems

<table>
<thead>
<tr>
<th>Evaluation method/instrument</th>
<th>Criteria (variables)</th>
<th>Metrics</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviews, Questionnaires</td>
<td>Usability, Perceived usefulness, Intention to use, user goals, knowledge of the domain, Background and Hyperspace experience, preferences trust and privacy issues, Appropriateness of adaptation</td>
<td>Accuracy of recommendations, accuracy of retrieval, AiAI: Administrator interaction, Adaptivity index</td>
<td>(Gena, 2005b); (Van Velsen et al., 2008); (Masthoff, 2006); (Raibulet and Masciadri, 2009)</td>
</tr>
</tbody>
</table>
Table 1 Evaluation data set for evaluating adaptive recommender systems (continued)

<table>
<thead>
<tr>
<th>Evaluation method/instrument</th>
<th>Criteria (variables)</th>
<th>Metrics</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>User observation, the</td>
<td>Usability, User behaviour, user goal, knowledge of domain, background and hypertension experience, user interests individual traits (e.g. cognitive or learning style), environment (e.g. location, locale, software, hardware), user situation awareness</td>
<td>behavioural complexity, reliability metrics, precision, software size and length metrics</td>
<td>(Gupta and Grover, 2004); (Magoulas and Dimakopoulos, 2005); (Brusilovsky, 2001)</td>
</tr>
<tr>
<td>systematic observation,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verbal protocol, Data mining,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Play with layer, Simulated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>users, Cross-validation,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristic evaluation, Play</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with layer, Simulated users,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross validation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristic evaluation,</td>
<td>Usability of interface adaptation &amp; user, domain and interface knowledge, user</td>
<td>pQoR: performance</td>
<td></td>
</tr>
<tr>
<td>expert review, parallel</td>
<td>performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>design, cognitive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>walkthroughs, social-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>technical models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wizard of Oz simulation,</td>
<td>Early prototype evaluations, evaluation before implementation</td>
<td>pIA: performance on Adaptivity</td>
<td>(Masthoff, 2006)</td>
</tr>
<tr>
<td>scenario-based design,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prototypes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usability testing,</td>
<td>Interface (and content) adaptation, usage data (user history), user cognitive</td>
<td>MpAC: Minimum personalisation adaptive cost</td>
<td>(Magoulas and Dimakopoulos, 2005)</td>
</tr>
<tr>
<td>experimental evaluation</td>
<td>workload, groups of users</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultural probes, Focus</td>
<td>preferences, user interests, user skills and capabilities, user performance</td>
<td>AvgpACF: Average personalisation adaptive cost per</td>
<td>(Santos, 2008); (Masthoff, 2006); (Paramythis et al., 2010)</td>
</tr>
<tr>
<td>group, User-as-wizard</td>
<td></td>
<td>functionalty</td>
<td></td>
</tr>
<tr>
<td>Heuristic evaluation,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive walk through,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulated users, Play with</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>layer, User test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creative brainstorming</td>
<td>privacy, transparency, appropriateness, appreciation, trust and privacy issues, user</td>
<td>MpOCF: Minimum personalisation Overall Cost</td>
<td>(Van Velsen et al., 2008)</td>
</tr>
<tr>
<td>sessions, Focus group,</td>
<td>experience, user satisfaction, usability, user behaviour, intention to use, perceived</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User-as-wizard</td>
<td>usefulness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative, Grounded</td>
<td>To combine qualitative evaluation, to discover new theories</td>
<td>ApOC: Adaptive personalisation Overall Cost</td>
<td>(Diaz et al., 2008); (Gena, 2005)</td>
</tr>
<tr>
<td>Theory, Cognitive walkthrough,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heuristic evaluation,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User test, Play with layer,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooperative evaluation,</td>
<td>Evaluation of vertical or horizontal prototype, Collaboration with real users during</td>
<td>DSAI: Domain specific Adaptivity index</td>
<td>(Gena, 2005)</td>
</tr>
<tr>
<td>verbal protocols, Focus group</td>
<td>final evaluation step</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Participants can choose one or multiple choices or can freely answer in writing.

- **Experimental evaluation**: Current evaluation approaches recommend experimental methods (techniques) in lab settings as a way of coping with adaptive systems complexity and identifying the aspects of these systems that require improvement. Furthermore interviews, user observations and usability testing. In interviews (i.e. structured, fixed questions or semi-structure), participants normally are asked in person by an interviewer. The manner in which interview results were reported, make it seem that evaluators consider interviews to be inferior to statistical data.

- **Usability testing methods** are used in user-centred interaction design to evaluate a system by testing it on users. It focuses on measuring the systems capacity to meet its intended purpose. In total, 40 evaluation methods were mentioned in the studies.

- **Measurement criteria (adaptive variables)**: Adaptive variables refer to features of the user that are used as a source of the adaptation (Triantafillou et al., 2007). In total, 50 variables were mentioned in the studies and were grouped into different categories of variables concerning (i.e. attitude and experience, actual use, system adoption and system output).
Usability was most frequently measured, followed by user satisfaction which is a subjective variable which can be influenced by factors (i.e. system effectiveness, user effectiveness, user characteristics, effort and effectiveness).

Perceived usefulness, user performance and intention to use: Keinonen defines usability as ‘a characteristic related to: (a) the products design process (b) the product itself (c) use of product (d) user experience of the product and user expectation’ (Nokelainen, 2006). These are attributes which can be measured through subjective user experience.

Metrics: In the studies we analysed that accuracy of recommendations metric was the most frequently used, followed by accuracy of retrieval.

Recently, a survey on evaluations of adaptive systems from 2000 to date was conducted. Table 2 presents the survey question. If the reader wishes to fill in the survey is online (see http://surveymonkey.com/s/Q2DSDF8). Over 100 domain experts from the User Modelling, Adaptation and Personalisation (UMAP), Adaptive Hypermedia and Adaptive Recommender communities have completed the questionnaire. The analysed results will be used to supplement and enhance the data set described above which was created from the research literature. This data set is used to feed the recommendation process (Figure 2).

Figure 2  Process by which a recommendation is produced (see online version for colours)
Table 2  Survey questions

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>General questions</td>
<td>Have You Developed an Adaptive System in the Past (from 2000 to 2011)</td>
</tr>
<tr>
<td>Questions aimed at identifying evaluation data</td>
<td></td>
</tr>
<tr>
<td>If You Have Developed an Adaptive System:</td>
<td></td>
</tr>
<tr>
<td>- What was improved by Adaptivity?</td>
<td></td>
</tr>
<tr>
<td>- What is the Variation Type of the Adaptive System You have Developed</td>
<td></td>
</tr>
<tr>
<td>- Please Tick the Meta Data Models Your System Uses. If You Conducted A Whole-System Evaluation,</td>
<td></td>
</tr>
<tr>
<td>- What Evaluation Methods did you use?</td>
<td></td>
</tr>
<tr>
<td>- What criteria did you use?</td>
<td></td>
</tr>
<tr>
<td>If you conducted evaluations of specific metadata models of adaptive system,</td>
<td></td>
</tr>
<tr>
<td>(For each model evaluated, please indicate which evaluation methods and criteria you used).</td>
<td></td>
</tr>
<tr>
<td>- What Evaluation Methods did you use? During this Evaluation (Conducted above),</td>
<td></td>
</tr>
<tr>
<td>- What Evaluation methods, Criteria and Metrics did You Use to Measure Performance against these criteria?</td>
<td></td>
</tr>
<tr>
<td>Perceived quality, Usability and Usefulness</td>
<td>Which of the following features of the Framework would you find (consider) useful? i) Recommendation on how to combine different methods, metrics and metrics to evaluate a personalised recommender adaptive system, ii) repository of state-of-the-art review of UCE and layered evaluation of PIR systems</td>
</tr>
</tbody>
</table>

3.2 Uses-case scenario: process of recommendation

All the recommendations are delivered by the proposed recommender framework based on the data collected in the evaluation data set (Section 3.1). This includes the evaluation approaches, methods, metrics and measurement criteria recommended to users.

Suppose user X who is a novice user and Y is an expert; both want to get recommendations on how to evaluate a new adaptive recommender TEL system and do not know which evaluation approach, method, metric and criteria to use.

In the initial stage of recommendation, we ask user X to: (a) select evaluation purpose, the system characteristics and variation type of their system from a list. In the next step, we recommend an evaluation approach to user X. User Y is provided with a list option of select evaluation approach from a list. The next step involves recommending list evaluation approaches, methods, metrics and measurement criteria to both users. Then the next step the recommended outputted resources are bundled (i.e. methods, metrics and criteria).

Both users are presented with an interface which also enables them to view a ranked list of evaluation studies of adaptive TEL systems, evaluation approaches, methods, metrics and criteria published from 2000 to date. Each study has a title, author, citation, reference and pdf of the study. Each evaluated system mentioned in this study is described in terms of: system name, the functions it fulfils, the purpose, application area, evaluation, approaches, methods, metrics, criteria, adaptation benefit and a URL of the developer. Each metadata model of these systems is described in terms of: name, purpose, description, model type, evaluation methods, data type, evaluation metrics and model relations. Throughout this process both users are provided with explanations on
The recommended resources are produced. These explanations are generated to give feedback about the process used to derive the recommendation conclusions and are significant in: (a) providing transparency and expose the reasoning and data behind a recommendation; (b) helping to inspire user trust and loyalty; (c) increasing satisfaction and (d) making it quicker and easier for users to find what they want persuade the user to try and use the recommended (Ricci et al., 2010). Our goals of providing explanations in the recommendation are to provide: (a) transparency, validity and trustworthiness; (b) persuasiveness, effectiveness and efficiency; (c) satisfaction and relevance; (d) comprehensibility and education (Masthoff, 2006; Tintarev, 2007). The whole process from recommendation is depicted in Figure 2.

4 The proposed hybrid recommender evaluation framework

Personalised TEL recommender frameworks, systems and prototypes are software tools that aim to support learners in decision-making processes while interacting with large information spaces. These applications provide a personalised means for users to find and evaluate items of interest. These tools recommend items to users based on implicit or explicit preferences. Recommender applications help users overcome the information overload problem by exposing the most significant, relevant and interesting items.

4.1 Recommendation problem addressed

Currently very few recommender evaluation frameworks for adaptive TEL exist. Due to the complexity of such systems and usability issues, the evaluation of these systems requires correct evaluation procedure. Novice users encounter difficulties in deciding what approach to use and how to perform evaluations. To address these problems, we use recommendation technology to enhance the appropriateness of suggestions of evaluations procedures (i.e. techniques) of adaptive TEL systems. A review of evaluation procedures was conducted and the results were used create the data set (Table 1). Fifteen domain experts from AH, AIR and adaptive recommender communities were interviewed. The results of the review and feedback from the interviews were used to specify, design and develop the proposed framework. The framework is subdivided into four major subsections: (a) a hybrid (i.e. case-based and knowledge-based) recommender system built upon the educational evaluation data set discussed in Section 3; (b) a search engine consisting of three search components which allow users to search for: evaluation techniques of nine metadata models (i.e. user, domain, content, strategy, presentation, navigation, device, task and system) of adaptive systems and UCE and layered evaluation studies of adaptive systems from 2000 to date; (c) a TEL support centre which has a UCE methodology which illustrates to learners how to use these UCE techniques; (d) a taxonomy which defines these UCE techniques. Learners are also provided with personalised information, which is translated according to individual learning needs. The focus of this paper is subsection (a)

4.2 Architectural design of the recommender system

The architectural and technical design (Figures 3 and 4) of the framework is designed as a three-tier architecture in which the user interfaces, functional process logic, computer data storage and data access are implemented and maintained as independent modules.
Our goal is to allow any of the three tiers to be upgraded or replaced independently as requirements change. The architecture and technical design has a web-based interactive and collaborative interface consisting of: (a) the presentation layer which is the topmost level of the application which displays information related to services. This tier communicates with other tiers by outputting results to the browser/client tier and all other tiers in the framework; (b) the business logic layer which is pulled out from the presentation tier and, has its own layer, control the frameworks’ functionality by performing detailed processing; (c) the data persistence layer which keeps data neutral and independent from the frameworks server or business logic. Giving data its own layer greatly improves scalability and performance of the framework.

Figure 3  Architectural designs (see online version for colours)

Figure 4  Technical design (see online version for colours)
The different models (user, system, learning portals and studies on UCE and layered evaluations) are populated with the harvested data (Section 3). The user model stores profile. The recommendation engine uses these models to process data during the process of recommendations (Figure 2).

The collaborative nature of this framework facilitates the sharing of information among people in four scientific communities (i.e. TEL, AH, AIR and recommender systems). The recommendation engine reconciles the user model, the system characteristics and the previous evaluation studies to produce a list of relevant studies for that individual.

4.2.1 Process of recommendation

During recommendation process we apply implicit recommendation techniques to personalise and recommend evaluation: methods, metrics and criteria and approaches in a bundle to a user. Supported user tasks include:

Start

Step 1: User selects the system variation type (adaptive recommender systems, personalised information retrieval systems, etc).

Step 2: The user selects from a list of evaluation procedures, the purpose of the evaluation and system characteristics.

Step 3: If the user is a non-expert, the systems recommends approach, otherwise expert users can select a pre-existing list.

Step 4: Using the selected variation type of a system of step 1, the algorithm does the following:

1) Select all the systems belonging to the variation type selected in step 1.
2) Select all the evaluations that have been carried out on the systems in previous step.
3) Using the evaluation approach defined in step 2, the system retrieves all the methods, metrics and criteria from the database along with their evaluation results.
4) All the evaluation results for each method, metric and criteria are stored in a list.
5) Each result has a success score and a flag as to whether this evaluation was carried out specifically for this system or not. If it was it is given extra weight in the scoring process. When all the results for each method, metric and criteria are collated they are added up and the list is sorted by score.
6) The results are presented as a percentage of the highest score in the list which will always have 90%.
7) Then the results are recommended to users in a bundle (i.e. method, metric and criteria).
8) If the methods, metrics and criteria in the list match the methods, metrics and criteria being used in the current evaluation then they are highlighted in the list.
9) Each result as a further flag indicating whether the evaluation was carried out specifically for the considered system or not.
10) Provision of explanation mechanism which explains the recommendations from step 1 to 10

End
4.3 Evaluating the recommender framework

Personalised recommender frameworks have proven to be very effective in the handling of information overload and improved end user experience. It is difficult to evaluate such frameworks since these evaluations involve pure subjective assessments. The full evaluation of this framework is ongoing as the execution of the evaluation of the framework is comprehensive. The evaluation will cover five types of experiments. The first three experiments that are motivated by evaluation protocols in areas such as information retrieval (Shani and Gunawardana, 2011):

- Offline experiments, using pre-collected or simulated data to test the performance of candidate algorithms;
- User studies, where a small group of subjects use a system in a controlled environment and report on their experience;
- Real-life testing, where a system is tested under realistic conditions during its normal operation with its actual users.

The fourth and fifth experiments will be task-based which are significant in evaluating the overall performance and usefulness of the framework. In this case, end users will be presented with a list of tasks that are specific to the particular domain chosen for the experiment. To evaluate the overall performance of the framework we base our technique on internal quality which consists of six characteristics selected from the state-of-the-art based upon other evaluations: (a) functionality, concerned with what the framework does to fulfil user needs; (b) reliability, evaluating the frameworks capability to maintain a specified level of performance; (c) usability, assessing how understandable and usable the framework is (d) efficiency, evaluating the capability of the framework to exhibit the required performance with regard to the amount of resources needed; (e) maintainability, which is concerned with the framework’s capability to be modified; and finally (f) portability, which will involve measuring the frameworks capability to be used in a distributed TELE. The evaluation metrics will be quantitative and qualitative. First, the quantitative measures will determine the efficiency and performance of the framework. The qualitative perspective of the framework will be measured by using System Usability Test (SUS). The characteristics were selected from other successful evaluations conducted from the state-of-art-review.

4.4 Educational benefit of the framework

Researchers and developers are provided with an interactive and collaborative user interface which is translated and personalised into user’s choice of language. TEL practices cater to students and teachers who use many different learning tools and environments and have experience of interaction derived with open, ubiquitous and socially oriented services. The process of learning in formal education no longer takes place solely in traditional, educator-centric settings. Interactive, learner-centric experiences are being used to support learner collaboration, knowledge acquisition and reflection. Learner enquiry, activity and engagement are key requirements in such experiences; and TEL applications are being designed and utilised to meet these requirements. TEL is expected to make a radical difference to education, specifically, the quality and effectiveness of the learning experience with one of its key contributions
being ‘learning’. TEL methods have been known to change the deployment of the most important resource in the education system: teachers’ and the learners’ time (Mulwa et al., 2010a) and also in cost savings to performance and strategic benefits (Rainsford and Murphy, 2005). In most cases, learners using these technologies are able to receive instant and personalised feedback, active engagement, reusable learning materials and a safe environment where one can learn from one’s mistakes and be able to access huge amounts of beneficial material on-demand. These technologies make learning more flexible in terms of time, space and place. A high quality structured educational data set will help support learning.

5 Conclusions and future work

This paper contributed a hybrid (case-based and knowledge-based) recommendation framework built upon an educational evaluation data set. It is crucial that software developers and end user evaluators evade well-known pitfalls and that writers of future evaluation reports increase their empirical value, by reporting the used approaches, methodologies, techniques and results in such a fashion that replication of the study is possible. The data set has the ability to grow overtime as the framework itself provides a mechanism for published authors to add their evaluation cases to the data set. The key aspect being this information is not arbitrary selections from novice end users but peer reviewed informed choices from published researchers. In future a full evaluation will be conducted.

Acknowledgements

This research is based upon works supported by Science Foundation Ireland (grant number: 07/CE/11142) as part of the Centre for Next Generation Localisation (see www.cngl.ie). The authors are grateful for the reviewers’ suggestions.

References


A recommender framework for the evaluation of end user experience


